Fantuan's Academia

FANTUAN'S STATISTICS NOTES SERIES VOL.4

Advanced Optimization Techniques

Author: Jingxuan Xu
Lecturer: Prof. Sinho Chewi

August 27, 2025

Contents

All the Sections with * are hard sections and can be skipped without losing coherence.

This note is a **partial** scribe of S&DS632 course: Advanced Optimization Techniques taught by Prof. Sinho Chewi at Yale University. This course has a lecture note https://chewisinho.github.io/opt_notes.pdf, I just finished all the exercises and steps missing, and added some of my own understanding.

Chapter 1

Introduction

The basic problem of optimization is to compute an approximate minimizer of a given function $f : \mathcal{X} \to \mathbb{R}$. In this note, $\mathcal{X}$ is always taken to be a subset of $\mathbb{R}^d$, but other possibilities are possible.

1.1 Black-Box Optimization and the Oracle Model

1. Black-Box Optimization

In **black-box optimization**, we assume that we can evaluate f, and possibly its derivatives, at any chosen point $x \in \mathcal{X}$.

- Advantage: *Generality*, most time we can evaluate f.
- Disadvantage: *Generality*, no additional structural information about f bringing computational savings.

2. Oracle Model

It is nonsense to talk about optimizing a single function f, since we can just make the algorithm 'output x_*', the minimizer. Therefore, we talk about a class of functions $\mathcal{F}$ of interest, and we require our algorithms to succeed on every $f \in \mathcal{F}$.

The algorithm 'knows' $\mathcal{F}$, but does not know which particular $f \in \mathcal{F}$ it is trying to optimize (otherwise we run into the same issue discussed above). The role of the **oracle** is to act as an intermediary between the algorithm and the function. We assume that the algorithm is allowed to ask certain questions (*queries*) to the oracle for f, e.g.,

- a **zeroth-order oracle** accepts a query point $x \in \mathbb{R}^d$ and outputs $f(x)$.
- a **first-order oracle** accepts a query point $x \in \mathbb{R}^d$ and outputs $(f(x), \nabla f(x))$.
- a **second-order oracle** accepts a query point $x \in \mathbb{R}^d$ and outputs $(f(x), \nabla f(x), \nabla^2 f(x))$.

Remark: the **query complexity** of $\mathcal{F}$ for a particular choice of oracle, as a function of the prescribed tolerance ϵ, is then informally defined to be the minimum number N such that there exists an algorithm which, for any $f \in \mathcal{F}$, makes N queries to the oracle for f and outputs a point x with $f(x) - \min f \leqslant \epsilon$.
However, **query complexity is not the same as computational complexity**. Indeed, query complexity only counts the number of interactions with the oracle, and the algorithm is allowed to perform unlimited computations between interactions.

1.2 Unsolvability of Global Optimization

We will see in this section that the general optimization problems are 'unsolvable', in the sense that there is a lower bound on any algorithm such that it converges so low with no hope to succeed. This section is a scribe of **Lectures on Convex Optimization**[11] by Yurii Nesterov Section 1.1.3.

In order to optimize efficiently, we need to place assumptions on f, ideally minimal ones. For example, we can assume that f is continuous. However, we are interested in *quantitative* rates of convergence for algorithms, and for this purpose, a *qualitative* assumption such as continuity is not enough. A quantitative form of continuity is to assume that f is *L-Lipschitz* in the ℓ_∞ norm:

$$|f(x) - f(y)| \leqslant L \max_{i \in [d]} |x_i - y_i|, \quad \forall\, x, y \in \mathcal{X} \tag{1.1}$$

Also, for concreteness, let us take $\mathcal{X} = [0,1]^d$. In the language of the framework above, we consider the class

$$\mathcal{F} = \{f : [0,1]^d \to \mathbb{R} \,|\, f \text{ is L-Lipschitz in the } \ell_\infty \text{ norm}\}$$

To solve this problem, we use the **uniform grid algorithm**, which only uses zeroth-order oracle, and have one parameter p.

Uniform Grid Algorithm:

1. Form p^d points:
$$x_\alpha = \left(\frac{2i_1 - 1}{2p}, \frac{2i_2 - 1}{2p}, \cdots, \frac{2i_n - 1}{2p}\right)^T$$
where $\alpha \equiv (i_1, i_2, \cdots, i_n) \in \{1, \cdots, p\}^d$.
2. Among all points x_α, find the point $\bar{x}$ with the minimal value of the objective function.
3. The pair $(\bar{x}, f(\bar{x}))$ is the output of the method.

The efficiency of this algorithm then depends on p.

Lemma 1.2.1: Efficiency of Uniform Grid Algorithm

Let $f_\star$ be a global optimal value of f in the class $\mathcal{F}$. Then,

$$f(\bar{x}) - f_\star \leqslant \frac{L}{2p}$$

Proof. For $\alpha = (i_1, \cdots, i_n)$, define

$$X_\alpha = \left\{ x \in \mathbb{R}^d : \|x - x_\alpha\|_\infty \leqslant \frac{1}{2p} \right\}$$

which is the set of closest points with x_α. We have

$$\bigcup_{\alpha \in \{1, \cdots, p\}^d} X_\alpha = [0,1]^d \tag{1.2}$$

Let $x_\star$ be a global solution of the problem. Then, by Equation 2.3, there exists a multi-index α^* such that $x_\star \in X_{\alpha^\star}$. Thus, $\|x_\star - x_{\alpha^\star}\|_\infty \leqslant 1/2p$. Finally,

$$f(\bar{x}) - f(x_\star) \leqslant f(x_{\alpha^\star}) - f(x_\star) \leqslant L\|x_{\alpha^*} - x_\star\| \leqslant \frac{L}{2p}$$

where the first inequality comes from that $\bar{x}$ has the smallest value over all grids, and the second comes from Equation 2.2. □

The upper bound of query complexity of this algorithm is shown below.

Lemma 1.2.2: Upper Bound of Query Complexity

The query complexity of uniform grid algorithm applied to the class $\mathcal{F}$ with precision ϵ is at most $(\lfloor \frac{L}{2\epsilon} \rfloor + 1)^d$.

Proof. Take $p = \lfloor \frac{L}{2\epsilon} \rfloor + 1$. Then, $p \geqslant \frac{L}{2\epsilon}$, in view of Lemma 1.2.1, we have

$$f(\bar{x}) - f_\star \leqslant \frac{L}{2p} \leqslant \frac{L}{2}\frac{2\epsilon}{L} \leqslant \epsilon$$

Note that we need to call the oracle at p^d points. □

Now we show a lower bound, which indicates that we cannot perform better than this upper bound.

Theorem 1.2.3: Unsolvability of Global Optimization

For any $0 < \epsilon < L/2$ and any deterministic algorithm, the query complexity of minimizing the function in class $\mathcal{F}$ with precision ϵ using a zeroth-order oracle is at least $\lfloor \frac{L}{2\epsilon} \rfloor^d$.

Proof. We prove this using the *resisting oracle*, which tries to create the 'worst possible' problem for each particular method.

Let $p = \lfloor \frac{L}{2\epsilon} \rfloor \geqslant 1$. Assume that there exists a method which needs $N < p^d$ calls of oracle to solve any problem from $\mathcal{F}$. Let us apply this method to the resisting strategy: *return* $f(x) = 0$ *at any test point* x. Then, this method can find only $\bar{x} \in [0,1]^d$ such that $f(\bar{x}) = 0$.

However, since $N < p^n$, there exists an $\hat{\alpha}$ such that there were no test points in the box $X_{\hat{\alpha}}$. Define $x_\star = x_{\hat{\alpha}}$, and consider the function

$$\bar{f}(x) = \min\{0, L\|x - x_\star\|_\infty - \epsilon\}$$

Clearly, this function is ℓ_∞-Lipschitz continuous with constant L, and its global optimal value is $-\epsilon$. Moreover, $\bar{f}(\cdot)$ differs from zero only inside the box $X_{\hat{\alpha}}$ (this is because, only inside this box, we have $\|x - x_\star\|_\infty \leqslant 1/2p$, and we have chance to get $L\|x - x_\star\|_\infty \ - \epsilon < 0$). Therefore, $f\bar{(\cdot)}$ is equal to zero at all test points of our method.

Sine the precision of the method is ϵ, we come to the conclusion: If the number of calls of the oracle is less than p^d, then the precision of the result cannot be better than ϵ. Thus, the desired result is proved. □

1.3 The Role of Convexity

Theorem 1.2.3 told us, for $\epsilon < L/4$, the query complexity grows exponentially with the dimension. It is also robust: variants of the result can be proven when the notion of Lipschitzness is w.r.t. the ℓ_2 norm; when the oracle is taken to be a first-order oracle; when the algorithm is allowed to be randomized; etc. The message is clear: in order for optimization to be tractable in the worst case, we must impose some structural assumptions.

The black-box oracles we have been considering are local in nature: given a query point $x \in \mathbb{R}^d$, the oracle reveals some information about the behavior of f in a local neighborhood of x. Assumptions such as Lipschitzness effectively govern how large this local neighborhood is. But ultimately, to render optimization tractable, we must ensure that local information yields global consequences. As justified in this section, a key assumption that makes this possible is **convexity**.

Definition 1.3.1: Convex Set

A subset $C \subseteq \mathbb{R}^d$ is **convex** if for all $x, y \in C$ and all $t \in [0,1]$, we have $(1-t)x + ty \in C$.

Definition 1.3.2: (Strongly) Convex Function

Let C be convex and let $\alpha \geqslant 0$. A function $f : C \to \mathbb{R}$ is **α-convex** if for all $x, y \in C$ and all $t \in [0,1]$,

$$f((1-t)x + ty) \leqslant (1-t)f(x) + tf(y) - \frac{\alpha}{2}t(1-t)\|y - x\|^2 \tag{1.3}$$

When $\alpha = 0$, the function is **convex**. When $\alpha > 0$, we call it **strongly convex**.

The definition above has the advantage that it does not require f to be differentiable. However, for the purposes of checking and utilizing convexity, it is convenient to have the following equivalent reformulations. We focus on $C = \mathbb{R}^d$.

Proposition 1.3.3: Equivalent Formulation of Convexity

Let $C = \mathbb{R}^d$ and $\alpha \geqslant 0$.

1. If f is continuously differentiable, then α-convexity is equivalent to

$$f(y) \geqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2}\|y - x\|^2, \quad \text{for all } x, y \in \mathbb{R}^d \tag{1.4}$$

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geqslant \alpha\|y - x\|^2 \quad \text{for all } x, y \in \mathbb{R}^d \tag{1.5}$$

2. If f is twice continuously differentiable, then α-convexity is equivalent to

$$\langle v, \nabla^2 f(x) v\rangle \geqslant \alpha \|v\|^2 \quad \text{for all } v, x \in \mathbb{R}^d \tag{1.6}$$

Proof. Assuming that f is continuously differentiable.
(3) $\implies$ (1.4): Rearranging Equation 3, for $t > 0$,

$$\begin{aligned} & f((1-t)x+ty) \leqslant (1-t)f(x) + tf(y) - \frac{\alpha}{2}t(1-t)\|y-x\|^2 \\ \implies\quad & tf(y) \geqslant tf(x) + f((1-t)x+ty) - f(x) + \frac{\alpha}{2}t(1-t)\|y-x\|^2 \\ \implies\quad & f(y) \geqslant f(x) + \frac{f(x+t(y-x))-f(x)}{t} + \frac{\alpha(1-t)}{2}\|y-x\|^2 \end{aligned}$$

Sending $t \searrow 0$, we have Equation 1.4:

$$f(y) \geqslant f(x) + \langle \nabla f(x), y-x\rangle + \frac{\alpha}{2}\|y-x\|^2$$

(1.4) $\implies$ (4.8): Swap x and y in Equation 1.4, we have

$$f(x) \geqslant f(y) - \langle \nabla f(y), y-x\rangle + \frac{\alpha}{2}\|y-x\|^2$$

Add this equation to Equation 1.4, we have

$$\begin{aligned} & f(x)+f(y) \geqslant f(x)+f(y) + \langle \nabla f(x), y-x\rangle - \langle \nabla f(y), y-x\rangle + \alpha\|y-x\|^2 \\ \implies\quad & \langle \nabla f(y) - \nabla f(x), y-x\rangle \geqslant \alpha\|y-x\|^2 \end{aligned}$$

(4.8) $\implies$ (3): By Fundamental Theorem of Calculus, along the path $\gamma(s) = x + s(y-x)$, $s \in [0,1]$, for $v := y-x$, we have

$$f(y) - f(x) = \int_\gamma \nabla f \cdot \mathrm{d}\gamma = \int_0^1 \langle \nabla f(\gamma(s)), \gamma'(s)\rangle\,\mathrm{d}s = \int_0^1 \langle \nabla f(x+sv), v\rangle\,\mathrm{d}s \tag{1.7}$$

Similarly, by Fundamental Theorem of Calculus, along the path from x to $x+tv = (1-t)x+ty$, we have $\gamma(s) = x+stv$, then

$$f((1-t)x+ty) - f(x) = \int_\gamma \nabla f \cdot \mathrm{d}\gamma = \int_0^1 \langle \nabla f(\gamma(s)), \gamma'(s)\rangle\,\mathrm{d}s = \int_0^1 \langle \nabla f(x+stv), tv\rangle\,\mathrm{d}s \tag{1.8}$$

Using (1.8$-t$1.7), we have

$$f((1-t)x+ty)-(1-t)f(x)-tf(y) = \int_0^1 \langle \nabla f(x+stv), tv\rangle\,\mathrm{d}s - t\int_0^1 \langle \nabla f(x+sv), v\rangle\,\mathrm{d}s = -t\int_0^1 \langle \nabla f(x+sv) - \nabla f(x+stv), v\rangle\,\mathrm{d}s$$

By Equation 4.8, since $x+sv-(x+stv) = s(1-t)v$, we finally arrive

$$\begin{aligned} f((1-t)x+ty)-(1-t)f(x)-tf(y) &= -t\int_0^1 \langle \nabla f(x+sv) - \nabla f(x+stv), v\rangle\,\mathrm{d}s \\ &= -t\int_0^1 \frac{1}{s(1-t)}\langle \nabla f(x+sv) - \nabla f(x+stv), s(1-t)v\rangle\,\mathrm{d}s \end{aligned}$$

$$
\begin{aligned}
&\leqslant -t\int_0^1 \frac{1}{s(1-t)}\alpha s^2(1-t)^2\|v\|^2\,\mathrm{d}s && \text{(By equation 4.8)}\\
&= -t\int_0^1 \alpha s(1-t)\|v\|^2\,\mathrm{d}s\\
&= -\frac{\alpha}{2}t(1-t)\|v\|^2
\end{aligned}
$$

which is exactly the Equation 3.

Assume f is twice continuously differentiable.
(4.8) $\Longrightarrow$ (1.6): Let $y = x + \epsilon v$ in Equation 4.8, we have

$$
\langle \nabla f(x+\epsilon v) - \nabla f(x), \epsilon v\rangle \geqslant \alpha\epsilon^2\|v\|^2
$$

Divide by ϵ^2 on both sides, we have

$$
\left\langle \frac{\nabla f(x+\epsilon v) - \nabla f(x)}{\epsilon}, v\right\rangle \geqslant \alpha\|v\|^2
$$

Take $\epsilon \to 0$, we have that

$$
\langle v, \nabla^2 f(x) v\rangle \geqslant \alpha\|v\|^2
$$

which is the form of Equation 1.6.
(1.6) $\Longrightarrow$ (4.8): Apply Fundamental Theorem of Calculus on $\nabla f(x)$, with integration path $\gamma(s) = x + s(y-x)$, $s \in [0,1]$, we have

$$
\begin{aligned}
\langle \nabla f(y) - \nabla f(x), y - x\rangle &= \int_0^1 \langle \nabla^2 f(x + s(y-x))(y-x), y-x\rangle\,\mathrm{d}s\\
&\geqslant \int_0^1 \alpha\|y-x\|^2\,\mathrm{d}s = \alpha\|y-x\|^2 && \text{(By Equation 1.6)}
\end{aligned}
$$

which is the form of Equation 4.8. □

They have each own interpretations: for $\alpha = 0$, Equation 3 states that f lies below each of its secant lines between the intersection points; Equation 1.4 states that f globally lies above each of its tangent lines; Equation 4.8 states that ∇f is a monotone vector field; Equation 1.6 states about the curvature.

Before describing the feature of convexity that local information yields global conclusions, we recall some basic facts about optimization. For simplicity, we consider unconstrained optimization throughout.

Lemma 1.3.4: Existance of Minimizer

Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuous and its level sets be bounded. Then, there exists a global minimizer of f.

Proof. Let $x_0 \in \mathbb{R}^d$ and let $\mathcal{K} = \{f \leqslant f(x_0)\}$ denote the level set. By the continuity assumption, $\mathcal{K}$ is closed and bounded, thus compact. Let $\{x_n\}_{n\in\mathbb{N}}$ be a minimizing sequence, $f(x_n) \to \inf f$. By compactness, it admits a convergent subsequence $\{x_{n_q}\}_{n_q\in\mathbb{N}}$, which converges to $x_\star \in \mathbb{R}^d$. By continuity, we have $f(x_\star) = \lim_{n\to\infty} f(x_{n_q}) = \inf f$. □

Next we state the necessary and sufficient conditions for optimality.

Lemma 1.3.5: Necessary Conditions for Optimality

Let $f : \mathbb{R}^d \to \mathbb{R}$ be minimized at $x_\star$.

1. If f is continuously differentiable, then $\nabla f(x_\star) = 0$.
2. If f is twice continuously differentiable, then $\nabla^2 f(x_\star) \succcurlyeq 0$.

Proof. 1. Let $v \in \mathbb{R}^d$ and $\epsilon > 0$. Then, $f(x_\star + \epsilon v) - f(x_\star) \geqslant 0$ since $x_\star$ is the minimizer. If f is continuously differentiable, we have, by Fundamental Theorem of Calculus, with integration path $\gamma(t) = x_\star + t\epsilon v$, $t \in [0, 1]$,

$$f(x_\star + \epsilon v) - f(x_\star) = \int_0^1 \langle \nabla f(x_\star + \epsilon tv), \epsilon v\rangle \, \mathrm{d}t = \epsilon \int_0^1 \langle \nabla f(x_\star + \epsilon tv), v\rangle \, \mathrm{d}t \geqslant 0$$

which shows that $\int_0^1 \langle \nabla f(x_\star + \epsilon tv), v\rangle \, \mathrm{d}t \geqslant 0$. By continuity of ∇f, sending $\epsilon \to 0$ proves that $\int_0^1 \langle \nabla f(x_\star), v\rangle \, \mathrm{d}t \geqslant 0$ for all $v \in \mathbb{R}^d$, which entails $\nabla f(x_\star) = 0$.

2. If f is twice continuously differentiable, we can use the Fundamental Theorem of Calculus twice, this time on the path $\gamma(s) = x_\star + st\epsilon v$, $s \in [0, 1]$, we then have

$$\int_0^1 \int_0^1 \langle \nabla^2 f(x_\star + \epsilon stv)v, v\rangle \, \mathrm{d}s \, \mathrm{d}t \geqslant 0$$

By continuity of $\nabla^2 f$, sending $\epsilon \to 0$, we have $\langle \nabla^2 f(x_\star)v, v\rangle \geqslant 0$ for all $v \in \mathbb{R}^d$, which shows that ∇f is positive semidefinite.

□

These are necessary conditions, but not sufficient. An easy counterexample is $f(x) = x^3$. The issue is that the proof of Lemma 1.3.5 is entirely local, so the same conclusion holds even if $x_\star$ is only assumed to be a local minimizer. On the other hand, under the assumption of convexity, the first-order necessary condition becomes sufficient.

Lemma 1.3.6: Sufficient Condition for Optimality

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and continuously differentiable, and let $\nabla f(x_\star) = 0$. Then, $x_\star$ is a global minimizer of f. In particular, every local minimizer of f is a global minimizer.

Proof. By equation 1.4, set $\alpha = 0$ and $x = x_\star$, we have

$$f(y) \geqslant f(x_\star) + \langle \nabla f(x_\star), y - x_\star\rangle = f(x_\star), \qquad \text{for all } y \in \mathbb{R}^d$$

which shows that $x_\star$ is a global minimizer. □

The minimizer is unique if f is strictly convex.

Definition 1.3.7: Strictly Convex

$f : \mathbb{R}^d \to \mathbb{R}$ is **strictly convex** if for all distinct $x, y \in \mathbb{R}^d$ and $t \in (0, 1)$, we have

$$f((1-t)x + ty) < (1-t)f(x) + tf(y)$$

Lemma 1.3.8: Uniqueness of Minimizer

Let $f : \mathbb{R}^d \to \mathbb{R}$ be strictly convex. Then, if f admits a minimizer $x_\star$, it is unique.

Proof. If we have two distinct minimizers $x_\star$, $\tilde{x_\star}$, so that $f(x_\star) = f(\tilde{x_\star})$. Then, strict convexity would imply

$$f\left(\frac{1}{2}x_\star + \frac{1}{2}\tilde{x_\star}\right) < \frac{f(x_\star) + f(\tilde{x_\star})}{2} = f(x_\star)$$

which is a contradiction since $f(x_\star)$ is the minimal value. □

We see that strongly convex implies strictly convex.

Lemma 1.3.9: Strong Convexity implies Strict Convexity

If f is strongly convex, then it is strictly convex.

Proof. If f is α-convex with $\alpha > 0$, then for distinct $x, y \in \mathbb{R}^d$, and $t \in (0, 1)$, we have

$$f((1-t)x + ty) \leqslant (1-t)f(x) + tf(y) - \frac{\alpha}{2}t(1-t)\|y - x\|^2 < (1-t)f(x) + tf(y)$$

□

Finally, we see that by Equation 1.4, f grows at least quadratically at ∞, which implies that it has bounded level sets. We have the following corollary:

Corollary 1.3.10: Existance and Uniqueness of Minimizer

Let $f : \mathbb{R}^d \to \mathbb{R}$ be strongly convex and continuously differentiable. Then, it admits a unique minimizer $x_\star$, which is characterized by $\nabla f(x_\star) = 0$.

1.4 The Role of Smoothness

When discussing algorithms, we also need a dual condition, an upper bound on the Hessian, which is called *smoothness*.

Definition 1.4.1: Smoothness

Let $\beta \geqslant 0$. We say that $f : \mathbb{R}^d \to \mathbb{R}$ is **β-smooth** if it is continuously differentiable and

$$f(y) \leqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{\beta}{2}\|y - x\|^2, \quad \text{for all } x, y \in \mathbb{R}^d \tag{1.9}$$

The same equivalent way of saying smoothness can be derived as what we have shown for convexity.

Proposition 1.4.2: Equivalent Formulation of Smoothness

Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable and $\beta \geqslant 0$. Then, f is β-smooth if and only if

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \leqslant \beta \|y - x\|^2, \quad \text{for all } x, y \in \mathbb{R}^d \tag{1.10}$$

If f is twice continuously differentiable, then

$$\langle v, \nabla^2 f(x) v \rangle \leqslant \beta \|v\|^2, \quad \text{for all } v, x \in \mathbb{R}^d \tag{1.11}$$

If f is convex, β-smooth, and twice continuously differentiable, then $0 \preccurlyeq \nabla^2 f \preccurlyeq \beta I$. This implies that the gradient ∇f is β-Lipschitz:

$$\|\nabla f(y) - \nabla f(x)\| \leqslant \beta \|y - x\|, \quad \text{for all } x, y \in \mathbb{R}^d \tag{1.12}$$

This is because: with Fundamental Theorem of Calculus,

$$\nabla f(y) - \nabla f(x) = \int_0^1 \nabla^2 f(x + t(y - x))(y - x) \, \mathrm{d}t$$

Take norm on both sides, and use triangular inequality,

$$\|\nabla f(y) - \nabla f(x)\| \leqslant \int_0^1 \|\nabla^2 f(x + t(y - x))\|_{\mathrm{op}} \|y - x\| \, \mathrm{d}t \leqslant \int_0^1 \beta \|y - x\| \, \mathrm{d}t = \beta \|y - x\|$$

This β-Lipschitz gradient shows that, the smoothness condition provides that the gradient does not change 'too fast'. This remains true even without assuming twice differentiability. We will see this later in Section 2.2.

Chapter 2

Gradient Method

2.1 Continuous Case: Gradient Flow

We first study the continuous time version of gradient descent via gradient flow. We let $(x_t)_{t\geqslant 0}$ denote the gradient flow for f:

$$\dot{x}_t = -\nabla f(x_t) \tag{GF}$$

This is an ODE. For our purpose, suppose f is twice continuously differentiable and do not worry about showing that (GF) is well-posed. We use the notation

$$x_\star = \mathrm{argmin} f, \quad f_\star = \min f = f(x_\star)$$

Generally, we always assume that f admits a minimizer.

It is continuous version of gradient descent since it always decreases the function value.

Lemma 2.1.1: Descent Property of GF

For any $f : \mathbb{R}^d \to \mathbb{R}$, the gradient flow $(x_t)_{t\geqslant 0}$ of f satisfies

$$\partial_t f(x_t) = -\|\nabla f(x_t)\|^2 \leqslant 0$$

Proof. By chain rule, $\partial_t f(x_t) = \langle \nabla f(x_t), \dot{x}_t \rangle = -\langle \nabla f(x_t), \nabla f(x_t) \rangle = -\|\nabla f(x_t)\|^2 \leqslant 0$. □

Now, we assume convexity. We show that under strong convexity (this is needed because it attains unique minimizer), the gradient flow contracts. Before doing this, we need a lemma. We will encounter a *differential inequality* (an inequality which holds between a quantity and its derivative(s)) in our proof of contraction. This is a common strategy for analyzing ODEs/PDEs, and it can be loosely viewed as the continuous-time analogue of induction. The **Grönwall's Ineuqality** is useful for handling such inequalities.

Lemma 2.1.2: Grönwall's Inequality

Suppose that $u : [0,T] \to \mathbb{R}$ is a continuously differentiable curve that satisfies the differential inequality

$$\dot{u}(t) \leqslant Au(t) + B(t), \quad t \in [0,T]$$

Then, it holds that

$$u(t) \leqslant u(0)\exp\left(At\right) + \int_0^t B(s)\exp(A(t-s))\,\mathrm{d}s, \quad t \in [0,T]$$

Proof. Differentiating $t \mapsto \exp(-At)u(t)$,

$$\partial_t[\exp(-At)u(t)] = -A\exp(-At)u(t) + \exp(-At)\dot{u}(t) = \exp(-At)[-Au(t) + \dot{u}(t)] \leqslant \exp(-At)B(t)$$

By Fundamental Theorem of Calculus, integrate both sides,

$$\exp(-At)u(t) - u(0) \leqslant \int_0^t B(s)\exp(-As)\,\mathrm{d}s$$

Move $u(0)$ to RHS, and multiply by $\exp(At)$ both sides, we have

$$u(t) \leqslant u(0)\exp(At) + \int_0^t B(s)\exp(A(t-s))\,\mathrm{d}s$$

which is the form we desired. □

Now we can see the contraction result.

Theorem 2.1.3: Contraction of GF

Let $f : \mathbb{R}^d \to \mathbb{R}$ be α-convex. Let $(y_t)_{t\geqslant 0}$ be another gradient flow for f, i.e., $\dot{y}_t = -\nabla f(y_t)$ (e.g., different starting point). Then, for all $t \geqslant 0$,

$$\|y_t - x_t\| \leqslant \exp(-\alpha t)\|y_0 - x_0\|$$

Particularly, for $y_t = x_\star$, the flow conveges to the minimizer

$$\|x_t - x_\star\| \leqslant \exp(-\alpha t)\|x_0 - x_\star\|$$

Proof. We differentiate the squared distance between the two flows,

$$\partial_t(\|y_t - x_t\|^2) = 2\langle y_t - x_t, \dot{y}_t - \dot{x}_t\rangle = -2\langle y_t - x_t, \nabla f(y_t) - \nabla f(x_t)\rangle \leqslant -2\alpha\|y_t - x_t\|^2$$

where the last inequality holds from α-convexity (Equation 4.8). Apply Grönwall's inequality with $u(t) = \|y_t - x_t\|^2$, $A = -2\alpha$ and $B = 0$, we have

$$\|y_t - x_t\|^2 \leqslant \|y_0 - x_0\|^2\exp(-2\alpha t) \quad \Longrightarrow \quad \|y_t - x_t\| \leqslant \exp(-\alpha t)\|y_0 - x_0\|$$

Particularly, we can take $y_t = x_\star$ for all $t \geqslant 0$ since at this point $\dot{y}_t = -\nabla f(y_t) = 0$ by Lemma 1.3.5. □

The next result is about convergence in function value, and unlike the previous theorem, it yields convergence for the case $\alpha = 0$ as well.

Theorem 2.1.4: Convergence of GF in Function Value

Let $f : \mathbb{R}^d \to \mathbb{R}$ be α-convex, $\alpha \geqslant 0$. Then, for all $t \geqslant 0$,

$$f(x_t) - f_\star \leqslant \frac{\alpha}{2(\exp(\alpha t) - 1)} \|x_0 - x_\star\|^2$$

When $\alpha = 0$, the RHS should be interpreted as limiting value as $\alpha \to 0$, which is $\frac{1}{2t}\|x_0 - x_\star\|^2$.

Proof. We differentiate $t \mapsto \|x_t - x_\star\|^2$. This time we use Equation 1.4,

$$\begin{aligned} \partial_t(\|x_t - x_\star\|^2) &= 2\langle x_t - x_\star, \dot{x}_t\rangle = 2\langle \nabla f(x_t), x_\star - x_t\rangle \\ &\leqslant -\alpha\|x_t - x_\star\|^2 - 2(f(x_t) - f_\star) \qquad \text{(Equation 1.4 with } y = x_\star \text{ and } x = x_t) \end{aligned}$$

Use Grönwall's inequality with $u(t) = \|x_t - x_\star\|^2$, $A = -\alpha$ and $B(t) = -2(f(x_t) - f_\star)$, we have

$$0 \leqslant \|x_t - x_\star\|^2 \leqslant \exp(-\alpha t)\|x_0 - x_\star\|^2 - 2\int_0^t \exp(-\alpha(t-s))(f(x_s) - f_\star)\,\mathrm{d}s \tag{2.1}$$

By the descent property Lemma 2.1.1, we have $f(x_s) \geqslant f(x_t)$ for all $s \leqslant t$, thus,

$$\begin{aligned} \int_0^1 \exp(-\alpha(t-s))(f(x_s) - f_\star)\,\mathrm{d}s &\geqslant (f(x_t) - f_\star)\int_0^t \exp(-\alpha(t-s))\,\mathrm{d}s \\ &= (f(x_t) - f_\star)\frac{1 - \exp(-\alpha t)}{\alpha} \end{aligned}$$

Substitute this into the previous equation 2.1, we have

$$0 \leqslant \exp(-\alpha t)\|x_0 - x_\star\|^2 - 2\int_0^t \exp(-\alpha(t-s))(f(x_s) - f_\star)\,\mathrm{d}s \leqslant \exp(-\alpha t)\|x_0 - x_\star\|^2 - 2(f(x_t) - f_\star)\frac{1 - \exp(-\alpha t)}{\alpha}$$

Rearranging this result, we have

$$\begin{aligned} & 0 \leqslant \|x_0 - x_\star\|^2 - 2(f(x_t) - f_\star)\frac{\exp(\alpha t) - 1}{\alpha} \\ \Longrightarrow \quad & f(x_t) - f_\star \leqslant \frac{\alpha}{2(\exp(\alpha t) - 1)}\|x_0 - x_\star\|^2 \end{aligned}$$

which completes the proof. For the case $\alpha = 0$, we go back to Equation 2.1, and see that then it becomes

$$0 \leqslant \|x_0 - x_\star\|^2 - 2\int_0^t (f(x_s) - f_\star)\,\mathrm{d}s \leqslant \|x_0 - x_\star\|^2 - 2\int_0^t (f(x_t) - f_\star)\,\mathrm{d}s = \|x_0 - x_\star\|^2 - 2t(f(x_t) - f_\star)$$

After rearranging, we see that $f(x_t) - f_\star \leqslant \frac{1}{2t}\|x_0 - x_\star\|^2$. □

When $\alpha > 0$, Theorem 2.1.4 shows that $f(x_t) - f_\star = O(\exp(-\alpha t))$. When $\alpha = 0$, the rate becomes $f(x_t) - f_\star = O(1/t)$. Actually, the rates in Theorem 2.1.4 are not sharp.

Theorem 2.1.5: Sharp Bound of GF in Function Value

Let $f : \mathbb{R}^d \to \mathbb{R}$ be α-convex, $\alpha \geqslant 0$. Then, for all $t \geqslant 0$,

$$f(x_t) - f_\star \leqslant \frac{\alpha}{2(\exp(2\alpha t) - 1)} \|x_0 - x_\star\|^2$$

When $\alpha = 0$, the RHS should be interpreted as limiting value as $\alpha \to 0$, which is $\frac{1}{4t}\|x_0 - x_\star\|^2$.

Proof. This proof is taken from Liang, Mitra and Wibisono (2024)[7].

STEP I: First consider the case $\alpha = 0$. We consider the function below.

$$\mathcal{L}_t = t^2\|\nabla f(x_t)\|^2 + 2t(f(x_t) - f_\star) + \|x_t - x_\star\|^2$$

Take derivative w.r.t. t, we have

$$\begin{aligned}
\dot{\mathcal{L}}_t &= \underbrace{2t\|\nabla f(x_t)\|^2 + 2t^2\left\langle \nabla f(x_t), \partial_t(\nabla f(x_t))\right\rangle}_{\text{Term 1}} + \underbrace{2(f(x_t) - f_\star) + 2t\langle \nabla f(x_t), \dot{x}_t\rangle}_{\text{Term 2}} + \underbrace{2\langle x_t - x_\star, \dot{x}_t\rangle}_{\text{Term 3}} \\
&= 2t\|\nabla f(x_t)\|^2 + 2t^2\langle \nabla f(x_t), \nabla^2 f(x_t)\cdot \dot{x}_t\rangle + 2(f(x_t) - f_\star) - 2t\|\nabla f(x_t)\|^2 + 2\langle x_\star - x_t, \nabla f(x_t)\rangle \\
&= -2t^2\langle \nabla f(x_t), \nabla^2 f(x_t)\cdot \nabla f(x_t)\rangle + 2(f(x_t) - f_\star) + 2\langle x_\star - x_t, \nabla f(x_t)\rangle \\
&\leqslant -2t^2\langle \nabla f(x_t), \nabla^2 f(x_t)\cdot \nabla f(x_t)\rangle + 2(f(x_t) - f_\star) + 2(f_\star - f(x_t)) && \text{(By convexity)} \\
&= -2t^2\langle \nabla f(x_t), \nabla^2 f(x_t)\cdot \nabla f(x_t)\rangle \leqslant 0 && \text{(By Equation (1.6))}
\end{aligned}$$

where we use the convexity criterion $f(y) \geqslant f(x) + \langle \nabla f(x), y - x\rangle$ with $y = x_\star$ and $x = x_t$. Since $\mathcal{L}_t$ is decreasing, we have $\mathcal{L}_0 \geqslant \mathcal{L}_t$. Then,

$$\begin{aligned}
&\|x_0 - x_\star\|^2 \geqslant t^2\|\nabla f(x_t)\|^2 + 2t(f(x_t) - f_\star) + \|x_t - x_\star\|^2 \\
\implies\ & t^2\|\nabla f(x_t)\|^2 \leqslant \|x_0 - x_\star\|^2 - 2t(f(x_t) - f_\star) - \|x_t - x_\star\|^2 \\
\implies\ & t^2\|\nabla f(x_t)\|^2 \leqslant \|x_0 - x_\star\|^2 && \text{(Since } f(x_t) - f_\star \geqslant 0, \|\cdot\|^2 \geqslant 0\text{)} \\
\implies\ & \|\nabla f(x_t)\|^2 \leqslant \frac{1}{t^2}\|x_0 - x_\star\|^2
\end{aligned}$$

Using Equation 1.4 with $y = x_\star$ and $x = x_t$, we have

$$f(x_\star) - f(x_t) \geqslant \langle \nabla f(x_t), x_\star - x_t\rangle \quad \implies \quad 2t(f(x_t) - f_\star) \leqslant 2t\langle \nabla f(x_t), x_t - x_\star\rangle$$

Using Cauchy-Schwartz inequality and $ab \leqslant \frac{a^2}{2} + \frac{b^2}{2}$, we have

$$\begin{aligned}
2t(f(x_t) - f_\star) &\leqslant 2t\langle \nabla f(x_t), x_t - x_\star\rangle \\
&\leqslant 2t\|\nabla f(x_t)\|\|x_t - x_\star\| && \text{(Cauchy-Schwartz)} \\
&\leqslant t^2\|\nabla f(x_t)\|^2 + \|x_t - x_\star\|^2 && (ab \leqslant \tfrac{a^2}{2} + \tfrac{b^2}{2},\ a = t\|\nabla f(x_t)\|, b = \|x_t - x_\star\|)
\end{aligned}$$

Then, by $\mathcal{L}_0 \geqslant \mathcal{L}_t$, we have

$$4t(f(x_t) - f_\star) \leqslant t^2\|\nabla f(x_t)\|^2 + 2t(f(x_t) - f_\star) + \|x_t - x_\star\|^2 = \mathcal{L}_t \leqslant \mathcal{L}_0 = \|x_0 - x_\star\|^2$$

This shows the upper bound

$$f(x_t) - f_\star \leqslant \frac{1}{4t}\|x_0 - x_\star\|^2$$

To prove the bound is sharp, use the function $f(x) = \frac{R}{2t}\max\{0, x\}$ with $x_0 = R$. In this case, $\dot{x}_s = -\frac{R}{2t}$, thus $x_s = R - \frac{R}{2t}s$, and it hits the stationary point $x = 0$ at $s = 2t$. Therefore, $f_\star = x_\star = 0$. When $s = t$, we have

$$f(x_t) - f_\star = f\left(-\frac{R}{2} + R\right) = \frac{R^2}{4t} = \frac{1}{4t}\|x_0 - x_\star\|^2$$

which attains the equality.

STEP II: Now consider $\alpha > 0$. Consider more general function

$$\mathcal{L}_t = A_t\|\nabla f(x_t)\|^2 + 2B_t(f(x_t) - f_\star) + \|x_t - x_\star\|^2$$

We want $\dot{\mathcal{L}}_t \leqslant -\alpha\mathcal{L}_t$. Take derivative, we have

$$\begin{aligned}
\dot{\mathcal{L}}_t &= \dot{A}_t\|\nabla f(x_t)\|^2 + 2A_t\langle\nabla f(x_t), \nabla^2 f(x_t)\cdot\dot{x}_t\rangle + 2\dot{B}_t(f(x_t) - f_\star) + 2B_t\langle\nabla f(x_t), \dot{x}_t\rangle + 2\langle x_t - x_\star, \dot{x}_t\rangle \\
&= \dot{A}_t\|\nabla f(x_t)\|^2 - 2A_t\langle\nabla f(x_t), \nabla^2 f(x_t)\cdot\nabla f(x_t)\rangle + 2\dot{B}_t(f(x_t) - f_\star) - 2B_t\|\nabla f(x_t)\|^2 + 2\langle x_\star - x_t, \nabla f(x_t)\rangle \\
&\leqslant \dot{A}_t\|\nabla f(x_t)\|^2 - 2\alpha A_t\|\nabla f(x_t)\|^2 + 2\dot{B}_t(f(x_t) - f_\star) - 2B_t\|\nabla f(x_t)\|^2 + 2\langle x_\star - x_t, \nabla f(x_t)\rangle && \text{(Equation (1.6))} \\
&\leqslant \left(\dot{A}_t - 2\alpha A_t - 2B_t\right)\|\nabla f(x_t)\|^2 + 2\dot{B}_t(f(x_t) - f_\star) + 2(f_\star - f(x_t)) - \alpha\|x_\star - x_t\|^2 && \text{(Equation (1.4))} \\
&= \left(\dot{A}_t - 2\alpha A_t - 2B_t\right)\|\nabla f(x_t)\|^2 + \left(2\dot{B}_t - 2\right)(f(x_t) - f_\star) - \alpha\|x_\star - x_t\|^2 \\
&= -\alpha\mathcal{L}_t
\end{aligned}$$

Here, we have one-to-one correspondance:

$$\dot{A}_t - 2\alpha A_t - 2B_t = -\alpha A_t \tag{2.2}$$

$$2\dot{B}_t - 2 = -2\alpha B_t \tag{2.3}$$

We first deal with Equation 2.3. This can be rearranging as

$$\dot{B}_t + \alpha B_t = 1$$

Multiplying both side by the integration factor $\exp(\int \alpha\,\mathrm{d}t) = \exp(\alpha t)$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\exp(\alpha t)B_t\right) = \exp(\alpha t)$$

Integrate both sides, we have

$$\exp(\alpha t)B_t = \frac{1}{\alpha}\exp(\alpha t) + C \quad \Longrightarrow \quad B_t = \frac{1}{\alpha} + C\exp(-\alpha t)$$

Analogous to the case $\alpha = 0$, we want $B_0 = 0$, which yields $C = -\frac{1}{\alpha}$. The final solution for B_t is thus

$$B_t = \frac{1 - \exp(-\alpha t)}{\alpha}$$

Now we deal with Equation 2.2. Substituting B_t into it, we have

$$\dot{A}_t - \alpha A_t = \frac{2(1 - \exp(-\alpha t))}{\alpha}$$

Similarly, multiply by the integration factor $\exp(\int -\alpha \, \mathrm{d}t) = \exp(-\alpha t)$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\exp(-\alpha t)A_t\right) = \frac{2(\exp(-\alpha t) - \exp(-2\alpha t))}{\alpha} \quad \Longrightarrow \quad \exp(-\alpha t)A_t = \frac{\exp(-2\alpha t) - 2\exp(-\alpha t)}{\alpha^2} + C$$

Set $A_0 = 0$, we have $C = \frac{1}{\alpha^2}$. Therefore, the solution is

$$A_t = \frac{\exp(\alpha t) - 2 + \exp(-\alpha t)}{\alpha^2}$$

Using Grönwall's inequality with $u(t) = \mathcal{L}_t$, $A = -\alpha$ and $B = 0$, we have

$$\mathcal{L}_t \leqslant \mathcal{L}_0 \exp(-\alpha t)$$

which is equivalent to

$$A_t\|\nabla f(x_t)\|^2 \leqslant A_t\|\nabla f(x_t)\|^2 + 2B_t(f(x_t) - f_\star) + \|x_t - x_\star\|^2 = \mathcal{L}_t \leqslant \mathcal{L}_0\exp(-\alpha t) = \exp(-\alpha t)\|x_0 - x_\star\|^2$$

Therefore, divide both sides by A_t, we have

$$\|\nabla f(x_t)\|^2 \leqslant \frac{\exp(-\alpha t)}{A_t}\|x_0 - x_\star\|^2 = \frac{\alpha^2}{\exp(2\alpha t)(1 - \exp(-\alpha t))^2}\|x_0 - x_\star\|^2$$

Use the convexity condition (1.4) with $y = x_\star$ and $x = x_t$, we have

$$\begin{aligned}
& f_\star - f(x_t) \geqslant \langle \nabla f(x_t), x_\star - x_t\rangle + \frac{\alpha}{2}\|x_t - x_\star\|^2 \\
\Longrightarrow \quad & f(x_t) - f_\star + \frac{\alpha}{2}\|x_t - x_\star\|^2 \leqslant \langle \nabla f(x_t), x_t - x_\star\rangle \\
& \leqslant \|\nabla f(x_t)\|\|x_t - x_\star\| \qquad (3)
\end{aligned}$$

where the last line follows from Cauchy-Schwartz. Using the fact $a^2 + b^2 \geqslant 2ab$ and the above relation, for any $C_t \geqslant 0$, we have

$$\begin{aligned}
A_t\|\nabla f(x_t)\|^2 + C_t\|x_t - x_\star\|^2 &\geqslant 2\sqrt{A_t C_t}\|\nabla f(x_t)\|\|x_t - x_\star\| \\
&\geqslant 2\sqrt{A_t C_t}\left((f(x_t) - f_\star) + \frac{\alpha}{2}\|x_t - x_\star\|^2\right) \qquad \text{(Equation (3))}
\end{aligned}$$

We can then decompose $\mathcal{L}_t$ as

$$\begin{aligned}\mathcal{L}_t &= A_t\|\nabla f(x_t)\|^2 + C_t\|x_t - x_\star\|^2 + 2B_t(f(x_t) - f_\star) + (1 - C_t)\|x_t - x_\star\|^2 \\ &\geqslant 2\sqrt{A_tC_t}\left((f(x_t) - f_\star) + \frac{\alpha}{2}\|x_t - x_\star\|^2\right) + 2B_t(f(x_t) - f_\star) + (1 - C_t)\|x_t - x_\star\|^2 \\ &= 2\left(\sqrt{A_tC_t} + B_t\right)(f(x_t) - f_\star) + \left(\alpha\sqrt{A_tC_t} + 1 - C_t\right)\|x_t - x_\star\|^2\end{aligned}$$

Therefore, for our purpose as what we did in the case $\alpha = 0$, which is to construct an equation such that $K_t(f(x_t) - f_\star) \leqslant \mathcal{L}_t$, we choose C_t such that $\alpha\sqrt{A_tC_t} + 1 - C_t = 0$. Choosing the positive solution, we have

$$\sqrt{C_t} = \frac{\alpha\sqrt{A_t} + \sqrt{\alpha^2 A_t + 4}}{2}$$

We have previously seen that $\dot{A}_t = \alpha A_t + 2B_t = \frac{1}{\alpha}(\exp(\alpha t) - \exp(-\alpha t))$. Moreover,

$$A_t(\alpha^2 A_t + 4) = \frac{\exp(\alpha t) - 2 + \exp(-\alpha t)}{\alpha^2}(\exp(\alpha t) + 2 + \exp(-\alpha t)) = \frac{(\exp(\alpha t) - \exp(-\alpha t))^2}{\alpha^2}$$

We have that

$$\begin{aligned}2\left(B_t + \sqrt{A_tC_t}\right) &= 2B_t + \alpha A_t + \sqrt{A_t(\alpha^2 A_t + 4)} \\ &= \frac{\exp(\alpha t) - \exp(-\alpha t)}{\alpha} + \frac{\exp(\alpha t) - \exp(-\alpha t)}{\alpha} \\ &= \frac{2}{\alpha}(\exp(\alpha t) - \exp(-\alpha t))\end{aligned}$$

Therefore, with this choice of C_t, we have

$$\mathcal{L}_t \geqslant \frac{2}{\alpha}(\exp(\alpha t) - \exp(-\alpha t))(f(x_t) - f_\star)$$

Using the previous conclusion from Grönwall's Inequality, i.e., $\mathcal{L}_t \leqslant \mathcal{L}_0 \exp(-\alpha t)$, we have

$$\begin{aligned}&\frac{2(\exp(\alpha t) - \exp(-\alpha t))}{\alpha}(f(x_t) - f_\star) \leqslant \mathcal{L}_t \leqslant \mathcal{L}_0 \exp(-\alpha t) = \exp(-\alpha t)\|x_0 - x_\star\|^2 \\ \implies\ & f(x_t) - f_\star \leqslant \frac{\alpha}{2(\exp(2\alpha t) - 1)}\|x_0 - x_\star\|^2\end{aligned}$$

which concludes the proof. □

Next, we observe that convexity is not needed for convergence in function value. Due to the descent property (Lemma 2.1.1), it is enough to have a lower bound on the norm of the gradient to ensure that we make sufficient progress.

Definition 2.1.6: Polyak-Łojasiewicz (PŁ) Inequality

Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable and $\alpha > 0$. We say that f satisfies a **Polyak-Łojasiewicz (PŁ) inequality** with constant α if

$$\|\nabla f(x)\|^2 \geqslant 2\alpha(f(x) - f(x_\star)), \qquad \text{for all } x \in \mathbb{R}^d$$

The next statement is an immediate corollary of Lemma 2.1.1, and Grönwall's Inequality.

Corollary 2.1.7: Convergence of (GF) under PŁ inequality

Let $f : \mathbb{R}^d \to \mathbb{R}$ satisfy PŁ inequality with constant $\alpha > 0$.Then, for all $t \geqslant 0$,

$$f(x_t) - f_\star \leqslant (f(x_0) - f_\star) \exp(-2\alpha t)$$

Proof. By descent proof and PŁ inequality, we have

$$\frac{\mathrm{d}}{\mathrm{d}t} f(x_t) = \langle \nabla f(x_t), \dot{x}_t \rangle = -\|\nabla f(x_t)\|^2 \leqslant -2\alpha(f(x_t) - f(x_\star)) \quad \Longrightarrow \quad \frac{\mathrm{d}}{\mathrm{d}t}(f(x_t) - f(x_\star)) \leqslant -2\alpha(f(x_t) - f(x_\star))$$

Therefore, we can use Grönwall's inequality with $A = -2\alpha$ and $B = 0$ to see that

$$f(x_t) - f(x_\star) \leqslant (f(x_0) - f(x_\star)) \exp(-2\alpha t)$$

which completes the proof. □

Next we show the key property of PŁ inequality.

Proposition 2.1.8: Strong Convexity ⇒ PŁ ⇒ Quadratic Growth

Let $f : \mathbb{R}^d \to \mathbb{R}$ and $\alpha > 0$. The following implications hold.

1. If f is α-convex, then f satisfies PŁ inequality with constant α.
2. If f satisfies PŁ inequality with constant α, then it satisfies the following *quadratic growth* property:

$$f(x) - f_\star \geqslant \frac{\alpha}{2} \inf_{x_\star \in \mathcal{X}_\star} \|x - x_\star\|^2, \qquad \text{for all } x \in \mathbb{R}^d$$

where $\mathcal{X}_\star$ denotes the set of minimizers of f.

Proof. 1. Set $y = x_\star$ in Equation 1.4, we have

$$\begin{aligned} -(f(x) - f_\star) &\geqslant \langle \nabla f(x), x_\star - x \rangle + \frac{\alpha}{2}\|x - x_\star\|^2 \\ &\geqslant -\|\nabla f(x)\|\|x_\star - x\| + \frac{\alpha}{2}\|x - x_\star\|^2 && \text{(Cauchy-Schwartz)} \\ &\geqslant -\frac{1}{2\alpha}\|\nabla f(x)\|^2 \end{aligned}$$

where the last step uses $ab \leqslant \frac{\lambda}{2}a^2 + \frac{1}{2\lambda}b^2$ for all $\lambda > 0$. Here we choose $\alpha = \|x - x_\star\|$, $b = \|\nabla f(x)\|$, and $\lambda = \alpha$.

2. Let $(x_t)_{t \geqslant 0}$ denote the gradient flow for f started at $x_0 = x$. For simplicity, we assume gradient flow converges to a point $x_\star$. This assumption can be avoided, as proved in Karimi, Nutini and Schmidt (2016)[5] Appendix A.

By Corollary 2.1.7, since the function value converges under PŁ, and the gradient flow converges to $x_\star$, we have $x_\star \in \mathcal{X}_\star$. First observe that

$$\partial_t(\|x_t - x_0\|^2) = 2\|x_t - x_0\|\partial_t(\|x_t - x_0\|) = 2\langle x_t - x_0, \dot{x}_t\rangle = -2\langle x_t - x_0, \nabla f(x_t)\rangle \leqslant 2\|\nabla f(x_t)\|\|x_t - x_0\|$$

where the last inequality follows by Cauchy-Schwarz. This indicates that

$$\partial_t\|x_t - x_0\| \leqslant \|\nabla f(x_t)\|$$

Differentiate the following quantity: $\mathcal{L}_t = \sqrt{\frac{\alpha}{2}}\|x_t - x_0\| + \sqrt{f(x_t) - f_\star}$, using the inequality above, we have

$$\dot{\mathcal{L}}_t \leqslant \sqrt{\frac{\alpha}{2}}\|\nabla f(x_t)\| + \frac{\partial_t(f(x_t) - f_\star)}{2\sqrt{f(x_t) - f_\star}} = \sqrt{\frac{\alpha}{2}}\|\nabla f(x_t)\| + \frac{\langle \nabla f(x_t), \dot{x}_t\rangle}{2\sqrt{f(x_t) - f_\star}} = \sqrt{\frac{\alpha}{2}}\|\nabla f(x_t)\| - \frac{\|\nabla f(x_t)\|^2}{2\sqrt{f(x_t) - f_\star}}$$

Using PŁ, we have $\|\nabla f(x_t)\|^2 \geqslant 2\alpha(f(x_t) - f_\star)$, so

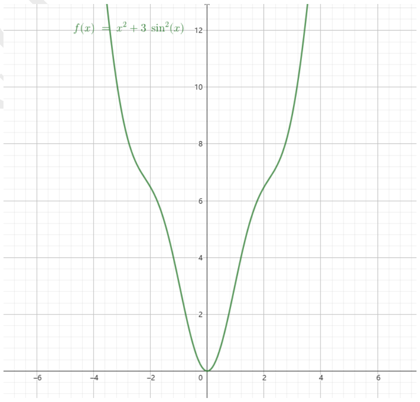
$$\dot{\mathcal{L}}_t \leqslant \|\nabla f(x_t)\|\left(\sqrt{\frac{\alpha}{2}} - \frac{\|\nabla f(x_t)\|}{2\sqrt{f(x_t) - f_\star}}\right) \leqslant \|\nabla f(x_t)\|\left(\sqrt{\frac{\alpha}{2}} - \frac{\sqrt{2\alpha}\sqrt{f(x_t) - f_\star}}{2\sqrt{f(x_t) - f_\star}}\right) = 0$$

Thus, $\mathcal{L}_t$ is decreasing. Since $\mathcal{L}_0 = \sqrt{f(x_0) - f_\star}$ and $\mathcal{L}_\infty = \sqrt{\frac{\alpha}{2}}\|x_0 - x_\star\|$, we have $\mathcal{L}_0 \geqslant \mathcal{L}_\infty$, which concludes our result since x_0 is arbitrary.

□

Remark: *PŁ is truly a weaker condition, PŁ does not implies convexity.* A simple example is

$$f(x) = x^2 + 3\sin^2 x$$



It is obviously not convex from the graph, but its gradient goes quickly and we can gradually reach the optimal point.

Finally, we did not assume convexity, and have the result below.

Theorem 2.1.9: Stationary Point Nonconvex Case

For any $f : \mathbb{R}^d \to \mathbb{R}$,

$$\min_{s\in[0,t]} \|\nabla f(x_s)\| \leqslant \sqrt{\frac{f(x_0) - f_\star}{t}}$$

Proof. Descent lemma 2.1.1 shows that $\partial_t f(x_t) = -\|\nabla f(x_t)\|^2$, so by Fundamental theorem of calculus,

$$\min_{s\in[0,t]} \|\nabla f(x_s)\|^2 \leqslant \frac{1}{t}\int_0^t \|\nabla f(x_s)\|^2 \,\mathrm{d}s = \frac{f(x_0) - f(x_t)}{t} \leqslant \frac{f(x_0) - f_\star}{t}$$

where the result follows by taking the square root. □

Remark: This implies that there exists a sequence of times $\{t_n\}_{n\in\mathbb{N}} \nearrow \infty$ such that $\|\nabla f(x_{t_n})\| \to 0$. Indeed, from the theorem above, $\min_{s\in[n,2n]} \|\nabla f(x_s)\| = O(1/n^{1/2})$, so we can choose $t_n \in [n, 2n]$. However, **the gradient flow may not converge** (this can be true indeed, for example, with a $1/n$ rate of decrease). Famouly, it is a result of Łojasiewicz (1963)[8] that for *real analytic* f, if the gradient flow remains bounded, then it does converge, and hence necessarily to a stationary point. Such stationary point, however, may not be a globla minimizer.

2.2 Discrete Case: Gradient Descent

The **gradient descent algorithm** is a discretization of gradient flow

$$x_{n+1} = x_n - h\nabla f(x_n) \tag{GD}$$

Throughout the section, we assume that f is twice continuously differentiable and β-smooth. We will show that the results for gradient flow tranfers to gradient descent. We use the single step notation

$$x^+ = x - h\nabla f(x)$$

Lemma 2.2.1: Descent Lemma

For any β-smooth $f : \mathbb{R}^d \to \mathbb{R}$. If $h \leqslant 1/\beta$, then

$$f(x^+) - f(x) \leqslant -\frac{h}{2}\|\nabla f(x)\|^2$$

Proof. By the smoothness inequality 1.9, we have

$$f(x^+) \leqslant f(x) + \langle \nabla f(x), x^+ - x\rangle + \frac{\beta}{2}\|x^+ - x\|^2 = f(x) - h\|\nabla f(x)\|^2 + \frac{\beta h^2}{2}\|\nabla f(x)\|^2$$

If $h \leqslant 1/\beta$, then the coefficient of $\|\nabla f(x)\|^2$ becomes $h(\beta h/2 - 1) \leqslant h(1/2 - 1) = -h/2$. The result follows. □

Remark: We continue the discussion in section 1.4 now. At the end of Section 1.4, we mentioned that if f is convex, β-smooth, and twice continuously differentiable, then the gradient ∇f is β-Lipschitz. However, *the requirement of twice* *differentiable is actually redundant.* Define the function

$$g(y) = f(y) - \langle \nabla f(x), y \rangle$$

This function is still β-smooth, since adding a linear term will not affect the smoothness. Moreover, it is minimized at x, since $\nabla g(y) = \nabla f(y) - \nabla f(x)$, where it equals to zero if $y = x$, and $\nabla g^2(y) = \nabla^2 f(y) \succcurlyeq 0$ since f is convex. If we apply descent lemma 2.2.1 onto function g with $h = 1/\beta$, we have

$$g(x) - g(y) \leqslant g(y^+) - g(y) \leqslant -\frac{1}{2\beta}\|\nabla g(y)\|^2$$

Substitute back the form of g, we have

$$f(x) - \langle \nabla f(x), x \rangle - f(y) + \langle \nabla f(x), y \rangle \leqslant -\frac{1}{2\beta}\|\nabla f(y) - \nabla f(x)\|^2 \quad \Longrightarrow \quad f(y) \geqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2\beta}\|\nabla f(y) - \nabla f(x)\|^2 \tag{2.4}$$

Change the role of x and y, we have

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

Adding the two equations,

$$\begin{aligned} & 0 \geqslant \langle \nabla f(x) - \nabla f(y), y - x \rangle + \frac{1}{\beta}\|\nabla f(y) - \nabla f(x)\|^2 \\ \Longrightarrow \quad & \|\nabla f(y) - \nabla f(x)\|^2 \leqslant \beta \langle \nabla f(y) - \nabla f(x), y - x \rangle \end{aligned} \tag{2.5}$$

Finally, by Cauchy-Schwarz inequality,

$$\|\nabla f(y) - \nabla f(x)\|^2 \leqslant \beta \langle \nabla f(y) - \nabla f(x), y - x \rangle \leqslant \beta \|\nabla f(x) - \nabla f(y)\| \|y - x\|$$

which yields $\|\nabla f(y) - \nabla f(x)\| \leqslant \beta \|y - x\|$, i.e., the gradient is β-Lipschitz. We didn't use Hessian in this proof.

We introduce condition number for further results.

Definition 2.2.2: Condition Number

Let f be α-convex and β-smooth. Then, the **condition number** of f is defined to be the ratio $\kappa = \beta/\alpha \geqslant 1$.

Remarks:

- The condition number *must* greater than 1. This is because, by equation 1.6 and 1.11, we actually have $\alpha I \preccurlyeq \nabla^2 f \preccurlyeq \beta I$, and it is impossible when $\alpha > \beta$.
- When $f = \frac{1}{2}\langle x, Ax \rangle$ with A symmetric, then α, β correspond to the minimum and maximum eigenvalues of A, respectively (which can be seen also from that $\alpha I \preccurlyeq \nabla^2 f \preccurlyeq \beta I$).

Now we recover contraction.

Theorem 2.2.3: Contraction of Gradient Descent

Let f be α-convex and β-smooth. For all $x, y \in \mathbb{R}^d$ and step size $h \leqslant 1/\beta$,

$$\|y^+ - x^+\| \leqslant (1 - \alpha h)^{1/2} \|y - x\|$$

Proof. Expand the square,

$$\|y^+ - x^+\|^2 = \|(y - h\nabla f(y)) - (x - h\nabla f(x))\|^2 = \|y - x\|^2 - 2h\langle y - x, \nabla f(y) - \nabla f(x)\rangle + h^2\|\nabla f(y) - \nabla f(x)\|^2$$

Using the equation 2.5 above,

$$\begin{aligned}
\|y^+ - x^+\|^2 &\leqslant \|y - x\|^2 - 2h\langle y - x, \nabla f(y) - \nabla f(x)\rangle + h^2\beta\langle y - x, \nabla f(y) - \nabla f(x)\rangle && \text{(Equation 2.5)} \\
&\leqslant \|y - x\|^2 - 2h\langle y - x, \nabla f(y) - \nabla f(x)\rangle + h\langle y - x, \nabla f(y) - \nabla f(x)\rangle && (h \leqslant 1/\beta) \\
&= \|y - x\|^2 - h\langle y - x, \nabla f(y) - \nabla f(x)\rangle \\
&\leqslant \|y - x\|^2 - \alpha h\|y - x\|^2 = (1 - \alpha h)\|y - x\|^2 && \text{(Equation 4.8)}
\end{aligned}$$

Taking square root on both sides, we get the desired result. □

Remark: In particular, if we take $y = x_\star$, $h = 1/\beta$, and iterate, it yields,

$$\|x_N - x_\star\| \leqslant \left(1 - \frac{\alpha}{\beta}\right)^{\frac{N}{2}} \|x_0 - x_\star\| = \left(1 - \frac{1}{\kappa}\right)^{\frac{N}{2}} \|x_0 - x_\star\| \leqslant \exp\left(-\frac{N}{2\kappa}\right) \|x_0 - x_\star\|$$

where the inequality follows from the fact that $1 - t \leqslant e^{-t}$. Thus, to obtain $\|x_N - x_\star\| \leqslant \epsilon$, it suffices to take $N \geqslant 2\kappa \log\left(\|x_0 - x_\star\|/\epsilon\right)$.

This bound can also be improved.

Theorem 2.2.4: Sharp Bound of Gradient Descent Contraction

Let f be α-convex and β-smooth. For all $x, y \in \mathbb{R}^d$ and step size $h = 2/(\alpha + \beta)$, the sharp rate is

$$\|y^+ - x^+\| \leqslant \frac{\kappa - 1}{\kappa + 1} \|y - x\|$$

Proof. Let $T := \mathrm{id} - h\nabla f$ denote the one step gradient descent mapping. By the fundamental theorem of calculus,

$$\|y^+ - x^+\| = \|T(y) - T(x)\| = \left\|\int_0^1 \nabla T((1 - t)x + ty)(y - x)\,\mathrm{d}t\right\| \leqslant \left(\int_0^1 \|\nabla T((1 - t)x + ty)\|_{\mathrm{op}}\,\mathrm{d}t\right) \|y - x\|$$

Since f is α-convex and β-smooth, we have $\alpha I \preccurlyeq \nabla f(x) \preccurlyeq \beta I$. Thus, $(1 - h\beta)I \preccurlyeq \nabla T(x) \preccurlyeq (1 - h\alpha)I$, for all $x \in \mathbb{R}^d$. Thus,

$$\|\nabla T(x)\|_{\mathrm{op}} \leqslant \max\{|1 - h\alpha|, |1 - h\beta|\}$$

This expression is minimized by setting h so that $1 - h\alpha = h\beta - 1$, which yields $h = 2/(\alpha + \beta)$, and

$$\|y^+ - x^+\| \leqslant \left(1 - \frac{2\alpha}{\alpha + \beta}\right) \|y - x\| = \frac{\kappa - 1}{\kappa + 1}\|y - x\|$$

which completes the proof. □

Remark: For large κ, the contraction factor is

$$\|x_N - x_\star\| \leqslant \left(1 - \frac{2}{\kappa + 1}\right)^N \|y - x\| \lesssim \exp\left(-\frac{2N}{\kappa}\right)$$

which improves the complexity implied by Theorem 2.2.3 by a factor of nearly 4.

Now we recover the function value convergence. Before doing this, we introduce the discrete version of Grönwall's inequality.

Lemma 2.2.5: Discrete Grönwall's Inequality

Suppose that for some $A > 0$,

$$u_{n+1} \leqslant Au_n + B_n, \qquad \text{for } n = 0, 1, \cdots, N-1$$

Then,

$$u_N \leqslant A^N u_0 + \sum_{n=1}^{N} A^{N-n} B_{n-1}$$

Proof. Multiply the given inequality by $A^{-(n+1)}$,

$$A^{-(n+1)}u_{n+1} \leqslant A^{-n}u_n + A^{-(n+1)}B_n, \quad n = 0, 1, \cdots, N-1$$

This will form a telescoping sum

$$\begin{aligned} \sum_{n=0}^{N-1} \left(A^{-(n+1)}u_{n+1} - A^{-n}u_n\right) &= A^{-1}u_1 - u_0 + A^{-2}u_2 - A^{-1}u_1 + \cdots + A^{-N}u_N - A^{-(N-1)}u_{N-1} \\ &= A^{-N}u_N - u_0 \leqslant \sum_{n=0}^{N-1} A^{-(n+1)}B_n = \sum_{n=1}^{N} A^{-n}B_{n-1} \end{aligned}$$

Multiply both sides by A^N, we have

$$u_N \leqslant A^N u_0 + \sum_{n=1}^{N} A^{N-n}B_{n-1}$$

which is the desired result. □

Theorem 2.2.6: Convergence of Gradient Descent in Function Value

Let f be α-convex and β-smooth. For any step size $h \leqslant 1/\beta$,

$$\|x^+ - x_\star\|^2 \leqslant (1-\alpha h)\|x - x_\star\|^2 - 2h(f(x^+) - f_\star)$$

Therefore,

$$f(x_N) - f_\star \leqslant \frac{\alpha}{2\left[(1-\alpha h)^{-N} - 1\right]}\|x_0 - x_\star\|^2$$

When $\alpha = 0$, the RHS should be interpreted as its limiting value as $\alpha \to 0$, i.e., $\frac{1}{2Nh}\|x_0 - x_\star\|^2$.

Proof. Expanding the square,

$$\begin{aligned}\|x^+ - x_\star\|^2 &= \|x - x_\star - h\nabla f(x)\|^2 = \|x - x_\star\|^2 - 2h\langle \nabla f(x), x - x_\star\rangle + h^2\|\nabla f(x)\|^2 \\ &\leqslant \|x - x_\star\|^2 - 2h(f(x) - f_\star) - \alpha h\|x - x_\star\|^2 + h^2\|\nabla f(x)\|^2 && \text{(Equation 1.4)} \\ &= (1-\alpha h)\|x - x_\star\|^2 - 2h(f(x) - f_\star) + h^2\|\nabla f(x)\|^2\end{aligned}$$

For $h \leqslant 1/\beta$, using the descent lemma 2.2.1, we have

$$\|x^+ - x_\star\|^2 \leqslant (1-\alpha h)\|x - x_\star\|^2 - 2h(f(x) - f_\star) - 2h(f(x^+) - f(x)) = (1-\alpha h)\|x - x_\star\|^2 - 2h(f(x^+) - f_\star)$$

which proves the first inequality. To prove the second inequality, we apply discrete Groönwall's inequality with $u_n = \|x_n - x_\star\|^2$, $A = 1 - \alpha h$, and $B_n = -2h(f(x_{n+1}) - f_\star)$, which yields

$$\begin{aligned}0 \leqslant \|x_N - x_\star\|^2 &\leqslant (1-\alpha h)^N\|x_0 - x_\star\|^2 - 2h\sum_{n=1}^{N}(1-\alpha h)^{N-n}(f(x_n) - f_\star) \\ \Longrightarrow \quad 2h\sum_{n=1}^{N}(1-\alpha h)^{N-n}(f(x_n) - f_\star) &\leqslant (1-\alpha h)^N\|x_0 - x_\star\|^2\end{aligned}$$

For $h \leqslant 1/\beta$, the descent lemma 2.2.1 implies $f(x_n) - f_\star \geqslant f(x_N) - f_\star$, so

$$f(x_N) - f_\star \leqslant f(x_n) - f_\star \leqslant \frac{\|x_0 - x_\star\|^2}{2h\sum_{n=1}^{N}(1-\alpha h)^{-n}}$$

where we use the sum formula for geometric progression to get

$$2h\sum_{n=1}^{n}(1-\alpha h)^{-n} = 2h\frac{(1-\alpha h)^{-1}(1-(1-\alpha h)^{-N})}{1-(1-\alpha h)^{-1}} = 2h\frac{(1-(1-\alpha h)^{-N})}{-\alpha h} = \frac{2((1-\alpha h)^{-N} - 1)}{\alpha}$$

Therefore, the final inequality is

$$f(x_N) - f_\star \leqslant \frac{\alpha\|x_0 - x_\star\|^2}{2((1-\alpha h)^{-N} - 1)}$$

which is our desired result. For $\alpha = 0$, the Grönwall's inequality actually yields

$$2Nh(f(x_n) - f_\star) \leqslant \|x_0 - x_\star\|^2$$

and using the descent lemma, we have

$$f(x_N) - f_\star \leqslant \frac{\|x_0 - x_\star\|^2}{2Nh}$$

which completes the proof. □

Remarks:

- The proof of the first inequality goes through even if we replace $x_\star$ by any other point $z \in \mathbb{R}^d$, i.e.,

$$\|x^+ - z\|^2 \leqslant (1 - \alpha h)\|x - z\|^2 - 2h(f(x^+) - f(z)), \quad \forall\, z \in \mathbb{R}^d \tag{2.6}$$

- For the second inequality, in particular, if we set $h = 1/\beta$, for $\alpha > 0$ it yields

$$f(x_N) - f_\star \leqslant \frac{\alpha\|x_0 - x_\star\|^2}{2((1 - 1/\kappa)^{-N} - 1)}$$

  and for $\alpha = 0$ it yields

$$f(x_N) - f_\star \leqslant \frac{\beta}{2N}\|x_0 - x_\star\|^2$$

Now we not assume convexity. We assume PŁ inequality.

Theorem 2.2.7: Convergence of Gradient Descent under PŁ

Let f be β-smooth and satisfy the PŁ inequality with constant α. Then, for all $h \leqslant 1/\beta$,

$$f(x_N) - f_\star \leqslant (1 - \alpha h)^N (f(x_0) - f_\star)$$

Proof. By descent lemma and PŁ inequality,

$$\begin{aligned} f(x^+) - f_\star = f(x) - f_\star + f(x^+) - f(x) &\leqslant f(x) - f_\star - \frac{h}{2}\|\nabla f(x)\|^2 && \text{(Descent Lemma)} \\ &\leqslant (1 - \alpha h)(f(x) - f_\star) \end{aligned}$$

which completes the proof. □

Finally, no convex, no PŁ, still we can get a stationary point.

Theorem 2.2.8: Nonconvex Case Stationary Point

Let f be β-smooth and $h \leqslant 1/\beta$. Then,

$$\min_{n=0,1,\cdots,N-1} \|\nabla f(x_n)\| \leqslant \sqrt{\frac{2(f(x_0) - f_\star)}{Nh}}$$

Proof. Telescoping the descent lemma,

$$f(x_N) - f(x_0) = \sum_{n=0}^{N-1}(f(x_{n+1}) - f(x_n)) \leqslant -\frac{h}{2}\sum_{n=0}^{N-1}\|\nabla f(x_n)\|^2$$

$$\implies \min_{n=0,1,\cdots,N-1} \|\nabla f(x_n)\|^2 \leqslant \frac{1}{N}\sum_{n=0}^{N-1} \|\nabla f(x_n)\|^2 \leqslant \frac{2(f(x_0)-f(x_N))}{Nh} \leqslant \frac{2(f(x_0)-f_\star)}{Nh}$$

Take square root, we get the desired result. □

Below is a summary table.

Assumptions	**Criterion**	**Iterations**
α-convex, β-smooth	$\|x_N - x_\star\| \leqslant \epsilon$	$O(\kappa \log(R/\epsilon))$
α-convex, β-smooth	$f(x_N) - f_\star \leqslant \epsilon$	$O(\kappa \log \alpha R^2/\epsilon))$
convex, β-smooth	$f(x_N) - f_\star \leqslant \epsilon$	$O(\beta R^2/\epsilon)$
α-PŁ, β-smooth	$f(x_N) - f_\star \leqslant \epsilon$	$O(\kappa \log(\Delta_0/\epsilon))$
β-smooth	$\min_{n=0,1,\cdots,N-1} \|\nabla f(x_n)\| \leqslant \epsilon$	$O(\beta \Delta_0/\epsilon^2)$

Table 2.1: Rates for step size $h = 1/\beta$. $R = \|x_0 - x_\star\|$, $\Delta_0 = f(x_0) - f_\star$

Chapter 3

Lower Bounds for Smooth Optimization

3.1 Reductions between Convex & Strongly Convex Settings

An algorithm *successfully optimize* a function class $\mathcal{F}$ in $\phi(\mathcal{F}, R, \epsilon)$ iterations if, given any $f \in \mathcal{F}$ and $x_0 \in \mathbb{R}^d$ with $\|x_0 - x_\star\| \leqslant R$, it outputs x with $f(x) - f_\star \leqslant \epsilon$ using no more than $\phi(\mathcal{F}, R, \epsilon)$ queries to a first-order oracle for f.

Lemma 3.1.1: Convex to Strongly Convex Reduction

Assume there is an algorithm which successfully optimizes the class of convex and β-smooth functions in $\phi(\beta R^2/\epsilon)$ iterations. Then, there is an explicit algorithm which successfully optimizes the class of α-convex and β-smooth functions on $O(\phi(8\kappa)\log(\alpha R^2/\epsilon))$ iterations.

Proof. Let f be α-convex and β-smooth, and apply the given algorithm to f to obtain a new point x_1 with tolerance ϵ_1. By quadratic growth in 2.1.8, we have

$$\frac{\alpha}{2}\|x_1 - x_\star\|^2 \leqslant f(x_1) - f_\star \leqslant \epsilon_1$$

Set $\epsilon_1 = \alpha R^2/8$, so that

$$\|x_1 - x_\star\|^2 \leqslant \frac{2\epsilon_1}{\alpha} = \frac{R^2}{4} \implies \|x_1 - x_\star\| \leqslant \frac{1}{2}R = \frac{1}{2}\|x_0 - x_\star\| \tag{3.1}$$

Therefore, we cut off half of the length per use of this algorithm. Each use need $\phi(8\kappa)$ iterations, since

$$\phi\left(\frac{\beta R^2}{\epsilon_1}\right) = \phi\left(\beta R^2 \frac{8}{\alpha R^2}\right) = \phi(8\kappa)$$

From Equation 3.1, if we now repeat this procedure $O(\log(\alpha R^2/\epsilon))$ times, we can reach a point $\tilde{x}$ satisfying $\tilde{R} = \|\tilde{x} - x_\star\| \leqslant \sqrt{\epsilon/\alpha}$. This is because, after M interation, if we want to get this precesion, we want

$$\|x_M - x_\star\| \leqslant \left(\frac{1}{2}\right)^M R \leqslant \sqrt{\frac{\epsilon}{\alpha}} \implies \left(\frac{1}{2}\right)^{2M} R^2 \leqslant \frac{\epsilon}{\alpha} \implies -2M\log 2 \leqslant \log\left(\frac{\epsilon}{\alpha R^2}\right) \implies M \gtrsim \log\left(\frac{\alpha R^2}{\epsilon}\right)$$

Finally, apply the algorithm one more time starting from $\tilde{x}$ with target accuracy ϵ to obtain a point x with $f(x) - f_\star \leqslant \epsilon$.

This need iterations $\phi(\kappa)$, since

$$\phi\left(\frac{\beta\tilde{R}^2}{\epsilon}\right) = \phi\left(\frac{\beta\epsilon/\alpha}{\epsilon}\right) = \phi(\kappa)$$

Therefore, it is in total $\phi(8\kappa)\log(\alpha R^2/\epsilon) + \phi(\kappa)$ iterations, which is in the scale of $O(\phi(8\kappa)\log(\alpha R^2/\epsilon))$. □

Remark: For example, in Theorem 2.2.6, we get a rate $O(\beta R^2/\epsilon)$ for $\alpha = 0$. Taking $\phi(x) = O(x)$, we recover the $\alpha > 0$ case with rate $O(\kappa \log \alpha R^2/\epsilon)$.

Lemma 3.1.2: Strongly Convex to Convex Reduction

Assume there is an algorithm which successfully optimizes the class of α-convex and β-smooth functions in $\phi(\kappa)\log(\alpha R^2/\epsilon)$ iterations. Then, there is an explicit algorithm which successfully optimizes the class of convex and β-smooth functions in $O(\phi(2\beta R^2/\epsilon))$ iterations.

Proof. Let f be convex and β-smooth. We apply the given algorithm to the regularized function $f_\delta = f + \frac{\delta}{2}\|\cdot - x_0\|^2$. Then, f_δ is δ-convex and $(\beta+\delta)$-smooth. We apply the algorithm until obtaining a point x such that $f_\delta(x) \leqslant \min f_\delta + \epsilon/2$. If $x_{\delta,\star}$ denotes the minimizer of f_δ, we have

$$f(x) \leqslant f_\delta(x) \leqslant f_\delta(x_{\delta,\star}) + \frac{\epsilon}{2} \leqslant f_\delta(x_\star) + \frac{\epsilon}{2} = f_\star + \frac{\delta}{2}\|x_0 - x_\star\|^2 + \frac{\epsilon}{2} = f_\star + \frac{\delta R^2}{2} + \frac{\epsilon}{2}$$

where the third inequality follows that $f_\delta(x_{\delta,\star}) \leqslant f_\delta(x)$ for all x. Now, if we set $\delta = \epsilon/R^2$, we get

$$f(x) - f_\star \leqslant \frac{\epsilon R^2}{2R^2} + \frac{\epsilon}{2} = \epsilon$$

and the desired tolerance is achieved.

It remains to estimate the complexity. Since $x_{\delta,\star}$ is minimizer, we have

$$f_\delta(x_{\delta,\star}) = f(x_{\delta,\star}) + \frac{\delta}{2}\|x_{\delta,\star} - x_0\|^2 \leqslant f(x_\star) + \frac{\delta}{2}\|x_\star - x_0\|^2 = f_\delta(x_\star)$$

Since $f(x_{\delta,\star}) \geqslant f(x_\star)$, we have $\|x_0 - x_{\delta,\star}\| \leqslant \|x_0 - x_\star\| = R$. Therefore, the initial distance to the minimizer of f_δ is bounded by R. Now, we can assume that $\epsilon \leqslant \beta R^2$ (this is because, by β-smoothness, we already have $f(x) - f(x_\star) \leqslant \langle \nabla f(x_\star), x - x_\star \rangle) + \frac{\beta}{2}\|x - x_\star\|^2 = \beta R^2/2$, where $\nabla f(x_\star) = 0$ by optimality condition). Therefore, the smooth ness of f_δ is bounded by $\beta + \delta = \beta + \epsilon/R^2 \leqslant 2\beta$. Then, the condition number of f_δ is bounded by

$$\kappa(f_\delta) = \frac{\beta + \delta}{\delta} \leqslant \frac{2\beta}{\delta} = \frac{2\beta R^2}{\epsilon}$$

Substitute in these quntities into the complexity of the given algorithm, we see that

$$\#\text{ of iterations } = \phi(\kappa(f_\delta))\log\frac{\alpha(f_\delta)\|x_0 - x_{\delta,\star}\|^2}{\epsilon/2} = \phi\left(\frac{2\beta R^2}{\epsilon}\right)\log\left(\frac{\epsilon/R^2 \cdot R^2}{\epsilon/2}\right) = \phi\left(\frac{2\beta R^2}{\epsilon}\right)\log 2 = O(\phi(2\beta R^2/\epsilon))$$

which completes the proof. □

Remark: For example, in Theorem 2.2.6, we get a rate $O(\kappa \log \alpha R^2/\epsilon)$ for $\alpha > 0$. Taking $\phi(x) = O(x)$, we recover the

$\alpha = 0$ case with rate $O(\beta R^2/\epsilon)$.

Taken together, it shows that 0-convex and strongly convex settings are essentially equivalent to each other. The next question is, what is the smallest possible ϕ?

3.2 Gradient Span Algorithm Lower Bounds

Although possible (e.g., in Nemirovski and Yudin (1983)[10]), it is difficult to get a lower bound for any algorithm interacting with first-order oracle. Therefore, we impose some natural restrction on the class of algorithms under consideration. It is also the standard approach in this field. In this section, we consider the *gradient span algorithm*.

Definition 3.2.1: Gradient Span Algorithm

An algorithm is called a **gradient span algorithm** if it deterministically generates a sequence of points $\{x_n\}_{n\in\mathbb{N}}$ such that for all $n \in \mathbb{N}$,

$$x_{n+1} \in x_0 + \operatorname{span}\{\nabla f(x_0), \cdots, \nabla f(x_n)\}$$

For example, gradient descent is a gradient span algorithm. Now we establish the lower bound.

Theorem 3.2.2: Lower Bound for Convex, Smooth Optimization

For any $1 \leqslant N \leqslant \frac{d-1}{2}$, $\beta > 0$, and $x_0 \in \mathbb{R}^d$, there exists a convex and β-smooth function $f : \mathbb{R}^d \to \mathbb{R}$ such that for any gradient span algorithm,

$$f(x_N) - f_\star \gtrsim \frac{\beta \|x_0 - x_\star\|^2}{N^2}$$

In other words, in order to obtain $f(x_N) - f_\star \leqslant \epsilon$, the number of iterations must satisfy

$$N \gtrsim \sqrt{\frac{\beta \|x_0 - x_\star\|^2}{\epsilon}}$$

Proof. To separate the notation of iteration points x_k and elements of $\mathbb{R}^d$, we denote $x = (x[1], x[2], \cdots, x[d]) \in \mathbb{R}^d$. By translating the problem, we may, without loss of generality, assume that $x[0] = 0$. Consider the following quadratic function:

$$f_n : \mathbb{R}^d \to \mathbb{R},\ n \leqslant d, \quad f_n(x) := \frac{\beta}{4}\left\{\frac{1}{2}\left[x[1]^2 + \sum_{k=1}^{n-1}(x[k] - x[k+1])^2 + x[n]^2\right] - x[1]\right\}$$

STEP I: Since f_n is quadratic, it is convex. Since for any $v \in \mathbb{R}^d$,

$$\nabla f_n(x) = \frac{\beta}{4}(2x[1]-x[2]-1, 2x[2]-x[1]-x[3], 2x[3]-x[2]-x[4], \cdots, 2x[n-1]-x[n-2]-x[n], 2x[n]-x[n-1], 0, \cdots, 0)$$

$$[\nabla^2 f_n(x)]_{n\times n} = \frac{\beta}{4}\begin{pmatrix} 2 & -1 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 2 \end{pmatrix} \quad \text{with all other elements being } 0$$

$$\begin{aligned}\langle v, \nabla^2 f_n(x) v\rangle &= \frac{\beta}{4}(v[1],\cdots,v[n]) \begin{pmatrix} 2v[1]-v[2] \\ 2v[2]-v[1]-v[3] \\ \cdots \\ 2v[n]-v[n-1]\end{pmatrix} = \frac{\beta}{4}\left(v_1^2 + \sum_{k=1}^{n-1}(v_k - v_{k+1})^2 + v_n^2\right) \\ &\leqslant \frac{\beta}{4}\left(v[1]^2 + \sum_{k=1}^{n-1}(2v[k]^2+2v[k+1]^2) + v[n]^2\right) \leqslant \beta\|v\|^2 \qquad ((a-b)^2 \leqslant 2a^2+2b^2)\end{aligned}$$

Therefore, each f_n is convex and β-smooth.

STEP II: Next, we prove by induction on n that when we apply a gradient span algorithm to f_d, the nth iterate x_n belongs to the subspace

$$\mathcal{V}_n = \{x \in \mathbb{R}^d : x_k = 0 \text{ for all } k = n+1,\cdots,d\}$$

- First, clearly $x_0 = 0 \in \mathcal{V}_0$.
- Suppose $x_k \in \mathcal{V}_k$ for all $k \leqslant n$.
- Note that

$$\begin{aligned}\nabla f_d(x_k) &= \frac{\beta}{4}(2x_k[1]-x_k[2]-1, 2x_k[2]-x_k[1]-x_k[3],\cdots,2x[d-1]-x[d-2]-x[d], 2x[d]-x[d-1]) \\ &= \frac{\beta}{4}\left(2x_k[1]-x_k[2]-1, 2x_k[2]-x_k[1]-x_k[3],\cdots,2x_k[k]-x_k[k-1], -x_k[k], 0,\cdots,0\right) \qquad (x_k \in \mathcal{V}_k) \\ &= \frac{\beta}{4}\left(x_k[1]e_1 + \sum_{j=1}^{k}(x_k[j]-x_k[j+1])(e_j - e_{j+1})\right) - \frac{\beta}{4}e_1 \in \mathcal{V}_{k+1}\end{aligned}$$

Hence, by the definition of gradient span algorithm,

$$x_{n+1} \in \operatorname{span}\{\nabla f_d(x_0),\cdots,\nabla f_d(x_n)\} \subseteq \mathcal{V}_{n+1}$$

which concludes the induction. This shows that each query only shows ‘one more dimension of information’, which is the intuition for why we have the lower bound.

STEP III: The next step is to estimate $(f_n)_\star := \min f_n$ for all n. By setting the gradient to zero, $\nabla f_n(x_{n,\star}) = 0$, we obtain the following system of equations:

$$\begin{cases} 2x_{n,\star}[1] - x_{n,\star}[2] = 1 \\ 2x_{n,\star}[k] - x_{n,\star}[k-1] - x_{n,\star}[k+1] = 0, \qquad \text{for } k = 2,3,\cdots,n-1 \\ 2x_{n,\star}[n] - x_{n,\star}[n-1] = 0 \end{cases}$$

The solution is $x_{n,\star}[k] = 1 - \frac{k}{n+1}$ for all $k \in [n]$. If we write $f_n(x) = \frac{\beta}{4}\left(\frac{1}{2}\langle x, A_n x\rangle - \langle e_1, x\rangle\right)$. Then, the system above reads

$$\nabla f_n(x_{n,\star}) = \frac{\beta}{4}\left(A_n x_{n,\star} - e_1\right) = 0 \quad\Longrightarrow\quad A_n x_{n,\star} = e_1$$

Therefore,

$$(f_n)_\star = f_n(x_{n,\star}) = \frac{\beta}{4}\left(\frac{1}{2}\langle x_{n,\star}, e_1\rangle - \langle e_1, x_{n,\star}\rangle\right) = -\frac{\beta}{8}\langle e_1, x_{n,\star}\rangle = -\frac{\beta}{8}\left(1 - \frac{1}{n+1}\right)$$

Moreover, since $f_N = f_d$ on $\mathcal{V}_N$, it follows that $f_d(x_N) = f_N(x_N) \geqslant (f_N)_\star$. Also, $\|x_0 - x_{n,\star}\|^2 = \|x_{n,\star}\|^2 \leqslant n$ since each entry is smaller than 1. Combining these together, we have

$$f_d(x_N) - (f_d)_\star \geqslant (f_N)_\star - (f_d)_\star = \frac{\beta}{8}\left(\frac{1}{N+1} - \frac{1}{d+1}\right) \geqslant \frac{\beta\|x_0 - x_{d,\star}\|^2}{8d}\left(\frac{1}{N+1} - \frac{1}{d+1}\right)$$

Choose $d \asymp N$, e.g., $d = 2N+1$, yields the final result. □

Remarks:

- It is surprising that the lower bound construction is a quadratic function. In some sense, quadratics are the hardest convex and smooth functions to optimize.
- The lower bound requires the dimension to be larger than the iteration count. This is crucial for the proof, which relies on the algorithm discovering one new dimension per iteration. There are better methods in low dimension.
- For $d = 2N+1$, let $x_\star$ denotes the minimizer of f_d, we have

  $$x_\star[k] = 1 - \frac{k}{2N+2}, \quad k = 1, 2, \cdots, 2N+1, \qquad \|x_0 - x_\star\|^2 = \|x_\star\|^2 = \sum_{k=1}^{2N+1}\left(1 - \frac{k}{2N+2}\right)^2$$

  Since $x_N \in \mathcal{V}_n$, we have $x_N[k] = 0$ for all $k > N$. Therefore,

  $$\|x_N - x_\star\|^2 = \sum_{k=1}^{N}(x_N[k] - x_\star[k])^2 + \sum_{k=N+1}^{2N+1}\left(1 - \frac{k}{2N+2}\right)^2 \geqslant \sum_{k=N+1}^{2N+1}\left(1 - \frac{k}{2N+2}\right)^2$$

  If we replace $k = N+1+t$, $t = 0, 1, \cdots, N$, we have

  $$\|x_N - x_\star\|^2 \geqslant \sum_{t=0}^{N}\left(1 - \frac{N+1+t}{2N+2}\right)^2 = \sum_{t=0}^{N}\left(\frac{N+1-t}{2N+2}\right)^2 = \frac{1}{(2N+2)^2}\sum_{s=1}^{N+1} s^2$$

  Using the fact that $\sum_{s=1}^{n} s^2 = \frac{n(n+1)(2n+1)}{6}$, we have

  $$\|x_N - x_\star\|^2 \geqslant \frac{(N+1)(N+2)(2N+3)}{6(2N+2)^2}$$

  Similarly, we can see that the original square distance

  $$\|x_0 - x_\star\|^2 = \frac{1}{(2N+2)^2}\sum_{k=1}^{2N+1}(2N+2-k)^2 = \frac{(2N+1)(2N+2)(4N+3)}{6(2N+2)^2}$$

  Take the ratio of the two, we see that

  $$\frac{\|x_N - x_\star\|^2}{\|x_0 - x_\star\|^2} \geqslant \frac{(N+1)(N+2)(2N+3)}{(2N+1)(2N+2)(4N+3)} \geqslant \frac{(N+1)(N+2)(2N+3)}{(2N+2)(2N+4)(4N+6)} = \frac{1}{8}$$

  which shows that $\|x_N - x_\star\|^2 \gtrsim \|x_0 - x_\star\|^2$. In other word, in the 0-convex case, it is not possible to make progress in the sense of distance to the minimizer by more than a constant factor.

In the more general setting, $d \geqslant 2N+1$. It then follows exactly the same process. In this case, we have

$$x_\star[k] = 1 - \frac{k}{d+1}, \quad k = 1, 2, \cdots, d, \qquad \|x_0 - x_\star\|^2 = \|x_\star\|^2 = \sum_{k=1}^{d} \left(1 - \frac{k}{d+1}\right)^2$$

Since $x_N \in \mathcal{V}_n$, we have $x_N[k] = 0$ for all $k > N$. Therefore,

$$\|x_N - x_\star\|^2 \geqslant \sum_{k=N+1}^{d} \left(1 - \frac{k}{d+1}\right)^2 = \sum_{t=0}^{d-N-1} \left(1 - \frac{N+1+t}{d+1}\right)^2 = \frac{1}{(d+1)^2} \sum_{t=0}^{N} (d-N-t)^2 = \frac{1}{(d+1)^2} \sum_{s=1}^{d-N} s^2$$

Using the fact that $\sum_{s=1}^{n} s^2 = \frac{n(n+1)(2n+1)}{6}$, we have

$$\|x_N - x_\star\|^2 \geqslant \frac{(d-N)(d-N+1)(2d-2N+1)}{6(d+1)^2}$$

Similarly, the original square distance is

$$\|x_0 - x_\star\|^2 = \frac{1}{(d+1)^2} \sum_{k=1}^{d} (d+1-k)^2 = \frac{d(d+1)(2d+1)}{6(d+1)^2}$$

Take the ratio, we can see that

$$\frac{\|x_N - x_\star\|^2}{\|x_0 - x_\star\|^2} \geqslant \frac{(d-N)(d-N+1)(2d-2N+1)}{d(d+1)(2d+1)} \asymp const$$

since $d \asymp N$ in this setting, which also shows that $\|x_N - x_\star\|^2 \gtrsim \|x_0 - x_\star\|^2$.

Finally, by applying the Lemma 3.1.1, we have the lower bound for α-convex and β-smooth class.

Theorem 3.2.3: Lower Bound for Strongly Convex, Smooth Optimization

For any $0 < \alpha < \beta$, and $\epsilon > 0$, any d sufficiently large, and any $x_0 \in \mathbb{R}^d$, there exists an α-convex and β-smooth function $f : \mathbb{R}^d \to \mathbb{R}$ such that for any gradient span algorithm, in order to obtain $f(x_N) - f_\star \leqslant \epsilon$, the number of iterations must satisfy

$$N \gtrsim \sqrt{\kappa} \log \frac{\alpha \|x_0 - x_\star\|^2}{\epsilon}$$

Proof 1. Since in Theorem 3.2.2 we have $N \gtrsim \sqrt{\beta R^2/\epsilon}$, we can apply Lemma 3.1.1 with $\phi(x) \asymp \sqrt{x}$, so that we have for strongly convex and smooth case, $N \gtrsim \sqrt{\kappa} \log(\alpha R^2/\epsilon)$. □

Proof 2. By translating the problem, we may assume $x_0 = 0$. Consider the function

$$f : \mathbb{R}^\infty \to \mathbb{R}, \quad f(x) := \frac{\beta - \alpha}{8} \left(x[1]^2 + \sum_{n=1}^{\infty} (x[n] - x[n+1])^2 - 2x[1] \right) + \frac{\alpha}{2} \|x\|^2$$

STEP I: We first show that this function is α-convex and β-smooth. Note that

$$\nabla f(x) = \frac{\beta - \alpha}{8} (2x[1] - x[2] - 1, 2x[2] - x[1] - x[3], 2x[3] - x[2] - x[4], \cdots) + \alpha x$$

$$\langle v, \nabla^2 f(x) v\rangle = \frac{\beta-\alpha}{4}\left(v[1]^2 + \sum_{k=1}^{\infty}(v[k]-v[k+1])^2\right) + \alpha\|v\|^2$$

For α-convexity, we have

$$\alpha\|v\|^2 \leqslant \frac{\beta-\alpha}{4}\left(v[1]^2 + \sum_{k=1}^{\infty}(v[k]-v[k+1])^2\right) + \alpha\|v\|^2 = \langle v, \nabla^2 f(x) v\rangle$$

For β-smoothness, we have

$$\begin{aligned}
\langle v, \nabla^2 f(x) v\rangle &= \frac{\beta-\alpha}{4}\left(v[1]^2 + \sum_{k=1}^{\infty}(v[k]-v[k+1])^2\right) + \alpha\|v\|^2 \\
&\leqslant \frac{\beta-\alpha}{4}\left(2v[1]^2 + 2\sum_{k=1}^{\infty}(v[k]^2+v[k+1]^2)\right) + \alpha\|v\|^2 && ((a-b)^2 \leqslant 2a^2+2b^2) \\
&= (\beta-\alpha)\|v\|^2 + \alpha\|v\|^2 = \beta\|v\|^2
\end{aligned}$$

STEP II: To find the optimal solution $x_\star$, we set $\nabla f(x) = 0$, and we have the following second-order difference equation:

$$\begin{cases}
\dfrac{\beta-\alpha}{4}\left(2x_\star[1] - x_\star[2] - 1\right) + \alpha x_\star[1] = 0 \\
\dfrac{\beta-\alpha}{4}\left(2x_\star[k] - x_\star[k+1] - x_\star[k+2]\right) + \alpha x_\star[k] = 0, \quad k = 2, 3, \cdots
\end{cases}$$

Divide both sides by α, and rearrange, we can get the standard form:

$$x_\star[k+1] - 2\frac{\kappa+1}{\kappa-1}x_\star[k] + x_\star[k-1] = 0, \quad \text{with initial condition } \frac{\beta-\alpha}{4}\left(2x_\star[1] - x_\star[2] - 1\right) + \alpha x_\star[1] = 0$$

The characteristic equation is

$$\lambda^2 - 2\frac{\kappa+1}{\kappa-1}\lambda + 1 = 0$$

which has two distinct real solution

$$\lambda = \frac{2\frac{\kappa+1}{\kappa-1} \pm \sqrt{4\left(\frac{\kappa+1}{\kappa-1}\right)^2 - 4}}{2} = \frac{\kappa+1}{\kappa-1} \pm \frac{2\sqrt{\kappa}}{\kappa-1} = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \text{ or } \frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}$$

Therefore, the general solution is

$$x_\star[k] = A\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k + B\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^k$$

However, since f is a quadratic function with positive second-order term, the norm of x must be bounded to get a finite function value. Therefore, B must equal to 0, otherwise $x_\star[k]$ is increasing with k, and it is not a minimization point of f. To find A, we apply the initial condition

$$\begin{aligned}
&\frac{\beta-\alpha}{4}\left(2A\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} - A\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2 - 1\right) + A\alpha\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} = 0 \\
\Longrightarrow\quad &\frac{\sqrt{\kappa}-1}{4}\left(\frac{(A-1)\kappa + (2A-2)\sqrt{\kappa} - 3A - 1}{\sqrt{\kappa}+1}\right) + A\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} = 0
\end{aligned}$$

which leads to $A = 1$. Therefore, the solution is

$$x_\star[k] = \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k \tag{3.2}$$

STEP III: Now we prove by induction that when we apply a gradient span algorithm to f, the nth iteration $x_n \in \mathcal{V}_n$. Clearly, $x_0 = 0 \in \mathcal{V}_0$, and suppose $x_k \in \mathcal{V}_k$ for all $k \leqslant n$. Then,

$$\begin{aligned}\nabla f(x_k) &= \frac{\beta-\alpha}{8}(2x[1]-x[2]-1, 2x[2]-x[1]-x[3], 2x[3]-x[2]-x[4], \cdots) + \alpha(x[1], x[2], \cdots) \\ &= \frac{\beta-\alpha}{8}(\underbrace{2x[1]-x[2]-1, \cdots, 2x[k]-x[k-1], -x[k]}_{k+1 \text{ terms}}, 0, 0\cdots) + \alpha(\underbrace{x[1], x[2], \cdots, x[k]}_{k \text{ terms}}, 0, 0, \cdots) \in \mathcal{V}_{k+1}\end{aligned}$$

where the second equality holds since $x_k \in \mathcal{V}_k$. Hence,

$$x_{n+1} \in \operatorname{span}\{\nabla f(x_0), \cdots, \nabla f(x_n)\} \subseteq \mathcal{V}_{n+1}$$

STEP IV: By α-strong convexity, we have that

$$f(x_N) - f_\star \geqslant \langle \underbrace{\nabla f(x_\star)}_{=0}, x_N - x_\star\rangle + \frac{\alpha}{2}\|x_N - x_\star\|^2 = \frac{\alpha}{2}\|x_N - x_\star\|^2 \tag{3.3}$$

We have closed-form solution for $x_\star$, and $x_N \in \mathcal{V}_N$, so we can plug in and get

$$\|x_N - x_\star\|^2 = \sum_{k=1}^{N}(x_N[k] - x_\star[k])^2 + \sum_{k\geqslant N+1} x_\star[k]^2 \geqslant \sum_{k\geqslant N+1} x_\star[k]^2$$

Substitute in the formula of $x_\star$ using Equation 3.2, we have

$$\|x_N - x_\star\|^2 \geqslant \sum_{k\geqslant N+1}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2k} = \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2N}\sum_{k\geqslant 1}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2k} = \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2N}\|x_0 - x_\star\|^2 \tag{3.4}$$

Combining Equation 3.3 and 3.4, we have finally

$$f(x_N) - f_\star \geqslant \frac{\alpha}{2}\|x_N - x_\star\|^2 \geqslant \frac{\alpha}{2}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2N}\|x_0 - x_\star\|^2$$

which completes the proof. If we want $f(x_N) - f_\star \leqslant \epsilon$, we have

$$\frac{\alpha}{2}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2N}\|x_0 - x_\star\|^2 \leqslant \epsilon \implies 2N\log\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right) \geqslant \log(\alpha R^2/2\epsilon)$$

Since on the LHS, we have the rate

$$\log\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right) = \log\left(1 + \frac{2}{\sqrt{\kappa}-1}\right) \approx \frac{2}{\sqrt{\kappa}}$$

by Taylor expansion. Thus, we have that $N \gtrsim \sqrt{\kappa}\log(\alpha R^2/\epsilon)$. □

Remark: The iteration complexity lower bounds in these two theorems are smaller than the bounds attained by gradient descent by a square root. As developed in the next sections, in fact the lower bounds are tight and gradient descent is suboptimal.

Chapter 4

Acceleration

We now show that the lower bounds of Theorem 3.2.2 and 3.2.3 can be attained via improvement of gradient flow. This is called *acceleration* phenomenon in optimization.

4.1 Quadratic Case: Conjugate Gradient Method

4.1.1 Algorithm and Convergence Rate

The objective function for this section is quadratic:

$$f : \mathbb{R}^d \to \mathbb{R}, \quad f(x) = \frac{1}{2}\langle x, Ax\rangle - \langle b, x\rangle$$

where A is a symmetric matrix, and $A \succ 0$. Note that minimizing f corresponds to solving the system of equations $Ax_\star = b$. Instead, we can use **conjugate gradient method**, which is defined as

$$x_{n+1} := \operatorname{argmin}\{f(x) \,|\, x \in x_0 + \operatorname{span}\{\nabla f(x_0), \nabla f(x_1), \cdots, \nabla f(x_n)\}\} \tag{CG}$$

The aim is to show that if f is quadratic, then CG can be rewritten as a simple iteration that uses one gradient query per step. Instead of working with $\{\nabla f(x_0), \nabla f(x_1), \cdots, \nabla f(x_n)\}$, it is more convenient to work with an *orthogonal set* $\{p_0, p_1, \cdots, p_n\}$. Here, orthogonality is w.r.t. the inner product $\langle\cdot,\cdot\rangle_A$, i.e., $\langle p_i, Ap_j\rangle = 0$ for all $i \neq j$. We start with $p_0 := \nabla f(x_0)$, and write $\mathcal{K}_n = \operatorname{span}\{p_0, p_1, \cdots, p_n\}$. We must address the following 2 questions:

- **Question 1:** Given $\mathcal{K}_n$ and x_n, how can we compute $x_{n+1} = \operatorname{argmin}_{x_0+\mathcal{K}_n} f$?
- **Question 2:** Given $\mathcal{K}_n$ and $\nabla f(x_{n+1})$, how can we compute p_{n+1} and thus $\mathcal{K}_{n+1}$?

Question 1: Assume inductively that $x_n = \operatorname{argmin}_{x_0+\mathcal{K}_{n-1}} f$. This means that $\langle\nabla f(x_n), p_k\rangle = 0$ for all $k < n$ (otherwise, if it has a nonzero component in $\mathcal{K}_{n-1}$, it can be moved along that direction and decrease the function value). The next point is taken to be $x_{n+1} = x_n + h_n p_n$, where h_n is chosen so that $\langle\nabla f(x_{n+1}), p_k\rangle = 0$ for all $k \leqslant n$, i.e., $\nabla f(x_{n+1})$ is orthogonal to $\mathcal{K}_n$. Since $\nabla f(x) = Ax - b$,

$$\langle\nabla f(x_{n+1}), p_k\rangle = \langle\nabla f(x_n + h_n p_n), p_k\rangle = \langle A(x_n + h_n p_n) - b, p_k\rangle = \langle\nabla f(x_n) + h_n Ap_n, p_k\rangle$$

For $k < n$, this equals zero by $\langle \nabla f(x_n), p_k \rangle = 0$ and $\langle p_n, p_k \rangle_A = 0$. Therefore, we only need to choose h_n so that it is also zero for $k = n$. Therefore, we choose

$$h_n = -\frac{\langle \nabla f(x_n), p_n \rangle}{\|p_n\|_A^2} \tag{4.1}$$

Question 2: We want to compute the Gram-Schmidt orthogonalization of $\nabla f(x_{n+1})$ w.r.t. $\{p_0, p_1, \cdots, p_n\}$ in the $\langle \cdot, \cdot \rangle_A$ inner product. We claim that we can choose p_{n+1} in the way below:

$$p_{n+1} = \nabla f(x_{n+1}) - \langle \nabla f(x_{n+1}), p_n \rangle_A \frac{p_n}{\|p_n\|_A^2} \tag{4.2}$$

That is, when doing Gram-Schmidt, we can just substract the part that is on the direction of p_n, i.e., $\nabla f(x_{n+1})$ is already A-orthogonal to p_k for $k < n$. To justify this, we need the following lemma.

Lemma 4.1.1: Krylov Subspaces

For all $n \in \mathbb{N}$,

$$\mathcal{K}_n = \operatorname{span}\{p_0, Ap_0, \cdots, A^n p_0\}$$

Proof. We prove by induction on n. We want to show that $\tilde{\mathcal{K}}_n = \operatorname{span}\{p_0, Ap_0, \cdots, A^n p_0\} = \mathcal{K}_n = \operatorname{span}\{p_0, p_1, \cdots, p_n\}$. For $n = 0$, $\tilde{\mathcal{K}}_0 = \mathcal{K}_0$ obviously. Suppose it holds at iteration n.
($\Longrightarrow$) We first need to show that $p_{n+1} \in \tilde{\mathcal{K}}_{n+1}$. However, by Equation 4.2, Since $p_n \in \tilde{\mathcal{K}}_{n+1}$ by induction assumption, we only need to show that $\nabla f(x_{n+1}) \in \tilde{\mathcal{K}}_{n+1}$. However, as discussed in Question 1,

$$\nabla f(x_{n+1}) = \nabla f(x_n + h_n p_n) = \nabla f(x_n) + h_n A p_n = \cdots = p_0 + h_0 A p_0 + \cdots + h_n A p_n \in \tilde{\mathcal{K}}_{n+1}$$

where the final equation is achieved iteratively.
($\Longleftarrow$) Conversely, we also need to show that $A^{n+1} p_0 \in \mathcal{K}_{n+1}$. Since $A^n p_0 \in \mathcal{K}_n$ by our induction hypothesis, we can write $A^n p_0 = \sum_{k=0}^n c_k p_k$, thus $A^{n+1} p_0 = \sum_{k=0}^n c_k A p_k$. By induction hypothesis, since $A p_k = A \sum_{i=0}^k c_i A^i p_0 = \sum_{i=0}^k c_i A^{i+1} p_0$ is a linear combination of $A p_0, \cdots, A^{k+1} p_0$, we have $A p_k$ belongs to $\mathcal{K}_n$ for all $k < n$. The only thing we need to prove is $A p_n \in \mathcal{K}_{n+1}$. However, by the discussion before, we have that

$$\begin{aligned} & \nabla f(x_{n+1}) = \nabla f(x_n) + h_n A p_n \\ \Longrightarrow \quad & A p_n = h_n^{-1} (\nabla f(x_{n+1}) - \nabla f(x_n)) \in \mathcal{K}_{n+1} \end{aligned}$$

since $\nabla f(x_{n+1})$ and $\nabla f(x_n)$, by definition of $\mathcal{K}_n$, belongs to $\mathcal{K}_{n+1}$ ($\mathcal{K}_n$ is just orthogonalization of ∇f's). □

The subspaces $\{\mathcal{K}_n\}_{n \in \mathbb{N}}$ are called **Krylov subspaces**. What we have done in the previous lemma is that, $A p_k \in \mathcal{K}_{k+1}$ for all $k < n$, hence $\langle \nabla f(x_{n+1}), p_k \rangle_A = \langle \nabla f(x_{n+1}), A p_k \rangle = 0$ using the fact shown before that $\nabla f(x_{n+1})$ is orthogonal (in the usual inner product) to $\mathcal{K}_n$. Therefore, Equation 4.2 is justified.

Combining all: To finalize our algorithm, we need some equations to simplify our previous rsults. Note first that since $\nabla f(x_{n+1})$ is orthogonal to all elements of $\mathcal{K}_n$, we have $\langle \nabla f(x_n), \nabla f(x_{n+1}) \rangle = 0$. Therefore,

$$\frac{\langle \nabla f(x_{n+1}), p_n \rangle_A}{\|p_n\|_A^2} = \frac{\langle \nabla f(x_{n+1}), \nabla f(x_{n+1}) - \nabla f(x_n) \rangle}{h_n \|p_n\|_A^2} = -\frac{\|\nabla f(x_{n+1})\|^2}{\langle \nabla f(x_n), p_n \rangle} \tag{4.3}$$

where the first equation comes from that $\nabla f(x_{n+1}) = \nabla f(x_n) + h_n A p_n$, and the second equation comes from that $\langle \nabla f(x_n), \nabla f(x_{n+1}) \rangle = 0$ and Equation 4.1. Moreover,

$$\|\nabla f(x_n)\|^2 = \langle \nabla f(x_n), \nabla f(x_n) \rangle = \langle \nabla f(x_n), p_n \rangle \tag{4.4}$$

since $\nabla f(x_n)$ is orthogonal to $\mathcal{K}_{n-1}$ and by Equation 4.2 where $p_n = \nabla f(x_n) - \alpha p_{n-1}$, so $\langle \nabla f(x_n), p_n \rangle = \langle \nabla f(x_n), \nabla f(x_n) - \alpha p_{n-1} \rangle = \langle \nabla f(x_n), \nabla f(x_n) \rangle$. Therefore, denote the residual $r_n := Ax_n - b = \nabla f(x_n)$, combining Equation 4.1 and 4.4, we have the iteration for x_n:

$$x_{n+1} = x_n + h_n p_n = x_n - \frac{\langle \nabla f(x_n), p_n \rangle}{\|p_n\|_A^2} p_n = x_n - \frac{\|r_n\|^2}{\langle p_n, Ap_n \rangle} p_n$$

Using $\nabla f(x_{n+1}) = \nabla f(x_n) + h_n A p_n$, we can get the iteration for r_n:

$$r_{n+1} = r_n + h_n A p_n = r_n - \frac{\langle \nabla f(x_n), p_n \rangle}{\|p_n\|_A^2} A p_n = r_n - \frac{\|r_n\|^2}{\langle p_n, Ap_n \rangle} A p_n$$

Finally, Combine equation 4.2, 4.3 and 4.4, we have the iteration for p_n:

$$p_{n+1} = \nabla f(x_{n+1}) - \langle \nabla f(x_{n+1}), p_n \rangle_A \frac{p_n}{\|p_n\|_A^2} = r_{n+1} + \frac{\|r_{n+1}\|^2}{\langle \nabla f(x_n), p_n \rangle} p_n = r_{n+1} + \frac{\|r_{n+1}\|^2}{\|r_n\|^2} p_n$$

where the first equality from 4.2, second from 4.3, and third from 4.4. We conclude the iteration algorithm:

Conjugate Gradient Descent for Quadratic optimization:

$$x_{n+1} = x_n - \frac{\|r_n\|^2}{\langle p_n, Ap_n \rangle} p_n, \qquad r_{n+1} = r_n - \frac{\|r_n\|^2}{\langle p_n, Ap_n \rangle} A p_n, \qquad p_{n+1} = r_{n+1} + \frac{\|r_{n+1}\|^2}{\|r_n\|^2} p_n$$

Remark: For gradiant descent, $x_n \in \mathcal{K}_n$, since $x_1 = x_0 - hp_0$, $x_2 = x_1 - h(Ax_1 - b) = x_1 - h(Ax_0 - hAp_0 - b) = x_1 - h(p_0 - hAp_0) \in x_0 + \text{span}\{p_0, Ap_0\}$. Thus the iteration is also in the Krylov subspaces. However, conjugate gradient is better than gradient descent since it finds the best next point among these span for each step.

Theorem 4.1.2: Termination of Conjugate Gradient

The conjugate gradient algorithm returns the exact minimizer in at most d iterations.

Proof. Note that if $p_{n+1} = 0$, then by Equation 4.2, $\nabla f(x_{n+1}) = \alpha p_n \in \mathcal{K}_n$. However, since $\nabla f(x_{n+1}) \perp \mathcal{K}_n$, we have $\nabla f(x_{n+1}) = 0$. Therefore, $x_{n+1} = x_\star$. Since an orthogonal set in $\mathbb{R}^d$ cannot have more than d elements, we have $p_{d+1} = 0$, which implies our theorem. □

Conjugate gradient also find an approximate minimizer at the accelerated rate.

Theorem 4.1.3: Acceelerated Convergence for Conjugate Gradient

Let $0 \prec \alpha I \preccurlyeq A \preccurlyeq \beta I$. Then, conjugate gradient outputs x_N satisfying $f(x_N) - f_\star \leqslant \epsilon$ in $N = O\left(\sqrt{\kappa} \log \frac{f(x_0) - f_\star}{\epsilon}\right)$ iterations.

Proof. By the descent lemma 2.2.1 and the definition of conjugate gradient, we have

$$f(x_{n+1}) \leqslant f\left(x_n - \frac{1}{\beta}\nabla f(x_n)\right) \leqslant f(x_n) - \frac{1}{2\beta}\|\nabla f(x_n)\|^2$$

where the first inequality uses the definition that conjugate gradient chooses the best point among the spanned space, and the second inequality follows from descent lemma 2.2.1 with $h = \frac{1}{\beta}$. Telescoping the sum, we have

$$\begin{aligned} & \sum_{n=0}^{N-1}(f(x_n) - f(x_{n+1})) = f(x_0) - f(x_N) \geqslant \frac{1}{2\beta}\sum_{n=0}^{N-1}\|\nabla f(x_n)\|^2 \\ \implies \quad & f(x_0) - f_\star \geqslant f(x_0) - f(x_N) \geqslant \frac{1}{2\beta}\sum_{n=0}^{N-1}\|\nabla f(x_n)\|^2 \end{aligned} \tag{4.5}$$

On the other hand, since for $k < n$, $x_{k+1} - x_k = h_k p_k \in \mathcal{K}_k \subseteq \mathcal{K}_{n-1}$, and $\nabla f(x_n) \perp \mathcal{K}_{n-1}$, we have $\nabla f(x_n) \perp x_{k+1} - x_k$ for $k < n$. Thus, along with convexity, we have

$$f_\star - f(x_n) \geqslant \langle \nabla f(x_n), x_\star - x_n\rangle = \langle \nabla f(x_n), x_\star - x_n\rangle + \sum_{k=0}^{n-1}\langle \nabla f(x_n), x_{k+1} - x_k\rangle = \langle \nabla f(x_n), x_\star - x_0\rangle \tag{4.6}$$

If we sum these inequalities and use orthogonality of the gradients,

$$\begin{aligned} N(f(x_N) - f_\star) &\leqslant \sum_{n=0}^{N-1}(f(x_n) - f_\star) && \text{(Conjugate gradient is descening by definition)} \\ &\leqslant \left\langle \sum_{n=0}^{N-1}\nabla f(x_n), x_0 - x_\star \right\rangle && \text{(Equation 4.6)} \\ &\leqslant \left\| \sum_{n=0}^{N-1}\nabla f(x_n) \right\| \|x_0 - x_\star\| && \text{(Cauchy Schwarz)} \\ &= \left(\sum_{n=0}^{N-1}\|\nabla f(x_n)\|^2\right)^{1/2} \|x_0 - x_\star\| && \text{(Orthogonality of the gradients)} \\ &\leqslant \left(\sum_{n=0}^{N-1}\|\nabla f(x_n)\|^2\right)^{1/2} \sqrt{\frac{2(f(x_0) - f_\star)}{\alpha}} && \text{(Strong convexity with } \nabla f(x_\star) = 0\text{)} \\ &\leqslant \sqrt{2\beta(f(x_0) - f_\star)}\sqrt{\frac{2(f(x_0) - f_\star)}{\alpha}} && \text{(Equation 4.5)} \\ &= 2\sqrt{\kappa}(f(x_0) - f_\star) \end{aligned}$$

Let N be such that $f(x_N) - f_\star \geqslant (f(x_0) - f_\star)/2$. The inequality above then implies that $N \leqslant 4\sqrt{\kappa}$. Thus, every $4\sqrt{\kappa}$ iterations, the objective gap decreases by a factor of 2. □

4.1.2 Relation with Polynomial Approximation

There is a classical link between conjugate gradient with *polynomial approximation*.

- By Lemma 4.1.1, $x_N - x_0 \in \mathcal{K}_{n-1}$ can be written in the form $x_N - x_0 = \sum_{n=0}^{N-1} c_n A^n p_0$. Since $p_0 = \nabla f(x_0) = Ax_0 - b$, and $x_\star$ is the minimizer ($Ax_\star = b$), we have $p_0 = Ax_0 - Ax_\star$, so $x_N - x_\star = x_N - x_0 + x_0 - x_\star =$

$\sum_{n=0}^{N-1} c_n A^{n+1}(x_0 - x_\star) + (x_0 - x_\star) = P_N(A)(x_0 - x_\star)$ where P_N is a polynomial of degree at most N satisfying $P_N(0) = 1$.

- Conversely, if Q_N is any other degree N polynomial with $Q_N(0) = 1$, then define $\tilde{x}_N := x_0 + A^{-1}(Q_N(A) - I)p_0 \in x_0 + \mathcal{K}_{N-1}$, it satisfies $\tilde{x}_N - x_\star = x_0 - x_\star + A^{-1}(Q_N(A) - I)p_0 = Q_N(A)(x_0 - x_\star)$.

This equivalence, together with the fact that the output x_N of conjugate gradient minimizes f over $x_0 + \mathcal{K}_{N-1}$, we have

$$f(x_N) - f_\star = \frac{1}{2}\|x_N - x_\star\|_A^2 \leqslant \frac{1}{2}\min\left\{\|Q_N(A)(x_0 - x_\star)\|_A^2 : Q_N \in \mathbb{R}_{\leqslant N}[X], Q_N(0) = 1\right\}$$

where $\mathbb{R}_{\leqslant N}[X]$ denotes the set of polynomials with real-valued coefficients and with degree at most N. The equality follows from

$$\begin{aligned}
f(x_N) - f_\star &= \frac{1}{2}\langle x_N, Ax_N\rangle - \langle b, x_N\rangle - \frac{1}{2}\langle x_\star, Ax_\star\rangle + \langle b, x_\star\rangle \\
&= \frac{1}{2}\langle x_N, Ax_N\rangle - \frac{1}{2}\langle Ax_\star, x_N\rangle - \frac{1}{2}\langle Ax_\star, x_N\rangle - \frac{1}{2}\langle x_\star, Ax_\star\rangle + \langle Ax_\star, x_\star\rangle \\
&= \frac{1}{2}\langle x_N, A(x_N - x_\star)\rangle - \frac{1}{2}\langle Ax_\star, x_N - x_\star\rangle \\
&= \frac{1}{2}\langle Ax_N, x_N - x_\star\rangle - \frac{1}{2}\langle Ax_\star, x_N - x_\star\rangle && (A \text{ is symmetric}) \\
&= \frac{1}{2}\|x_N - x_\star\|_A^2
\end{aligned}$$

Furthermore, since A and $Q_N(A)$ commute ($Q_N(A)$ is a function of A, matrix multiplication is commutative when multiplying by themselves), we have

$$\|Q_N(A)(x_0 - x_\star)\|_A^2 \leqslant \|Q_N(A)\|_{\text{op}}^2 \|x_0 - x_\star\|_A^2 \leqslant \left(\max_{[\lambda_{\min}(A), \lambda_{\max}(A)]} |Q_N|^2\right)\|x_0 - x_\star\|_A^2$$

This leads to the bound

Assume that $0 \prec \alpha I \preccurlyeq A \preccurlyeq \beta I$. Then, the output x_N of conjugate gradient satisfies:

$$f(x_N) - f_\star \leqslant \min\left\{\max_{\lambda \in [\alpha, \beta]} |Q_N(\lambda)|^2 : Q_N \in \mathbb{R}_{\leqslant N}[X], Q_N(0) = 1\right\}(f(x_0) - f_\star) \tag{4.7}$$

To bound the rate of convergence, we remain to exhibit a judicious polynomial Q_N. This is accomplished by the family of Chebyshev polynomials.

Definition 4.1.4: Chebyshev Polynomial

The degree-n **Chebyshev polynomial** T_n is defined so that $\cos(n\theta) = T_n(\cos\theta)$ for all $\theta \in \mathbb{R}$.

For the bound of convergence rate, we choose

$$Q_n(x) = T_n\left(\frac{\alpha + \beta - 2x}{\beta - \alpha}\right) \Big/ T_n\left(\frac{\alpha + \beta}{\beta - \alpha}\right), \quad Q_n(0) = T_n\left(\frac{\alpha + \beta}{\beta - \alpha}\right) \Big/ T_n\left(\frac{\alpha + \beta}{\beta - \alpha}\right) = 1$$

Note that for $x \in [-1, 1]$, we can write Definition 5.6 in the form

$$T_n(x) = \cos\left(n \cos^{-1} x\right) \tag{4.8}$$

Therefore,

$$\begin{aligned} T_{n+1}(x) + T_{n-1}(x) &= \cos((n+1)\cos^{-1} x) + \cos((n-1)\cos^{-1} x) \\ &= 2\cos(n\cos^{-1} x)\cos(\cos^{-1} x) && (\cos(u+v) + \cos(u-v) = 2\cos u \cos v) \\ &= 2x\cos(n\cos^{-1} x) = 2xT_n(x) \end{aligned}$$

This leads to the three-term recurrence

$$T_{n+1}(x) - 2xT_n(x) + T_{n-1}(x) = 0$$

The characteristic equation and the solution is

$$\lambda^2 - 2x\lambda + 1 = 0 \quad \Longrightarrow \quad \lambda = x \pm \sqrt{x^2 - 1}$$

Therefore,

$$T_n(x) = A\left(x + \sqrt{x^2-1}\right)^n + B\left(x - \sqrt{x^2-1}\right)^n$$

To determine A and B, we examine $T_0(x)$ and $T_1(x)$. By Equation 4.8, we have

$$T_0(x) = 1, \quad T_1(x) = x$$

Substitute in $T_n(x)$, we get $A = B = 1/2$. Therefore, we get the formula

$$T_n(x) = \frac{1}{2}\left(\left(x + \sqrt{x^2-1}\right)^n + \left(x - \sqrt{x^2-1}\right)^n\right), \quad x \in [-1, 1]$$

Indeed, this formula can be extended to $x \in \mathbb{R}$. Let $x = \cos\theta$, we have

$$T_n(x) = \cos(n\theta) = \frac{e^{in\theta} + e^{-in\theta}}{2}$$

Note that $e^{i\theta} + e^{-i\theta} = 2\cos\theta = 2x$ and $e^{i\theta}e^{-i\theta} = 1$, by Vieta's formulas, they are the two solutions of $\lambda^2 - 2x\lambda + 1$, which leads to

$$e^{i\theta} = x + \sqrt{x^2-1}, \quad e^{-i\theta} = x - \sqrt{x^2-1}$$

substitute this back to the formula of T_n, we obtain the same formula

$$T_n(x) = \frac{1}{2}\left(\left(x + \sqrt{x^2-1}\right)^n + \left(x - \sqrt{x^2-1}\right)^n\right), \quad x \in \mathbb{R}$$

Therefore, we can write

$$T_n\left(\frac{\alpha+\beta-2x}{\beta-\alpha}\right) = \cos\left(n\cos^{-1}\left(\frac{\alpha+\beta-2x}{\beta-\alpha}\right)\right) \in [-1, 1]$$

since $(\alpha+\beta-2x)/(\beta-\alpha) \in [-1,1]$ when $x \in [\alpha,\beta]$. Moreover,

$$\begin{aligned} T_n\left(\frac{\alpha+\beta}{\beta-\alpha}\right) = T_n\left(\frac{\kappa+1}{\kappa-1}\right) &= \frac{1}{2}\left(\left(\frac{\kappa+1}{\kappa-1}+\sqrt{\frac{4\kappa}{(\kappa-1)^2}}\right)^n + \left(\frac{\kappa+1}{\kappa-1}-\sqrt{\frac{4\kappa}{(\kappa-1)^2}}\right)^n\right) \\ &= \frac{1}{2}\left(\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^n + \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^n\right) \geqslant \frac{1}{2}\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^n \end{aligned}$$

Therefore, we have the relation

$$|Q_n(x)| = \left|T_n\left(\frac{\alpha+\beta-2x}{\beta-\alpha}\right)\right| \Big/ T_n\left(\frac{\alpha+\beta}{\beta-\alpha}\right) \leqslant \frac{1}{\frac{1}{2}\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^n} = 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^n$$

Since this holds for all $x \in [\alpha,\beta]$, we have

$$\max_{x\in[\alpha,\beta]} |Q_n(x)| \leqslant 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^n$$

Note that by combining this with 4.7, we obtain an exponential rate of convergence for conjugate gradient matching the lower bound of Theorem 3.2.3. Therefore, conjugate gradient is **optimal** in this sense.

Remark 1: Here we state the motivation for following sections. If we can compute $\tilde{x}_N = x_0 + A^{-1}(Q_N(A) - I)p_0$, then it incurs error at most $f(\tilde{x}_N) - f_\star \leqslant \left(\max_{\lambda\in[\alpha,\beta]} |Q_N(\lambda)|^2\right)(f(x_0) - f_\star)$. In particular, rather than using conjugate gradient, we can try to compute the polynomial Q_N define above that achieves the fast convegence rate. This can be computed by the three-term recurrece. Therefore, it leads to a wish of optimization algorithm of the form

$$x_{n+1} = c_0 A x_n + c_1 x_{n-1} + c_2 b$$

where $c_0, c_1, c_2 \in \mathbb{R}$ are fixed coefficients. Note taht unlike gradient descent, x_{n+1} depends on the previous two iterates. This is often referred to as *momentum*, and also forms the basis for acceleration for general convex functions.

Remark 2: Practicality of Conjugate Gradient Solving linear system $Ax = b$ via Gaussian elimination requires $O(d^3)$ operations and is numerically unstable. For well-conditioned matrices A, conjugate gradient method, instead, returns an approximate solution in $\tilde{O}(\sqrt{\kappa})$ iterations (the 'tilde-O notation' ignores logarithm terms), each of which requires a matrix-vector multiplication. A matrix-vector multiplication requires $O(d^2)$ time in the worst case, but can be faster if A is sparse. In practice, conjugate gradient is widely used, especially when combined with other strategies such as preconditioning.

4.2 General Case: Continuous Time

The **aceelerated gradient flow** is

$$\begin{aligned} \dot{x}_t &= p_t \\ \dot{p}_t &= -\nabla f(x_t) - \gamma_t p_t \end{aligned} \tag{AGF}$$

p_t is the *momentum* (for a particle with unit mass). The part $\dot{x}_t = p_t$, and $\dot{p}_t = -\nabla f(x_t)$ is called *Hamilton's equations*, and they are just reformulation of Newton's law of motion. Hamilton's equations conserve the energy

$H(x,p) = f(x) + \frac{1}{2}\|p\|^2$, and this is undesirable for an optimization algorithm. Thus, the second part $\dot{p}_t = -\gamma_t p_t$ adds a dissipative *friction force*. For f convex, it turns out that the right choice is $\gamma_t = 3/t$. This is mysterious at first sight and was obtained by taking the continuous-time limit of Nesterov's discrete algorithm in the next subsection. As in Chapter 2, we assume f is smooth, admits a minimizer, and AGF is well-posed.

Theorem 4.2.1: Convergence of AGF under Convexity

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and let $(x_t)_{t\geqslant 0}$ evolve along AGF with $\gamma_t = 3/t$ and $p_0 = 0$. Then, for all $t \geqslant 0$,

$$f(x_t) - f_\star \leqslant \frac{2\|x_0 - x_\star\|^2}{t^2}$$

Proof. Consider the auxiliary point $z_t := x_t + \frac{t}{2}p_t$, and the Lyapunov function

$$\mathcal{L}_t = \frac{t^2}{2}(f(x_t) - f_\star) + \|z_t - x_\star\|^2$$

Take derivative, we have

$$\begin{aligned}
\dot{\mathcal{L}}_t &= t(f(x_t) - f_\star) + \frac{t^2}{2}\langle \nabla f(x_t), \dot{x}_t\rangle + 2\langle z_t - x_\star, \dot{z}_t\rangle \\
&= t(f(x_t) - f_\star) + \frac{t^2}{2}\langle \nabla f(x_t), p_t\rangle + 2\left\langle z_t - x_\star, p_t + \frac{1}{2}p_t + \frac{t}{2}\left(-\nabla f(x_t) - \frac{3}{t}p_t\right)\right\rangle && \text{(Definition of AGF)} \\
&= t(f(x_t) - f_\star) + \frac{t^2}{2}\langle \nabla f(x_t), p_t\rangle - t\left\langle \nabla f(x_t), x_t + \frac{t}{2}p_t - x_\star\right\rangle \\
&= t(f(x_t) - f_\star) + \frac{t^2}{2}\langle \nabla f(x_t), p_t\rangle - \frac{t^2}{2}\langle \nabla f(x_t), p_t\rangle - t\langle \nabla f(x_t), x_t - x_\star\rangle \\
&= t(f(x_t) - f_\star) - t\langle \nabla f(x_t), x_t - x_\star\rangle \leqslant 0 && \text{(Convexity)}
\end{aligned}$$

Therefore, it is decreasing, and $\mathcal{L}_t \leqslant \mathcal{L}_0$. Substitute in, we have

$$\frac{t^2}{2}(f(x_t) - f_\star) \leqslant \frac{t^2}{2}(f(x_t) - f_\star) + \|z_t - x_\star\|^2 = \mathcal{L}_t \leqslant \mathcal{L}_0 = \|x_0 - x_\star\|^2$$

Rearrange, we have the desired result. □

Remark: Although the Lyapunov function above appears fortuitous, it can be derived in a reasonably systematic manner. Consider the most general Lyapunov function

$$\mathcal{L}_t = \|x_t - x_\star\|^2 + a_t\langle x_t - x_\star, p_t\rangle + b_t\|p_t\|^2 + c_t(f(x_t) - f_\star)$$

The goal is to choose a_t, b_t, c_t such that $\dot{\mathcal{L}}_t \leqslant 0$. The derivative is

$$\begin{aligned}
\dot{\mathcal{L}}_t &= 2\langle x_t - x_\star, \dot{x}_t\rangle + \dot{a}_t\langle x_t - x_\star, p_t\rangle + a_t\langle \dot{x}_t, p_t\rangle + a_t\langle x_t - x_\star, \dot{p}_t\rangle \\
&\quad + \dot{b}_t\|p_t\|^2 + 2b_t\langle p_t, \dot{p}_t\rangle + \dot{c}_t(f(x_t) - f_\star) + c_t\langle \nabla f(x_t), \dot{x}_t\rangle \\
&= 2\langle x_t - x_\star, p_t\rangle + \dot{a}_t\langle x_t - x_\star, p_t\rangle + a_t\|p_t\|^2 - a_t\left\langle x_t - x_\star, \nabla f(x_t) + \gamma_t p_t\right\rangle \\
&\quad + \dot{b}_t\|p_t\|^2 - 2b_t\left\langle p_t, \nabla f(x_t) + \gamma_t p_t\right\rangle + \dot{c}_t(f(x_t) - f_\star) + c_t\langle \nabla f(x_t), p_t\rangle
\end{aligned}$$

Since f is convex, we have $f_\star - f(x_t) \geqslant \langle \nabla f(x_t), x_\star - x_t\rangle$, substitute into equation,

$$\begin{aligned}\dot{\mathcal{L}}_t &= 2\langle x_t - x_\star, p_t\rangle + \dot{a}_t\langle x_t - x_\star, p_t\rangle + a_t\|p_t\|^2 + a_t\langle \nabla f(x_t), x_\star - x_t\rangle - a_t\langle \gamma_t p_t, x_t - x_\star\rangle \\ &\quad + \dot{b}_t\|p_t\|^2 - 2b_t\left\langle p_t, \nabla f(x_t) + \gamma_t p_t\right\rangle + \dot{c}_t(f(x_t) - f_\star) + c_t\langle \nabla f(x_t), p_t\rangle \\ &\leqslant 2\langle x_t - x_\star, p_t\rangle + \dot{a}_t\langle x_t - x_\star, p_t\rangle + a_t\|p_t\|^2 - a_t(f(x_t) - f_\star) - a_t\langle \gamma_t p_t, x_t - x_\star\rangle \\ &\quad + \dot{b}_t\|p_t\|^2 - 2b_t\left\langle p_t, \nabla f(x_t) + \gamma_t p_t\right\rangle + \dot{c}_t(f(x_t) - f_\star) + c_t\langle \nabla f(x_t), p_t\rangle \\ &= (2 + \dot{a}_t - a_t\gamma_t)\left\langle x_t - x_\star, p_t\right\rangle + (a_t + \dot{b}_t - 2b_t\gamma_t)\|p_t\|^2 + (\dot{c}_t - a_t)(f(x_t) - f_\star) + (c_t - 2b_t)\langle \nabla f(x_t), p_t\rangle\end{aligned}$$

Since the terms $\langle x_t - x_\star, p_t\rangle$ and $\langle \nabla f(x_t), p_t\rangle$ do not have definite signs, we choose coefficients to make these terms vanish, i.e.,

$$\begin{cases}\dot{a}_t - \dfrac{3}{t}a_t + 2 = 0 \\ c_t = 2b_t\end{cases} \tag{4.9}$$

For the first equation, we have integration factor $e^{\int -\frac{3}{t}\,\mathrm{d}t} = t^{-3}$, thus

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(t^{-3}a_t\right) = -2t^{-3} \quad\Longrightarrow\quad t^{-3}a_t = t^{-2} + \bar{a} \quad\Longrightarrow\quad a_t = t + \bar{a}t^3 \tag{4.10}$$

for some $\bar{a} \geqslant 0$. Now, the remaining terms state

$$\dot{\mathcal{L}}_t \leqslant (a_t + \dot{b}_t - 2b_t\gamma_t)\|p_t\|^2 + (\dot{c}_t - a_t)(f(x_t) - f_\star)$$

To make to our goal, i.e., $\dot{\mathcal{L}}_t \leqslant 0$, since $\|p_t\|^2, (f(x_t) - f_\star) \geqslant 0$, we need both coefficients smaller than 0, so

$$\begin{cases}a_t + \dot{b}_t - \dfrac{6}{t}b_t \leqslant 0 \\ \dot{c}_t - a_t = 2\dot{b}_t - a_t \leqslant 0\end{cases} \tag{4.11}$$

where the second equation follows from the second equation of 4.9. To make these both hold, we need $\dot{b}_t \leqslant \min\left\{\frac{a_t}{2}, \frac{6b_t}{t} - a_t\right\}$. Therefore,

$$3\dot{b}_t \leqslant 2 \cdot \frac{a_t}{2} + \frac{6b_t}{t} - a_t = \frac{6b_t}{t}$$

The integrating factor is $e^{\int -\frac{2}{t}\,\mathrm{d}t} = t^{-2}$, so we solve the equation, and get

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(t^{-2}b_t\right) \leqslant 0 \quad\Longrightarrow\quad t^{-2}b_t \leqslant \bar{b} \quad\Longrightarrow\quad b_t \leqslant \bar{b}t^2 \tag{4.12}$$

We consider the form $b_t = \bar{b}t^2$ for some $\bar{b} \geqslant 0$. To find $\bar{a}$ and $\bar{b}$, we come back to equation 4.11, and see that

$$\begin{cases}t + \bar{a}t^3 + 2\bar{b}t - \dfrac{6}{t}\bar{b}t^2 = \bar{a}t^3 + (-4\bar{b} + 1)t \leqslant 0 \\ (4\bar{b} - 1)t - \bar{a}t^3 \leqslant 0\end{cases}$$

The first equation yields $\bar{a} = 0$ and $\bar{b} \geqslant 1/4$. The second equation yields $\bar{b} \leqslant 1/4$. So we must take $\bar{a} = 0$ and $\bar{b} = 1/4$. With these choices, we finally have

$$a_t = t, \quad b_t = \frac{1}{4}t^2, \quad c_t = \frac{1}{2}t^2$$

Note that in this case $b_t = a_t^2/4$, so that in the Lyapunov function,

$$\begin{aligned}\mathcal{L}_t &= \|x_t - x_\star\|^2 + a_t\langle x_t - x_\star, p_t\rangle + b_t\|p_t\|^2 + c_t(f(x_t) - f_\star)\\ &= \left\|x_t - x_\star + \frac{a_t}{2}p_t\right\|^2 + c_t(f(x_t) - f_\star) \geqslant c_t(f(x_t) - f_\star)\end{aligned}$$

We get the Lyapunov function

$$\mathcal{L}_t = \left\|x_t - x_\star + \frac{t}{2}p_t\right\|^2 + \frac{t^2}{2}(f(x_t) - f_\star) = \|z_t - x_\star\|^2 + \frac{t^2}{2}(f(x_t) - f_\star)$$

which recovers the one in the proof of Theorem 4.2.1.

The strongly convex case is similar. We omit the process to find the correct Lyapunov function as above, and directly state it in the proof.

Theorem 4.2.2: Convergence of AGF under strong convexity

Let $f : \mathbb{R}^d \to \mathbb{R}$ be α-convex and let $(x_t)_{t\geqslant 0}$ evolve along AGF with $\gamma_t = 2\sqrt{\alpha}$ (does not depend on t), and $p_0 = 0$. For all $t \geqslant 0$,

$$f(x_t) - f_\star \leqslant 2\exp(-\sqrt{\alpha}t)(f(x_0) - f_\star)$$

Proof. Consider the auxiliary point $z_t = x_t + \frac{2}{\gamma}p_t$ and consider the Lyapunov function

$$\mathcal{L}_t = f(x_t) - f_\star + \frac{\alpha}{2}\|z_t - x_\star\|^2$$

Now we compute the derivative,

$$\begin{aligned}\dot{\mathcal{L}}_t &= \langle \nabla f(x_t), \dot{x}_t\rangle + \alpha\langle z_t - x_\star, \dot{z}_t\rangle\\ &= \langle \nabla f(x_t), p_t\rangle + \alpha\left\langle x_t + \frac{2}{\gamma}p_t - x_\star, p_t - \frac{2}{\gamma}\nabla f(x_t) - 2p_t\right\rangle && \text{(Definition of AGF)}\\ &= \langle \nabla f(x_t), p_t\rangle - \alpha\left\langle x_t + \frac{2}{\gamma}p_t - x_\star, p_t + \frac{2}{\gamma}\nabla f(x_t)\right\rangle\\ &= \left(1 - \frac{4\alpha}{\gamma^2}\right)\langle \nabla f(x_t), p_t\rangle - \alpha\left\langle x_t - x_\star, p_t + \frac{2}{\gamma}\nabla f(x_t)\right\rangle - \frac{2\alpha}{\gamma}\|p_t\|^2\\ &= -\alpha\langle x_t - x_\star, p_t\rangle - \sqrt{\alpha}\langle x_t - x_\star, \nabla f(x_t)\rangle - \sqrt{\alpha}\|p_t\|^2 && (\gamma = 2\sqrt{\alpha})\\ &\leqslant -\alpha\langle x_t - x_\star, p_t\rangle - \alpha\left\langle \frac{2}{\gamma}p_t, p_t\right\rangle - \sqrt{\alpha}(f(x_t) - f_\star) - \sqrt{\alpha}\frac{\alpha}{2}\|x_t - x_\star\|^2 && (\alpha\text{-Convexity})\\ &= -\alpha\langle z_t - x_\star, p_t\rangle - \sqrt{\alpha}\left(f(x_t) - f_\star + \frac{\alpha}{2}\left\|z_t - x_\star - \frac{2}{\gamma}p_t\right\|^2\right)\\ &= -\alpha\langle z_t - x_\star, p_t\rangle - \sqrt{\alpha}\left(f(x_t) - f_\star + \frac{\alpha}{2}\left(\|z_t - x_\star\|^2 - 2\left\langle z_t - x_\star, \frac{2}{\gamma}p_t\right\rangle + \left\|\frac{2}{\gamma}p_t\right\|^2\right)\right)\\ &= -\sqrt{\alpha}\left(f(x_t) - f_\star + \frac{\alpha}{2}\|z_t - x_\star\|^2\right) - \sqrt{\alpha}\frac{\alpha}{2}\left\|\frac{2}{\gamma}p_t\right\|^2\\ &\leqslant -\sqrt{\alpha}\mathcal{L}_t\end{aligned}$$

By continuous Grönwall's inequality, with $A = -\sqrt{\alpha}, B = 0$, we have

$$\mathcal{L}_t \leqslant \mathcal{L}_0 \exp(-\sqrt{\alpha} t)$$

Substitute the formula of $\mathcal{L}_t$, we have

$$\begin{aligned} f(x_t) - f_\star \leqslant f(x_t) - f_\star + \frac{\alpha}{2}\|z_t - x_\star\|^2 &\leqslant \exp(-\sqrt{\alpha}t)\left(f(x_0) - f_\star + \frac{\alpha}{2}\|x_0 - x_\star\|^2\right) && (p_0 = 0) \\ &\leqslant \exp(-\sqrt{\alpha}t)\left(f(x_0) - f_\star + f(x_0) - f_\star - \langle \underbrace{\nabla f(x_\star)}_{=0}, x_0 - x_\star\rangle\right) && \text{(Convexity)} \\ &= 2\exp(-\sqrt{\alpha}t)(f(x_0) - f_\star) \end{aligned}$$

which is the desired result. □

The table below shows the comparison:

	0-convex	α-convex
GF	$1/t$	$\exp(-2\alpha t)$
AGF	$1/t^2$	$\exp(-\sqrt{\alpha} t)$

It is strongly suggestive of the square root factor speed-up, i.e., acceleration. However, it is dangerous to deduce result in continuous time. For example, we can run continuous ODE faster, but may make it more unstable, therefore need a smaller step size for discretization.

4.3 General Case: Discrete Time

We could consider the discretization

$$\begin{aligned} x_{n+1} &\approx x_n + hp_{n+1} \\ p_{n+1} &\approx p_n - h\nabla f(x_n) - \gamma_n h p_n \end{aligned}$$

which is equivalent to the update

$$x_{n+1} = x_n - h^2 \nabla f(x_n) + (1 - \gamma_n h)(x_n - x_{n-1})$$

Or, if we left the coefficients unknown,

$$x_{n+1} = x_n - \eta_n \nabla f(x_n) + \theta_n (x_n - x_{n-1})$$

In other words, we take a gradient step and then apply momentum. This is known as **Polyak's heavy ball** method. Although it can be tuned to converge at the rate of conjugate gradient for quadratic objectives, this same tuning leads to divergence for general convex functions, see, Lessard et al.[6]. Or, we can *add momentum and then take a gradient step*,

$$x_{n+1} = x_n + \theta_n(x_n - x_{n-1}) - \eta_n \nabla f(x_n + \theta_n(x_n - x_{n-1}))$$

This is called the **Nesterov's accelerated gradient method**. We analyze this method with $x_{-1} = x_0$, and $\eta_n = 1/\beta$, thus

$$x_{n+1} = x_n + \theta_n(x_n - x_{n-1}) - \frac{1}{\beta}\nabla f(x_n + \theta_n(x_n - x_{n-1})) \tag{AGD}$$

Theorem 4.3.1: Convergence of AGD

Let f be convex and β-smooth. Define the sequence $\lambda_0 = 0$ and $\lambda_{n+1} = \frac{1}{2}(1 + \sqrt{1 + 4\lambda_n^2})$ for $n \in \mathbb{N}$. Set $\theta_n = (\lambda_n - 1)/\lambda_{n+1}$. Then, the accelerate gradient descent satisfies

$$f(x_N) - f_\star \leqslant \frac{2\beta\|x_0 - x_\star\|^2}{N^2}$$

Proof. $y_n = x_n + \theta_n(x_n - x_{n-1})$, so that $x_{n+1} = y_n - \frac{1}{\beta}\nabla f(y_n)$, mimicking a gradient descent. Recall from Equation 2.6 that for any point $z \in \mathbb{R}^d$, it holds that

$$\|x_{n+1} - z\|^2 \leqslant \|y_n - z\|^2 - \frac{2}{\beta}(f(x_{n+1}) - f(z))$$

Rearranging, we have

$$\begin{aligned} f(x_{n+1}) - f(z) &\leqslant \frac{\beta}{2}\left(\|y_n - z\|^2 - \|x_{n+1} - z\|^2\right) = \frac{\beta}{2}\left(\|y_n - z\|^2 - \|x_{n+1} - y_n + y_n - z\|^2\right) \\ &= \frac{\beta}{2}\left(\|y_n - z\|^2 - \|x_{n+1} - y_n\|^2 - \|y_n - z\|^2 - 2\langle x_{n+1} - y_n, y_n - z\rangle\right) \\ &= -\frac{\beta}{2}\|x_{n+1} - y_n\|^2 - \beta\langle x_{n+1} - y_n, y_n - z\rangle \end{aligned}$$

We apply this inequality with two points: $z = x_n$ and $z = x_\star$, so

$$\begin{aligned} f(x_{n+1}) - f(x_n) &\leqslant -\frac{\beta}{2}\|x_{n+1} - y_n\|^2 - \beta\langle x_{n+1} - y_n, y_n - x_n\rangle \\ f(x_{n+1}) - f_\star &\leqslant -\frac{\beta}{2}\|x_{n+1} - y_n\|^2 - \beta\langle x_{n+1} - y_n, y_n - x_\star\rangle \end{aligned}$$

Multiplying the first inequality by $\lambda_{n+1} - 1 \geqslant 0$ and adding it to the second, it implies

$$\begin{aligned} (\lambda_{n+1} - 1)(f(x_{n+1}) - f(x_n)) + f(x_{n+1}) - f_\star &\leqslant -\frac{\beta\lambda_{n+1}}{2}\|x_{n+1} - y_n\|^2 - \beta\langle x_{n+1} - y_n, \lambda_{n+1}y_n - (\lambda_{n+1} - 1)x_n - x_\star\rangle \\ &= -\frac{\beta}{2\lambda_{n+1}}\|\underbrace{\lambda_{n+1}(x_{n+1} - y_n)}_{=a}\|^2 - \frac{\beta}{\lambda_{n+1}}\langle\underbrace{\lambda_{n+1}(x_{n+1} - y_n)}_{=a}, \underbrace{\lambda_{n+1}y_n - (\lambda_{n+1} - 1)x_n - x_\star}_{=b}\rangle \\ &= \frac{\beta}{2\lambda_{n+1}}\left(\|\lambda_{n+1}y_n - (\lambda_{n+1} - 1)x_n - x_\star\|^2 - \|\lambda_{n+1}x_{n+1} - (\lambda_{n+1} - 1)x_n - x_\star\|^2\right) \end{aligned}$$

where the last line uses the identity $\|a\|^2 + 2\langle a, b\rangle = \|a + b\|^2 - \|b\|^2$. Our goal is to produce a telescoping sum, which is the case if we assume

$$\lambda_{n+1}x_{n+1} - (\lambda_{n+1} - 1)x_n = \lambda_{n+2}y_{n+1} - (\lambda_{n+2} - 1)x_{n+1}$$

We can see that

$$(\lambda_{n+1} + \lambda_{n+2} - 1)x_{n+1} = \lambda_{n+2}(x_{n+1} + \theta_{n+1}(x_{n+1} - x_n)) + (\lambda_{n+1} - 1)x_n$$

$$\begin{aligned}
&\implies (\lambda_{n+1} - \theta_{n+1}\lambda_{n+2} - 1)x_{n+1} = (\lambda_{n+1} - \theta_{n+1}\lambda_{n+2} - 1)x_n \\
&\implies \theta_{n+1} = (\lambda_{n+1} - 1)/\lambda_{n+2}
\end{aligned}$$

Therefore, multiplying the above inequality by λ_{n+1} and summing over n, we will have

$$\begin{aligned}
&\sum_{n=0}^{N-1} (\lambda_{n+1}(f(x_{n+1}) - f_\star) + \lambda_{n+1}(\lambda_{n+1} - 1)(f(x_{n+1}) - f(x_n))) \\
=&\sum_{n=0}^{N-1} (\lambda_{n+1}(f(x_{n+1}) - f_\star) + \lambda_{n+1}(\lambda_{n+1} - 1)(f(x_{n+1}) - f_\star + f_\star - f(x_n))) \\
=&\sum_{n=0}^{N-1} \left(\lambda_{n+1}^2(f(x_{n+1}) - f_\star) - \lambda_{n+1}(\lambda_{n+1} - 1)(f(x_n) - f_\star)\right) \\
\leqslant& \frac{\beta}{2}\sum_{n=0}^{N-1} \left(\|\lambda_{n+1}y_n - (\lambda_{n+1} - 1)x_n - x_\star\|^2 - \|\lambda_{n+1}x_{n+1} - (\lambda_{n+1} - 1)x_n - x_\star\|^2\right) \\
=& \frac{\beta}{2}\left(\|\lambda_1 y_0 - (\lambda_1 - 1)x_0 - x_\star\|^2 - \|\lambda_N x_N - (\lambda_N - 1)x_{N-1} - x_\star\|^2\right) && \text{(Telescoping)} \\
\leqslant& \frac{\beta}{2}\|\lambda_1 y_0 - (\lambda_1 - 1)x_0 - x_\star\|^2
\end{aligned}$$

We also want LHS telescoping. So we set $\lambda_n^2 = \lambda_{n+1}(\lambda_{n+1} - 1)$, which yields the recursion $\lambda_{n+1} = \frac{1}{2}\left(1 + \sqrt{1 + 4\lambda_n^2}\right)$ (we take the positive root so that λ is increasing). With $\lambda_0 = 0$, it yields $\lambda_1 = \frac{1}{2}(1 + \sqrt{1}) = 1$, and

$$\begin{aligned}
&\sum_{n=0}^{N-1} \left(\lambda_{n+1}^2(f(x_{n+1}) - f_\star) - \lambda_{n+1}(\lambda_{n+1} - 1)(f(x_n) - f_\star)\right) \\
=&\lambda_N^2(f(x_N) - f_\star) - \lambda_1(\lambda_1 - 1)(f(x_0) - f_\star) && \text{(Telescoping)} \\
=&\lambda_N^2(f(x_N) - f_\star) \leqslant \frac{\beta}{2}\|y_0 - x_\star\|^2 = \frac{\beta}{2}\|x_0 - x_\star\|^2 && (\lambda_1 = 1,\ x_{-1} = x_0)
\end{aligned}$$

This leads to

$$f(x_N) - f_\star \leqslant \frac{\beta\|x_0 - x_\star\|^2}{2\lambda_N^2}$$

Finally, we prove inductively that $\lambda_N \geqslant N/2$. The base case $N = 0$ and $N = 1$ satisfies obviously. Suppose that this is true for all $1, 2, \cdots, N$. Then,

$$\lambda_{N+1} = \frac{1}{2}\left(1 + \sqrt{1 + 4\lambda_N^2}\right) \geqslant \frac{1}{2}\left(1 + \sqrt{1 + N^2}\right) \geqslant N/2$$

Therefore, we get the final bound:

$$f(x_N) - f_\star \leqslant \frac{\beta\|x_0 - x_\star\|^2}{2\lambda_N^2} \leqslant \frac{2\beta\|x_0 - x_\star\|^2}{N^2}$$

which completes the proof. □

Remark:

- By applying the reduction in Lemma 3.1.1, we can get the bound for strongly convex case, i.e., $\phi(x) \asymp \sqrt{x}$, thus leading to a $O(\sqrt{\kappa}\log\frac{\alpha R^2}{\epsilon})$ complexity.

- Consider the gradient descent with changing step sizes $x_{n+1} = x_n - h\nabla f(x_n)$, the so-called *silver step size* schedule achieves the rate of Lemma 3.1.1 and Lemma 3.1.2 with $\phi(x) = x^{\log_\rho 2} \approx x^{0.786}$ with $\rho = 1+\sqrt{2}$, see Altschuler and Parrilo[1, 2]. This is a rate intermediate between the unaccelerate rate of gradient descent and the acceleerated rate of accelerated gradient descent.

Chapter 5

Non-smooth Convex Optimization

We will consider *constrained* and *non-smooth* optimization in this chapter. There are certain reasons to tackle these two together.

First, to minimize f over a convex set $\mathcal{C}$, it is equivalent to minimize $f + \chi_{\mathcal{C}}$ over all of $\mathbb{R}^d$, where $\chi_{\mathcal{C}}$ is the convex indicator function for $\mathcal{C}$:

$$\chi_{\mathcal{C}} := \begin{cases} 0, & x \in \mathcal{C}, \\ \infty, & x \notin \mathcal{C} \end{cases}$$

thus, a constrained problem is equivalent to a non-smooth problem with f allowed to taking value ∞.

Second, even if we do not formulate in this way, we have constraints $\mathcal{C} = \{f_i \leqslant 0 \text{ for all } i \in [m]\}$ will have the form $\mathcal{C} = \{\max_{i \in [m]} f_i \leqslant 0\}$, where $\max_{i\in[m]} f_i$ is a non-smooth function.

5.1 Convex Analysis

5.1.1 Lower Semicontinuity

Definition 5.1.1: Epigraph

The **epigraph** of $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is the subset of $\mathbb{R}^d \times \mathbb{R}$ defined as

$$\operatorname{epi} f := \{(x, t) \in \mathbb{R}^d \times \mathbb{R} : f(x) \leqslant t\}$$

Remark: f is convex if and only if $\operatorname{epi} f$ is a convex set.

Definition 5.1.2: Domain

The **domain** of a function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is the set

$$\operatorname{dom} f := \{x \in \mathbb{R}^d : f(x) < \infty\}$$

Convex f can still be quite pathological. Consider the following function:

$$f(x) = \begin{cases} 0, & \|x\| < 1, \\ \phi(x), & \|x\| = 1, \\ \infty, & \|x\| > 1 \end{cases} \tag{5.1}$$

where ϕ is an arbitrary non-negative function. Then, f is convex, but ϕ can be extremely pathological. To avoid this, the basic regularity property is that f is lower semicontinuous.

Definition 5.1.3: Lower Semicontinuous

A function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is **lower semicontinuous** if for all sequences $\{x_n\}_{n\in\mathbb{N}}$ converging to a point $x \in \mathbb{R}^d$, it holds that

$$f(x) \leqslant \liminf_{n\to\infty} f(x_n)$$

In other words, when we pass to the limit of a convergent sequence, the value of f can only drop down. One way to motivate this is that, we often consider suprema $f = \sup_{\omega\in\Omega} f_\omega$, where $\{f_\omega\}_{\omega\in\Omega}$ is a collection of continuous functions. When Ω is finite, the supremum is continuous. However, when Ω is infinite, it may not be continuous. The class of lower semicontinuous function is the smallest class of functions which contains all continuous functions and is closed under taking arbitrary suprema. More properties are explored below.

Proposition 5.1.4: Properties of Lower Semicontinuous Function

1. A function f is lower semicontinuous if and only if for all $c \in \mathbb{R}$, the level set $\{f \leqslant c\}$ is closed.

2. A supremum of lower semicontinuous functions is lower semicontinuous.

3. The function 5.1 is lower semicontinuous if and only if $\phi = 0$.

Proof. 1. ($\Longrightarrow$) Suppose f is lower semicontinuous. Consider the sequence $\{x_n\}_{n\in\mathbb{N}}$ where each $x_n \in \{f \leqslant c\}$, that is $f(x_n) \leqslant c$ for all $n \in \mathbb{N}$, and $x_n \to x$. Then, by lower semicontinuous, we have

$$f(x) \leqslant \liminf_{n\to\infty} f(x_n) \leqslant c$$

This shows that $x \in \{f \leqslant c\}$. Since $\{x_n\}$ is arbitrary, $\{f \leqslant c\}$ is closed.
($\Longleftarrow$) Suppose that the level sets $\{f \leqslant c\}$ are closed. Define $\{x_n\}_{n\in\mathbb{N}}$ an arbitrary sequence in $\mathbb{R}^d$ converging to x. Define $L := \liminf_{n\to\infty} f(x_n)$. We need to show that $f(x) \leqslant L$. By the definition of lim inf, for any $\epsilon > 0$, there exists a subsequence $\{x_{n_k}\}$ such that $f(x_{n_k}) \leqslant L + \epsilon$. Then, each x_{n_k} belongs to $\{f \leqslant L + \epsilon\}$. Since this set is closed by assumption, the limit $x \in \{f \leqslant L + \epsilon\}$, which means that $f(x) \leqslant L + \epsilon$. Since ϵ is arbitrary, we have $f(x) \leqslant L$, which completes the proof.

2. Let $\{f_\omega\}_{\omega\in\Omega}$ be a collection of lower semicontinuous functions. Let $f(x) = \sup_{\omega\in\Omega} f_\omega(x)$. The goal is to show that $f(x)$ is lower semicontinuous. Note that by the proof of 1, we only need to show that $\{f \leqslant c\}$ is closed for all $c \in \mathbb{R}$. We can write

$$\{f(x) \leqslant c\} = \bigcap_{\omega\in\Omega} \{x \in \mathbb{R}^d : f_\omega(x) \leqslant c\}$$

However, $f_\omega(x)$ are lower semicontinuous, so each set at RHS is closed. Since intersection of arbitrary sets is closed, we conclude that the level set of $f(x)$ is closed, i.e., it is lower semicontinuous.

3. ($\Longrightarrow$) Suppose $f(x)$ is lower semicontinuous. We can construct a convergent sequence $\{x_n\}_{n\in\mathbb{N}}$ with limit x such that $\|x_n\| < 1$ for all $n \in \mathbb{N}$ and $\|x\| = 1$. Then, since f is lower semicontinuous, we have $f(x) \leqslant \liminf_{n\to\infty} f(x_n) = 0$. f is nonnegative, so we must have $f(x) = 0$. Since this is true for arbitrary x, we have $\phi(x) = 0$.
   ($\Longleftarrow$) Suppose $\phi(x) = 0$. Then, $\{f \leqslant c\} = \emptyset$ for all $c < 0$; $\{f \leqslant c\} = \bar{B}_1(0)$ for all $0 \leqslant c < \infty$, where $\bar{B}_1(0)$ is the closed ball with radius 1 centered at 0. For $c = \infty$, we have $\{f \leqslant c\} = \mathbb{R}^d$. In all three cases, the level sets are closed. By 1, f is lower semicontinuous.

□

Follow from the proposition, f **is convex and lower semicontinuous if and only if its epigraph is closed and convex**. So, *when it comes to functions, we impose convexity and lower semicontinuity*; and *when it comes to sets, we impose convexity and closedness.* The following terminology is not standard.

Definition 5.1.5: Regular Function

A convex function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is **regular** if it is not identically equal to ∞, it is lower semicontinuous, and its domain has non-empty interior.

Since the domain of a convex function is a convex set, if it has empty interior then it must be contained in a lower dimensional affine space, and when we restrict to that space, the domain then has a non-empty interior; this is usually summarized by saying that any non-empty convex set has a non-empty *relative interior*. This is why we regard the condition that the domain has non-empty interior as “without loss of generality”.

We also note that in the proof of existence of a minimizer, it is really only lower semicontinuity that matters.

Lemma 5.1.6: Existence of Minimizer

Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be lower semicontinuous and its level sets be bounded. Then, there exists a global minimizer of f.

Proof. The proof is similar to that of Lemma 1.3.4. Since f is lower semicontinuous, the level sets are closed. Since its level sets are assumed to be bounded, they are compact. Let $\{x_n\}_{n\in\mathbb{N}}$ be a minimizing sequence, $f(x_n) \to \inf f$. By compactness, it admits a convergent subsequence $\{x_{n_q}\}_{n_q\in\mathbb{N}}$, which converges to $x_\star \in \mathbb{R}^d$. By lower semicontinuity, we have $f(x_\star) \leqslant \lim_{n\to\infty} f(x_{n_q}) = \inf f$, and this indicates $f(x_\star) = \inf f$. □

5.1.2 Subdifferential

We establish properties of regular convex functions in this subsection.

Lemma 5.1.7: Lipschitz Continuity

Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be convex and let $x_0 \in \operatorname{int}\operatorname{dom} f$. Then, f is locally Lipschitz continuous around x_0. i.e., $\exists\, L > 0$, s.t. $\forall\, x, y \in N_\epsilon(x_0)$, where $N_\epsilon(x_0)$ denotes a neighbourhood of x_0, we have $|f(x) - f(y)| \leqslant L\|x - y\|$.

Proof. By translating f, we may assume that $x_0 = 0$ without loss of generality. Since $0 \in \operatorname{int}\operatorname{dom} f$, we can find a simplex centered at the origin inside the domain, i.e., there exists $\epsilon > 0$ such that $\mathcal{C} = \operatorname{conv}\{\pm\epsilon e_k : k \in [d]\} \subseteq \operatorname{dom} f$.

We first show that f is bounded on $\mathcal{C}$.
Upper bound: Note that $f(\pm\epsilon e_k) < \infty$ for all $k \in [d]$ since they are in the domain of f. Moreover, denote the vertices $\{\epsilon e_1, \cdots, \epsilon e_d, -\epsilon e_1, \cdots, -\epsilon e_d\} = \{v_1, \cdots, v_{2d}\}$, then for any $x = \sum_{i=1}^{2d} \lambda_i v_i$, where $\sum_{i=1}^{2d} \lambda_i = 1$, $\lambda_i \geqslant 0$, we have

$$f(x) = f\left(\sum_{i=1}^{2d} \lambda_i v_i\right) \leqslant \sum_{i=1}^{2d} \lambda_i f(v_i) \leqslant \max\{f(\epsilon e_1), \cdots, f(\epsilon e_d), f(-\epsilon e_1), \cdots, f(-\epsilon e_d)\}$$

by convexity. Therefore, the maximum of f over $\mathcal{C}$ is attained at one of the vertices. So we have a finite upper bound.
Lower bound: By convexity, $f(x) \geqslant 2f(0) - f(-x) \geqslant 2f(0) - \max_{\mathcal{C}} f$ for all $x \in \mathcal{C}$.

Next, we show that f is Lipschitz on the smaller set $\mathcal{C}' := \operatorname{conv}\{\pm\frac{\epsilon}{2} e_k : k \in [d]\}$. The point is that there is a constant $c_{d,\epsilon} > 0$ such that for all $x, y \in \mathcal{C}'$, there is a point $y^+ \in \mathcal{C}$ such that the line segment from x to y is contained in the line segment from x to y^+, and the extension is not too short: $\|y^+ - x\| \geqslant c_{d,\epsilon}$. Then, by convexity,

$$f(y) = f\left(\frac{\|y^+ - y\|}{\|y^+ - x\|} x + \frac{\|y - x\|}{\|y^+ - x\|} y^+\right) \leqslant \frac{\|y^+ - y\|}{\|y^+ - x\|} f(x) + \frac{\|y - x\|}{\|y^+ - x\|} f(y^+)$$

hence,

$$\begin{aligned} f(y) - f(x) &\leqslant \frac{\|y^+ - y\| - \|y^+ - x\|}{\|y^+ - x\|} f(x) + \frac{\|y - x\|}{\|y^+ - x\|} f(y^+) \\ &\leqslant \frac{\|y - x\|}{\|y^+ - x\|}(f(y^+) - f(x)) && \text{(Triangular inequality)} \\ &\leqslant \frac{\sup_{\mathcal{C}} f - \inf_{\mathcal{C}} f}{c_{d,\epsilon}} \|y - x\| \end{aligned}$$

Interchanging x and y proves the Lipschitz bound. □

This shows that locally near x_0, $f(x)$ grows at most linearly in the distance $\|x - x_0\|$. This suggests that f is 'almost' differentiable at x_0. Anyways, we can find an appropriate substitute for differentiability.

Definition 5.1.8: Subgradient/Subdifferential

- Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be convex. We say that $p \in \mathbb{R}^d$ is a **subgradient** of f at x if for all $y \in \mathbb{R}^d$, we have
$$f(y) \geqslant f(x) + \langle p, y - x\rangle \tag{5.2}$$
- Denote the set of subgradients of f at x as $\partial f(x)$, called **subdifferential** of f at x, we also set
$$\partial f = \{(x, p) \in \mathbb{R}^d \times \mathbb{R}^d : p \in \partial f(x)\}$$

Remark:

- By definition, if $0 \in \partial f(x)$, then x is a global minimizer of f.

- If f is differentiable at $x_0 \in \operatorname{int}\operatorname{dom} f$, then $\partial f(x_0)$ is a singleton: $\partial f(x_0) = \{\nabla f(x_0)\}$. To prove this, first, by convexity, $f(y) \geqslant f(x_0) + \langle \nabla f(x_0), y - x_0 \rangle$, so $\nabla f(x_0) \in \partial f(x_0)$. Second, if $g \in \partial f(x_0)$, then

  $$f(y) - f(x_0) \geqslant \langle g, y - x_0 \rangle,\ f(x_0) - f(y) \geqslant \langle \nabla f(x_0), x_0 - y \rangle \quad \Longrightarrow \quad \langle \nabla f(x_0) - g, u \rangle \geqslant 0 \text{ for any direction } u$$

  This has to hold for all u, and so necessarily $g = \nabla f(x_0)$. This proves the uniqueness.

- Subdifferential can be multi-valued. The simplest example is $f : x \mapsto |x|$, for which $\partial f(0) = [-1, 1]$.

5.1.3 Constraint and Projection

When the constraint set $\mathcal{C}$ is simple, it is reasonable to suppose that we can compute the projection onto $\mathcal{C}$. We study some properties of this projection operator.

Definition 5.1.9: Projection

Let $\mathcal{C}$ be closed and convex. The **projection** onto $\mathcal{C}$ is the mapping $\pi_{\mathcal{C}} : \mathbb{R}^d \to \mathcal{C}$ defined by

$$\pi_{\mathcal{C}}(x) = \operatorname*{argmin}_{y \in \mathcal{C}} \|y - x\|^2$$

The existence of this argmin follows from that $\mathcal{C}$ is closed, and the uniqueness of minimizer follows from Lemma 1.3.8 since the objective function is strictly convex. When $\mathcal{C}$ is a linear subspace, $\pi_{\mathcal{C}}$ is linear. But generally, $\pi_{\mathcal{C}}$ is a non-linear operator.

Lemma 5.1.10: Characterization of Projection

Let $\mathcal{C}$ be closed and convex, and let $x \notin \mathcal{C}$. Then, $\pi_{\mathcal{C}}(x)$ is the unique point satisfying the following condition:

$$\langle \pi_{\mathcal{C}}(x) - x, x' - \pi_{\mathcal{C}}(x) \rangle \geqslant 0, \qquad \text{for all } x' \in \mathcal{C}$$

Proof. As in the proof of Lemma 1.3.5, the first-order necessary condition for optimality indicates

$$\langle \nabla \|\pi_{\mathcal{C}}(x) - x\|^2, v \rangle \geqslant 0 \quad \Longrightarrow \quad \langle 2(\pi_{\mathcal{C}}(x) - x), v \rangle \geqslant 0 \quad \Longrightarrow \quad \langle \pi_{\mathcal{C}}(x) - x, v \rangle \geqslant 0$$

for all $v \in \mathbb{R}^n$. However, because the optimization problem is constrained to lie in $\mathcal{C}$, this time we do not have the inequality for all v, but only for v of the directions $x' - \pi_{\mathcal{C}}(x)$ where $x' \in \mathcal{C}$. □

Lemma 5.1.11: Convex Projections are Non-expansive

Let $\mathcal{C}$ be closed and convex. Then, for all $x, y \in \mathbb{R}^d$,

$$\|\pi_{\mathcal{C}}(y) - \pi_{\mathcal{C}}(x)\| \leqslant \|y - x\|$$

Proof. By Lemma 5.1.10, we have

$$\langle \pi_{\mathcal{C}}(x) - x, \pi_{\mathcal{C}}(y) - \pi_{\mathcal{C}}(x) \rangle \geqslant 0$$

$$\langle \pi_{\mathcal{C}}(y) - y, \pi_{\mathcal{C}}(x) - \pi_{\mathcal{C}}(y) \rangle \geqslant 0$$

Adding these together, we have

$$\begin{aligned} & \langle \pi_{\mathcal{C}}(y) - \pi_{\mathcal{C}}(x), y - x + \pi_{\mathcal{C}}(x) - \pi_{\mathcal{C}}(y) \rangle \geqslant 0 \\ \implies \quad & \|\pi_{\mathcal{C}}(y) - \pi_{\mathcal{C}}(x)\|^2 \leqslant \langle \pi_{\mathcal{C}}(y) - \pi_{\mathcal{C}}(x), y - x \rangle \leqslant \|\pi_{\mathcal{C}}(y) - \pi_{\mathcal{C}}(x)\| \|y - x\| \end{aligned}$$

which completes the proof. □

5.1.4 Existence of Subdifferential

For the purpose of optimization, it is enough to have at least one subgradient. When this is true? To answer this, we first need two lemma.

Lemma 5.1.12: Lipschitz Continuity and Boundedness of Subgradient

Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuous and convex on a convex set $\mathcal{C}$. Then, f is Lipschitz continuous over $\mathcal{C}$ with constant L if and only if for every $x_0 \in \operatorname{int} \mathcal{C}$ and every $p \in \partial f(x_0)$, we have $\|p\| \leqslant L$.

Proof. ($\Longrightarrow$) Suppose f is Lipschitz continuous over $\mathcal{C}$. Then,

$$\forall\, x, y \in \mathcal{C}, \quad |f(x) - f(y)| \leqslant L\|x - y\|$$

Let $x_0 \in \operatorname{int} \mathcal{C}$ and $p \in \partial f(x_0)$. By definition of subdifferential, we have

$$f(y) \geqslant f(x_0) + \langle p, y - x_0 \rangle, \quad \forall\, y \in \mathcal{C}$$

Now, choose a unit vector u with $\|u\| = 1$ and choose a small $t > 0$ such that $x_0 + tu \in \mathcal{C}$. This is possible since x_0 is an interior point. Then,

$$f(x_0 + tu) \geqslant f(x_0) + t\langle p, u \rangle$$

The Lipschitz result can be written as $f(x_0 + tu) \leqslant f(x_0) + L\|tu\| = f(x_0) + Lt$ for $x = x_0 + tu$ and $y = x_0$. Combining these two, we have

$$f(x_0) + t\langle p, u \rangle \leqslant f(x_0) + Lt \quad \Longrightarrow \quad \langle p, u \rangle \leqslant L$$

Applying the same argument with $-u$, we can show that $-\langle p, u \rangle \leqslant L$, so $|\langle p, u \rangle| \leqslant L$. Therefore, we have

$$\|p\| = \sup_{\|u\|=1} |\langle p, u \rangle| \leqslant L$$

($\Longleftarrow$) Suppose for every $x_0 \in \operatorname{int} \mathcal{C}$ and every $p \in \partial f(x_0)$, we have $\|p\| \leqslant L$. We have

$$f(y) - f(x_0) \geqslant \langle p, y - x_0 \rangle \geqslant -\|p\| \|y - x_0\| \geqslant -L\|y - x_0\|$$

Changing the role of y and x_0, we have

$$f(x_0) - f(y) \geqslant -L\|y - x_0\|$$

Combining these two, we have $|f(y) - f(x_0)| \leqslant L\|y - x_0\|$, it is L-Lipschitz. □

Remark: In the proof here, the continuity assumption is only needed for the existence of subdifferential. So when we use it later, it can be ignored if the existence has been proven.

Lemma 5.1.13: Supporting Hyperplane

Let $\mathcal{C}$ be a closed and convex set, and let $x \in \partial\mathcal{C}$. Then, there exists a non-zero $p \in \mathbb{R}^d$ such that

$$\langle p, x\rangle \leqslant \inf_{x' \in \mathcal{C}} \langle p, x'\rangle$$

Proof. We first show that, if $\mathcal{C}$ is closed and convex, choose $x \notin \mathcal{C}$, we can separate $\mathcal{C}$ from x. Namely, by 5.2, the vector $p = \pi_{\mathcal{C}}(x) - x$ is non-zero and satisfies

$$\begin{aligned}
&\langle \pi_{\mathcal{C}}(x) - x, x' - \pi_{\mathcal{C}}(x)\rangle \geqslant 0 \quad \forall x' \in \mathcal{C} \quad \Longrightarrow \quad \langle p, x'\rangle \geqslant \langle p, \pi_{\mathcal{C}}(x)\rangle \quad \forall x' \in \mathcal{C} \quad \Longrightarrow \quad \inf_{x' \in \mathcal{C}} \langle p, x'\rangle \geqslant \langle p, \pi_{\mathcal{C}}(x)\rangle \\
\Longrightarrow \quad &\inf_{x' \in \mathcal{C}} \langle p, x'\rangle \geqslant \langle p, \pi_{\mathcal{C}}(x)\rangle = \langle \pi_{\mathcal{C}}(x) - x, \pi_{\mathcal{C}}(x) - x + x\rangle = \|\pi_{\mathcal{C}}(x) - x\|^2 + \langle p, x\rangle \geqslant \langle p, x\rangle
\end{aligned}$$

To prove the theorem, note that since $x \in \partial\mathcal{C}$, there is a sequence of points $\{x_n\}_{n\in\mathbb{N}}$ which lies outside of $\mathcal{C}$, such that $x_n \to x$. For each n, let p_n be the hyperplane that separates $\mathcal{C}$ from x_n. By normalizing, we may assume that $\|p_n\| = 1$. Since $\{p_n\}_{n\in\mathbb{N}}$ is a bounded sequence, it contains a subsequence which converges to some unit vector p. By taking limit, since inner product is a continuous function, the above inequality is still satisfied. □

Now we state the theorem.

Theorem 5.1.14: Existance of Subdifferential

Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a regular convex function. If $x_0 \in \operatorname{int}\operatorname{dom} f$, then $\partial f(x_0)$ is non-empty, bounded, convex and closed.

Proof. We first deal with non-emptyness. Since $(x_0, f(x_0)) \in \partial\mathrm{epi}f$, and $\mathrm{epi}f$ is closed and convex since f is regular. Therefore, there is a supporting hyperplane (p, q) such that

$$\langle p, x_0\rangle + qf(x_0) \leqslant \inf_{(x,t)\in\mathrm{epi}f} \{\langle p, x\rangle + qt\} \tag{5.3}$$

We can normalize the coefficients so that $\|p\|^2 + q = 1$, and we note that $q \geqslant 0$ since otherwise, the RHS would get arbitrarily small when t is large. If we choose the x RHS sufficiently close to x_0, and choose $t = f(x)$, we have

$$\begin{aligned}
&\langle p, x_0\rangle + qf(x_0) \leqslant \langle p, x\rangle + qf(x) \\
\Longrightarrow \quad &\langle p, x_0 - x\rangle \leqslant q(f(x) - f(x_0)) \leqslant Lq\|x - x_0\|
\end{aligned}$$

where the last inequality follows from the local Lipschitz property 5.1.7. Take $x = x_0 - \epsilon p$ for some small $\epsilon > 0$, we deduce that $\epsilon\|p\|^2 \leqslant \epsilon Lq\|p\|$, which indicates $\|p\| \leqslant Lq$. Hence from the normalizing condition, $q \neq 0$. Thus, for any $x \in \mathrm{dom}\ f$, we deduce that

$$f(x) \geqslant f(x_0) - \frac{1}{q}\langle p, x - x_0\rangle$$

by, again, taking $t = f(x)$ on the RHS of Equation 5.3. Thus, $-p/q \in \partial f(x_0)$ by Equation 5.2.

For closedness and convexity, note that $\partial f(x_0) = \{p \in \mathbb{R}^d : f(y) \geqslant f(x_0) + \langle p, y - x_0\rangle\}$, and it is a half-plane for each y, which is closed and convex. Therefore, $\partial f(x_0)$ is the intersection of these half-planes, and it is still closed and convex. Boundedness then follows from Lemma 5.1.12, where we use the local Lipschitz conclusion above. □

5.2 Projected Subgradient Descent

5.2.1 Set Constraint

This method assumes access to the projection mapping $\pi_{\mathcal{C}}$ for the set $\mathcal{C}$. This assumption is appropriate when the set $\mathcal{C}$ is particularly 'simple', e.g., $\mathcal{C} = \{\|\cdot\| \leqslant R\}$ is a ball, in which the projection can be computed in closed form. When $\mathcal{C}$ is more complex, e.g., a polytope, we need method beyond this.

Projected subgradient descent is the following algorithm:

$$x_{n+1} = \pi_{\mathcal{C}}\left(x_n - h\frac{p_n}{\|p_n\|}\right), \quad p_n \in \partial f(x_n) \tag{PSD}$$

Note that we use the normalized subgradient. The reason is that, if we think about the example of the absolute value function $|\cdot|$ with subdifferential $[-1,1]$ at the origin, we see that the magnitude of an arbitrary element of the subdifferential need not be informative. Instead, non-smooth optimization treat subgradients as separating directions: any minimizer must lie on one side of the hyperplane defined by the subgradient.

Theorem 5.2.1: Convergence of PSD under Convexity

Let f be convex and L-Lipschitz continuous on the closed convex set $\mathcal{C}$. Then, projected subgradient descent satisfies

$$f\left(\frac{1}{N}\sum_{n=0}^{N-1} x_n\right) - f_\star \leqslant \frac{1}{N}\sum_{n=0}^{N-1}(f(x_n) - f_\star) \leqslant \frac{L}{2Nh}\|x_0 - x_\star\|^2 + \frac{Lh}{2}$$

In particular, by setting $h = R/\sqrt{N}$, where R is an upper bound on $\|x_0 - x_\star\|$, it yields the convergence rate

$$f\left(\frac{1}{N}\sum_{n=0}^{N-1} x_n\right) - f_\star \leqslant \frac{LR}{\sqrt{N}}$$

Proof. The first inequality holds by convexity, we focus on the second. The idea is similar to the proof of Theorem 2.2.6, except that instead of smoothness, we use Lipschitzness to handle error term. Expanding the square,

$$\begin{aligned}
\|x_{n+1} - x_\star\|^2 &= \left\|\pi_{\mathcal{C}}\left(x_n - h\frac{p_n}{\|p_n\|}\right) - \pi_{\mathcal{C}}(x_\star)\right\|^2 && \text{(Defnition of PSD, and } x^\star \in \mathcal{C}\text{)} \\
&\leqslant \left\|x_n - h\frac{p_n}{\|p_n\|} - x_\star\right\|^2 && \text{(Lemma 5.1.11)} \\
&= \|x_n - x_\star\|^2 - \frac{2h}{\|p_n\|}\langle p_n, x_n - x_\star\rangle + h^2 && \text{(Expanding the square)} \\
&\leqslant \|x_n - x_\star\|^2 - \frac{2h}{\|p_n\|}(f(x_n) - f_\star) + h^2 && \text{(Equation 5.2)}
\end{aligned}$$

By Lemma 5.1.12, we have $\|p_n\| \leqslant L$ for all n, we sum the inequalities:

$$\begin{aligned}
&\frac{2h}{\|p_n\|}(f(x_n) - f_\star) \leqslant \|x_n - x_\star\|^2 - \|x_{n+1} - x_\star\|^2 + h^2 \\
\Longrightarrow\quad & f(x_n) - f_\star \leqslant \frac{\|p_n\|}{2h}\left(\|x_n - x_\star\|^2 - \|x_{n+1} - x_\star\|^2\right) + \frac{\|p_n\|h}{2} \leqslant \frac{L}{2h}\left(\|x_n - x_\star\|^2 - \|x_{n+1} - x_\star\|^2\right) + \frac{Lh}{2} \\
\Longrightarrow\quad & \frac{1}{N}\sum_{n=0}^{N-1}(f(x_n) - f_\star) \leqslant \frac{L}{2Nh}\left(\|x_0 - x_\star\|^2 - \|x_N - x_\star\|^2\right) + \frac{Lh}{2} \leqslant \frac{L}{2Nh}\|x_0 - x_\star\|^2 + \frac{Lh}{2}
\end{aligned}$$

Take $h = R/\sqrt{N}$, we have

$$f\left(\frac{1}{N}\sum_{n=0}^{N-1} x_n\right) - f_\star \leqslant \frac{L\sqrt{N}}{2NR}R^2 + \frac{LR}{2\sqrt{N}} = \frac{LR}{\sqrt{N}}$$

which completes the proof. □

Remark:

- The averaged iterate $\tilde{x}_N$ satisfies $f(\tilde{x}_N) - f_\star \leqslant \epsilon$ provided $N \geqslant L^2R^2/\epsilon^2$, which is substantially worse than the one for smooth case.
- Here descent lemma is not available either, so the guarantee only holds for the averaged iterate.

The analysis can also be performed under strong convexity.

Theorem 5.2.2: Convergence of PSD under Strong Convexity

Assume f is α-convex and L-Lipschitz continuous over the closed convex set $\mathcal{C}$. Then, projected subgradient descent satisfies

$$f(\tilde{x}_N) - f_\star \leqslant \frac{\alpha}{2\left((1 - \alpha h/L)^{-N} - 1\right)}\|x_0 - x_\star\|^2 + \frac{Lh}{2}$$

where $\tilde{x}_N$ is a suitable averaged iterate. In particular, by setting $h = \epsilon/L$, we can achieve $f(\tilde{x}_N) - f_\star \leqslant \epsilon$ in $O\left(\frac{L^2}{\alpha\epsilon}\log\left(\frac{\alpha R^2}{\epsilon}\right)\right)$ iterations.

Proof. **We first show that the Strong convexity condition 1.4 is still satisfied for non-smooth function when ∇f is substituted by the subgradient** p. Let $g(x) = f(x) - \frac{\alpha}{2}\|x\|^2$. Then, g is 0-convex. Therefore, for $q \in \partial g(x)$, by Equation 5.2,

$$g(y) \geqslant g(x) + \langle q, y - x\rangle$$

We prove that $\partial g(x) = \partial f(x) - \alpha x$. To do this, we write the equation above in f,

$$\begin{aligned}
& f(y) - \frac{\alpha}{2}\|y\|^2 \geqslant f(x) - \frac{\alpha}{2}\|x\|^2 - \alpha\langle x, y - x\rangle + \langle q + \alpha x, y - x\rangle \\
\Longrightarrow\quad & f(y) \geqslant f(x) + \frac{\alpha}{2}\|y - x\|^2 + \langle q + \alpha x, y - x\rangle \geqslant f(x) + \langle q + \alpha x, y - x\rangle \qquad (5.4)
\end{aligned}$$

This shows that for $p \in \partial f(x)$, $q = p - \alpha x \in \partial g(x)$, and the inequality also recovers Equation 1.4 with gradient substituted by subgradient. Therefore, by expanding the squares,

$$\|x_{n+1} - x_\star\|^2 = \left\|\pi_{\mathcal{C}}\left(x_n - h\frac{p_n}{\|p_n\|}\right) - \pi_{\mathcal{C}}(x_\star)\right\|^2 \qquad \text{(Defnition of PSD, and } x^\star \in \mathcal{C}\text{)}$$

$$\begin{aligned}
&\leqslant \left\| x_n - h\frac{p_n}{\|p_n\|} - x_\star \right\|^2 && \text{(Lemma 5.1.11)} \\
&= \|x_n - x_\star\|^2 - \frac{2h}{\|p_n\|}\langle p_n, x_n - x_\star \rangle + h^2 && \text{(Expanding the square)} \\
&\leqslant \|x_n - x_\star\|^2 - \frac{2h}{\|p_n\|}(f(x_n) - f_\star) - \frac{\alpha h}{\|p_n\|}\|x_n - x_\star\|^2 + h^2 && \text{(Equation 5.4)} \\
&= \left(1 - \frac{\alpha h}{\|p_n\|}\right)\|x_n - x_\star\|^2 - \frac{2h}{\|p_n\|}(f(x_n) - f_\star) + h^2 \\
&\leqslant \left(1 - \frac{\alpha h}{L}\right)\|x_n - x_\star\|^2 - \frac{2h}{L}(f(x_n) - f_\star) + h^2 && (\|p_n\| \leqslant L)
\end{aligned}$$

Use the discrete Grönwall with $A = 1 - \alpha h/L$ and $B_n = -\frac{2h}{L}(f(x_n) - f_\star) + h^2$, we have

$$0 \leqslant \|x_N - x_\star\|^2 \leqslant \left(1 - \frac{\alpha h}{L}\right)^N \|x_0 - x_\star\|^2 + \sum_{n=1}^{N}\left(1 - \frac{\alpha h}{L}\right)^{N-n}\left(-\frac{2h}{L}(f(x_{n-1}) - f_\star)\right) + h^2 \sum_{n=1}^{N}\left(1 - \frac{\alpha h}{L}\right)^{N-n}$$

Rearrange, we have

$$\sum_{n=1}^{N}\left(1 - \frac{\alpha h}{L}\right)^{-n}(f(x_{n-1}) - f_\star) \leqslant \frac{L}{2h}\|x_0 - x_\star\|^2 + \frac{Lh}{2}\sum_{n=1}^{N}\left(1 - \frac{\alpha h}{L}\right)^{-n}$$

Note that

$$\sum_{n=1}^{N}\left(1 - \frac{\alpha h}{L}\right)^{-n} = \frac{(1 - \alpha h/L)^{-1}\left(1 - (1 - \alpha h/L)^{-N}\right)}{1 - (1 - \alpha h/L)^{-1}} = \frac{(1 - \alpha h/L)^{-N} - 1}{\alpha h/L}$$

Multiply both sides by $1/\sum_{n=1}^{N}(1 - \alpha h/L)^{-n}$, we have

$$\frac{\sum_{n=1}^{N}\left(1 - \frac{\alpha h}{L}\right)^{-n}(f(x_{n-1}) - f_\star)}{\sum_{n=1}^{N}\left(1 - \frac{\alpha h}{L}\right)^{-n}} \leqslant \frac{\alpha}{2((1 - \alpha h/L)^{-N} - 1)}\|x_0 - x_\star\|^2 + \frac{Lh}{2}$$

This shows that $f(\tilde{x}_N) - f_\star$ is bounded by RHS with

$$\tilde{x}_N = \frac{\sum_{n=0}^{N-1}(1 - \alpha h/L)^{-n-1}x_n}{\sum_{n=0}^{N-1}(1 - \alpha h/L)^{-n-1}}$$

which is an averaged sum. Thus, $f(\tilde{x}_N) - f_\star \leqslant LHS \leqslant RHS$ by convexity. If we set $h = \epsilon/L$, to achieve $f(\tilde{x}_N) - f_\star \leqslant \epsilon$, we have

$$\begin{aligned}
&\frac{\alpha}{2\left((1 - \alpha\epsilon/L^2)^{-N} - 1\right)}R^2 + \frac{\epsilon}{2} \leqslant \epsilon \\
\implies\quad &(1 - \alpha\epsilon/L^2)^{-N} - 1 \geqslant \frac{\alpha R^2}{\epsilon} \\
\implies\quad &-N\log\left(1 - \frac{\alpha\epsilon}{L^2}\right) \gtrsim \log\left(\frac{\alpha R^2}{\epsilon}\right) \\
\implies\quad &N \gtrsim \frac{\log\left(\frac{\alpha R^2}{\epsilon}\right)}{-\log\left(1 - \frac{\alpha\epsilon}{L^2}\right)}
\end{aligned}$$

$$\implies N \gtrsim \frac{\log\left(\frac{\alpha R^2}{\epsilon}\right)}{\frac{\alpha\epsilon}{L^2}} = \frac{L^2}{\alpha\epsilon}\log\left(\frac{\alpha R^2}{\epsilon}\right) \qquad \text{(1st-order Taylor Approximation)}$$

Therefore, the algorithm achieves the complexity $O\left(\frac{L^2}{\alpha\epsilon}\log\left(\frac{\alpha R^2}{\epsilon}\right)\right)$. □

Remark: Under these assumptions, we also have $\|x_0 - x_\star\| \leqslant 2L/\alpha$. This is because:

- Under strong convexity, we have
$$f(x_0) \geqslant f_\star + \frac{\alpha}{2}\|x_0 - x_\star\|^2$$
- Under L-Lipschitz continuity, we have
$$f(x_0) - f_\star \leqslant L\|x_0 - x_\star\|$$
  Combining these two, we have
$$\frac{\alpha}{2}\|x_0 - x_\star\|^2 \leqslant L\|x_0 - x_\star\| \quad \implies \quad \|x_0 - x_\star\| \leqslant \frac{2L}{\alpha}$$

If we only assume that f is L-Lipschitz continuous over $B(x_\star, R)$, rather than on all of $\mathcal{C}$, it is still possible to show that $\min_{n=0,\cdots,N-1} f(x_n) - f_\star \leqslant LR/\sqrt{N}$, although the proof is more involved. See Nestorov[11] chapter 3.2.3.

5.2.2 Functional Constraints

Now we tackle a more general setting in which we separate out the constraints into a simple set $\mathcal{C}$ with access to projection operator, and additional functional constraints $\{f_i \leqslant 0 \,\forall\, i \in [m]\}$. Thus, we consider:

$$\min\{f(x) \mid x \in \mathcal{C}, f_i(x) \leqslant 0 \,\forall\, i \in [m]\}$$

Assume $f_1, \cdots, f_m$ are all regular convex functions, and write $f_{\max} = \max_{i\in[m]} f_i$.

The algorithm we consider is the **projected subgradient method with functional constraints**.

For $n = 0, 1, \cdots, N-1$:

- If $f_{\max}(x_n) \leqslant \epsilon$, set
$$x_{n+1} = \pi_{\mathcal{C}}\left(x_n - \frac{\epsilon}{\|p_n\|^2}p_n\right), \quad p_n \in \partial f(x_n)$$
- Otherwise, set
$$x_{n+1} = \pi_{\mathcal{C}}\left(x_n - \frac{f_{\max}(x_n)}{\|p_n\|^2}p_n\right), \quad p_n \in \partial f_{\max}(x_n)$$

The algorithm requires computing elements of the subdifferential for function $f_{\max}$. We therefore first identify this subdifferential.

Lemma 5.2.3: Subdifferential of a maximum

Let $f_1, \cdots, f_m$ be regular convex functions. Then, for all $x \in \mathbb{R}^d$,

$$\partial f_{\max}(x) = \operatorname{conv}\{\partial f_i(x) \mid i \in [m], f_i(x) = \max_{j \in [m]} f_j(x)\}$$

Proof. ($\supseteq$) Let $I_\star(x) = \{i \in [m] : f_i(x) = f_{\max}(x)\}$. If λ is a probability vector and $p_i \in \partial f_i(x)$ for all $i \in I_\star(x)$, then,

$$f_{\max}(y) \geqslant \sum_{i \in I_\star(x)} \lambda_i f_i(y) \geqslant \sum_{i \in I_\star(x)} \lambda_i (f_i(x) + \langle p_i, y - x\rangle) = f_{\max}(x) + \left\langle \sum_{i \in I_\star(x)} p_i, y - x \right\rangle$$

Hence, $\sum_{i \in I_\star(x)} \lambda_i p_i \in \partial f_{\max}(x)$.
($\subseteq$) Since the purpose of this lemma from the perspective of these notes is simply to compute an element of $\partial f_{\max}(x)$, we omit the proof of this direction. It can be proven, e.g., via Lagrangian duality or via more subdifferential theory. □

The convergence rate is established below. It is no more than the case without functional constraints.

Theorem 5.2.4: Convergence of PSD under functional constraints

Let $f, f_1, \cdots, f_m$ be convex and L-Lipschitz on the closed convex set $\mathcal{C}$. Then, project subgradient descent under functional constraints satisfies

$$\min\{f(x_n) \mid n = 0, 1, \cdots, N-1, f_{\max}(x_n) \leqslant \epsilon\} - f_\star \leqslant \epsilon$$

provided that

$$N \geqslant \frac{L^2 \|x_0 - x_\star\|^2}{\epsilon^2}$$

Proof. There are two cases for the algorithm. If the iteration n belongs to the first case, then as we saw in the proof of Theorem 5.2.1,

$$\|x_{n+1} - x_\star\|^2 \leqslant \|x_n - x_\star\|^2 - \frac{2\epsilon}{\|p_n\|^2}(f(x_n) - f_\star) + \frac{\epsilon^2}{\|p_n\|^2}$$

If $f(x_n) - f_\star \leqslant \epsilon$, then since $f_{\max}(x_n) \leqslant \epsilon$ by the definition of the first case, we have met the success condition. Otherwise, $f(x_n) - f_\star > \epsilon$, and the inequality above implies

$$\|x_{n+1} - x_\star\|^2 < \|x_n - x_\star\|^2 - \frac{\epsilon^2}{\|p_n\|^2} \leqslant \|x_n - x_\star\|^2 - \frac{\epsilon^2}{L^2}$$

If it is the second case, since $x_\star$ satisfies the functional constraints and x_n does not, the subgradient $p_n \in \partial f_{\max}(x_n)$ still acts as a separating hyperplane, and

$$\begin{aligned}\|x_{n+1} - x_\star\|^2 &= \left\| \pi_{\mathcal{C}}\left(x_n - \frac{f_{\max}(x_n)}{\|p_n\|^2} p_n\right) - \pi_{\mathcal{C}}(x_\star) \right\|^2 \leqslant \left\| x_n - \frac{f_{\max}(x_n)}{\|p_n\|^2} p_n - x_\star \right\|^2 && \text{(Non-expansivity)} \\ &= \|x_n - x_\star\|^2 - \frac{2 f_{\max}(x_n)}{\|p_n\|^2} \langle p_n, x_n - x_\star\rangle + \frac{f_{\max}(x_n)^2}{\|p_n\|^2}\end{aligned}$$

Since $\langle p_n, x_n - x_\star\rangle \geqslant f_{\max}(x_n) - f_{\max}(x_\star) \geqslant f_{\max}(x_n)$ where the inequality follows since $x_\star$ satisfies the functional constraints. We can further write the inequality as

$$\begin{aligned} \|x_{n+1} - x_\star\|^2 &\leqslant \|x_n - x_\star\|^2 - \frac{2 f_{\max}(x_n)}{\|p_n\|^2} f_{\max}(x_n) + \frac{f_{\max}(x_n)^2}{\|p_n\|^2} \\ &< \|x_n - x_\star\|^2 - \frac{\epsilon^2}{L^2} && (f_{\max}(x_n) > \epsilon) \end{aligned}$$

which is the same inequality as derived in case 1. Summing these inequalities, we have

$$\|x_N - x_\star\|^2 < \|x_0 - x_\star\|^2 - \frac{N\epsilon^2}{L^2}$$

For $N \geqslant L^2\|x_0 - x_\star\|^2/\epsilon^2$, we will have $\|x_N - x_\star\|^2 < 0$, which is impossible. Therefore, there must some step that hit case 1 before n, and the success condition is met. □

5.3 Cutting Plane Methods

Suppose we wish to minimize f over a bounded, closed, convex set $\mathcal{C}$. Let $\mathcal{C}_\star$ denote the set of minimizers. The idea is to construct a sequence of convex sets $(\mathcal{C}_n)_{n\in\mathbb{N}} = \mathcal{C}_0, \mathcal{C}_1, \mathcal{C}_2, \cdots$, which shrink towards $\mathcal{C}_\star$. The set $\mathcal{C}_n$ represents possible candidates for the solution to the problem at iteration n.

For $x_n \in \mathcal{C}_n$ and $p_n \in \partial f(x_n)$, the subgradient inequality 5.2 reads

$$0 \geqslant f(x_\star) - f(x_n) \geqslant \langle p_n, x_\star - x_n\rangle$$

Thus,

$$\mathcal{C}_\star \subseteq \mathcal{C}_n \cap \{x \in \mathbb{R}^d : \langle p_n, x\rangle \leqslant \langle p_n, x_n\rangle\}$$

We can take $\mathcal{C}_{n+1}$ to be any superset of the RHS above. However, we want to shrink as fast as possible, which uses the following lemma from convex geometry.

Lemma 5.3.1: Grünbaum Inequality

Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a convex body (i.e., a compact convext set with non-empty interior) and let $x_{\mathcal{C}}$ denote the centroid of $\mathcal{C}$, i,e, $x_{\mathcal{C}} = (\operatorname{vol}\mathcal{C})^{-1}\int_{\mathcal{C}} x \, \mathrm{d}x$. Then, for any half-space $\mathcal{H}$ containing $x_{\mathcal{C}}$,

$$\frac{\operatorname{vol}(C \cap \mathcal{H})}{\operatorname{vol}(\mathcal{C})} \geqslant \left(\frac{d}{d+1}\right)^d \geqslant \frac{1}{e}$$

Consequently, if we choose x_n to be the centroid of $\mathcal{C}_n$ and set

$$\mathcal{C}_{n+1} = \mathcal{C}_n \cap \{x : \langle p_n, x\rangle \leqslant \langle p_n, x_n\rangle\}, \quad x_n = x_{\mathcal{C}_n} \tag{CoGM}$$

then Grünbaum Inequality shows that,

$$\frac{\operatorname{vol}(\mathcal{C}_n \backslash \mathcal{C}_{n+1})}{\operatorname{vol}(\mathcal{C}_n)} \geqslant \frac{1}{e} \quad \Longrightarrow \quad \frac{\operatorname{vol}(\mathcal{C}_{n+1})}{\operatorname{vol}(\mathcal{C}_n)} \leqslant 1 - \frac{1}{e}$$

Thus, we cut away a constant fraction of the volume at each iteration. This is known as the **center of gravity method**.

However, it is not a practical method. The feasible set $\mathcal{C}_n$ at iteration n can be quite complicated, making it prohibitively expensive to compute its centroid. Centroids can be computed via Markov chain Monte Carlo (MCMC) methods for numerical integration, with guarantees available due to recent advances in log-concave sampling, but it is generally understood that this is a more difficult computational problem than the original convex optimization problem we set out to solve. Nevertheless, CoGM achieves the optimal complexity bound in the oracle model, so let us analyze its efficiency.

Theorem 5.3.2: Center of Gravity Bound

Let $D = \operatorname{diam} \mathcal{C}$ and let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and L-Lipschitz on $\mathcal{C}$. Then, CoGM satisfies

$$f(x_{N-1}) - f_\star \leqslant DL \left(1 - \frac{1}{e}\right)^{N/d}$$

Proof. By the argument above, at iteration N, $\operatorname{vol}(\mathcal{C}_N)/\operatorname{vol}(\mathcal{C}) \leqslant \lambda^N$, where we can take $\lambda = 1 - 1/e$. Now we consider the set $\hat{\mathcal{C}} = (1-t)x_\star + t\mathcal{C}$, where we choose t so that $\operatorname{vol}(\hat{\mathcal{C}}) > \operatorname{vol}(\mathcal{C}_N)$. Since $\operatorname{vol}(\hat{\mathcal{C}}) = t^d \operatorname{vol}(\mathcal{C})$, we can take any $t > \lambda^{N/d}$, so that $\operatorname{vol}(\mathcal{C}_N) \leqslant \lambda^N \operatorname{vol}(\mathcal{C}) < t^d \operatorname{vol}(\mathcal{C}) = \operatorname{vol}(\hat{\mathcal{C}})$. Then, with this choice, there must be a $\hat{x} \in \hat{\mathcal{C}} \backslash \mathcal{C}_N$. By the definition of $\mathcal{C}_N$,

$$\mathcal{C}_N = \{x \in \mathcal{C}_{N-1} : \langle p_{N-1}, x \rangle \leqslant \langle p_{N-1}, x_{N-1} \rangle\}$$

Thus, any point $x \notin \mathcal{C}_N$ must satisfy

$$\langle p_{N-1}, x \rangle > \langle p_{N-1}, x_{N-1} \rangle$$

Since $\hat{x} \notin \mathcal{C}_N$, and using the subgradient inequality 5.2, we have

$$f(\hat{x}) \geqslant f(x_{N-1}) + \langle p_{N-1}, \hat{x} - x_{N-1} \rangle > f(x_{N-1})$$

Therefore, we have $f(x_{N-1}) - f_\star \leqslant f(\hat{x}) - f_\star$. To study $f(\hat{x})$, note that $\hat{\mathcal{C}} = (1-t)x_\star + t\mathcal{C}$, which means that $\hat{x}$ can be written as $\hat{x} = (1-t)x_\star + ty$, where $y \in \mathcal{C}$. Therefore, by convexity,

$$\begin{aligned} f(x_{N-1}) - f_\star \leqslant f(\hat{x}) - f_\star = f((1-t)x_\star + ty) - f_\star \leqslant (1-t)f(x_\star) + tf(y) - f(x_\star) \\ = t(f(y) - f_\star) \leqslant t\left(\sup_{\mathcal{C}} f - f_\star\right) \leqslant tDL \end{aligned}$$

where the last inequality holds because of L-Lipschitzness. The result then follows by letting $t \searrow \lambda^{N/d}$. □

Thus, we can achieve $f(x_{N-1}) - f_\star \leqslant \epsilon$ in $O(d \log(DL/\epsilon))$ iterations. Compared to the projected subgradient method (Theorem 5.2.1), this result incurs only a logarithmic dependence on the ratio DL/ϵ, i.e., we can output a high-accuracy solution even for poorly conditioned convex sets. On the other hand, it incurs dependence on the dimension.

5.3.1 Ellipsoid Method

To make CoGM more practical, a famous example is the **ellipsoid method**. In this scheme, we take each set $\mathcal{C}_n$ an ellipsoid

$$\mathcal{C}_n = \{x \in \mathbb{R}^d : \langle x - x_n, \Sigma_n^{-1}(x - x_n)\rangle \leqslant 1\}$$

At the next iteration, we must find a new ellipsoid such that

$$\mathcal{C}_{n+1} \supseteq \mathcal{C}_n \cap \{x \in \mathbb{R}^d : \langle p_n, x\rangle \leqslant \langle p_n, x_n\rangle\}$$

We use the following lemma:

Lemma 5.3.3: Ellipsoid Lemma

Let $\mathcal{C}_n$ be the ellipsoid $\mathcal{C}_n = \{x \in \mathbb{R}^d : \langle x - x_n, \Sigma_n^{-1}(x - x_n)\rangle \leqslant 1\}$ and let $p_n \in \mathbb{R}^d$ be a non-zero vector. Define $\mathcal{C}_{n+1} = \{x \in \mathbb{R}^d : \langle x - x_{n+1}, \Sigma_{n+1}^{-1}(x - x_{n+1})\rangle \leqslant 1\}$, where

$$x_{n+1} = x_n - \frac{1}{d+1}\frac{\Sigma_n p_n}{\sqrt{\langle p_n, \Sigma_n p_n\rangle}}$$

$$\Sigma_{n+1} = \frac{d^2}{d^2-1}\left(\Sigma_n - \frac{2}{d+1}\frac{\Sigma_n p_n p_n^T \Sigma_n}{\langle p_n, \Sigma_n p_n\rangle}\right)$$

Then, for $d > 1$, $\mathcal{C}_{n+1}$ satisfies that $\mathcal{C}_{n+1} \supseteq \mathcal{C}_n \cap \{x \in \mathbb{R}^d : \langle p_n, x\rangle \leqslant \langle p_n, x_n\rangle\}$, and

$$\frac{\mathrm{vol}(\mathcal{C}_{n+1})}{\mathrm{vol}(\mathcal{C}_n)} = \sqrt{\frac{d-1}{d+1}\left(\frac{d^2}{d^2-1}\right)^d} = 1 - \Omega\left(\frac{1}{d}\right)$$

Proof. We need to first show that $\mathcal{C}_{n+1} \supseteq \mathcal{C}_n \cap \{x \in \mathbb{R}^d : \langle p_n, x\rangle \leqslant \langle p_n, x_n\rangle\}$. To do this, we first compute Σ_{n+1}^{-1}. By *Sherman–Morrison formula*,

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u}$$

We let

$$A = \Sigma_n, \quad u = -\frac{2}{d+1}\frac{1}{\langle p_n, \Sigma_n p_n\rangle}\Sigma_n p_n, \quad v = \Sigma_n p_n$$

in our case, so that

$$\begin{aligned}
\Sigma_{n+1}^{-1} &= \frac{d^2-1}{d^2}\left(\Sigma_n^{-1} - \frac{\Sigma_n^{-1}\frac{-2}{d+1}\frac{\Sigma_n p_n}{\langle p_n, \Sigma_n p_n\rangle}p_n^T\Sigma_n\Sigma_n^{-1}}{1 - p_n^T\Sigma_n\Sigma_n^{-1}\frac{2}{d+1}\frac{1}{\langle p_n, \Sigma_n p_n\rangle}\Sigma_n p_n}\right) \\
&= \frac{d^2-1}{d^2}\left(\Sigma_n^{-1} + \frac{\frac{2}{d+1}\frac{p_n p_n^T}{\langle p_n, \Sigma_n p_n\rangle}}{1 - \frac{2}{d+1}}\right) \\
&= \frac{d^2-1}{d^2}\Sigma_n^{-1} + \frac{(d+1)(d-1)}{d^2}\frac{2p_n p_n^T}{(d-1)\langle p_n, \Sigma_n p_n\rangle} \\
&= \frac{d^2-1}{d^2}\Sigma_n^{-1} + \frac{2(d+1)}{d^2}\frac{p_n p_n^T}{\langle p_n, \Sigma_n p_n\rangle}
\end{aligned}$$

Now, suppose $x \in \mathcal{C}_n \cap \{x \in \mathbb{R}^d : \langle p_n, x\rangle \leqslant \langle p_n, x_n\rangle\}$, our goal is to show that $x \in \mathcal{C}_{n+1}$. By brute-force calculation,

$$\begin{aligned}
&\langle x - x_{n+1}, \Sigma_{n+1}^{-1}(x - x_{n+1})\rangle \\
=&\left\langle x - x_n + \frac{1}{d+1}\frac{\Sigma_n p_n}{\sqrt{\langle p_n, \Sigma_n p_n\rangle}}, \left(\frac{d^2-1}{d^2}\Sigma_n^{-1} + \frac{2(d+1)}{d^2}\frac{p_n p_n^T}{\langle p_n, \Sigma_n p_n\rangle}\right)\left(x - x_n + \frac{1}{d+1}\frac{\Sigma_n p_n}{\sqrt{\langle p_n, \Sigma_n p_n\rangle}}\right)\right\rangle \\
=&\left\langle x - x_n + \frac{1}{d+1}\frac{\Sigma_n p_n}{\sqrt{\langle p_n, \Sigma_n p_n\rangle}}, \frac{d^2-1}{d^2}\Sigma_n^{-1}(x-x_n) + \frac{d-1}{d^2\sqrt{\langle p_n, \Sigma_n p_n\rangle}}p_n + \frac{2(d+1)\langle p_n, x - x_n\rangle}{d^2\langle p_n, \Sigma_n p_n\rangle}p_n + \frac{2}{d^2\sqrt{\langle p_n, \Sigma_n p_n\rangle}}p_n\right\rangle \\
=&\left\langle x - x_n + \frac{1}{d+1}\frac{\Sigma_n p_n}{\sqrt{\langle p_n, \Sigma_n p_n\rangle}}, \frac{d^2-1}{d^2}\Sigma_n^{-1}(x-x_n) + \frac{d+1}{d^2\sqrt{\langle p_n, \Sigma_n p_n\rangle}}p_n + \frac{2(d+1)\langle p_n, x - x_n\rangle}{d^2\langle p_n, \Sigma_n p_n\rangle}p_n\right\rangle \\
=&\frac{d^2-1}{d^2}\langle x - x_n, \Sigma_n^{-1}(x-x_n)\rangle + \frac{d+1}{d^2\sqrt{\langle p_n, \Sigma_n p_n\rangle}}\langle x - x_n, p_n\rangle + \frac{2(d+1)\langle p_n, x - x_n\rangle}{d^2\langle p_n, \Sigma_n p_n\rangle}\langle x - x_n, p_n\rangle \\
&+ \frac{d-1}{d^2\sqrt{\langle p_n, \Sigma_n p_n\rangle}}\underbrace{\langle \Sigma_n p_n, \Sigma_n^{-1}(x-x_n)\rangle}_{=\langle p_n, x - x_n\rangle} + \frac{d+1}{d^2(d+1)} + \frac{2}{d^2\sqrt{\langle p_n, \Sigma_n p_n\rangle}}\langle p_n, x - x_n\rangle \\
=&\frac{d^2-1}{d^2}\langle x - x_n, \Sigma_n^{-1}(x-x_n)\rangle + \frac{2(d+1)}{d^2\sqrt{\langle p_n, \Sigma_n p_n\rangle}}\langle x - x_n, p_n\rangle + \frac{2(d+1)\langle p_n, x - x_n\rangle}{d^2\langle p_n, \Sigma_n p_n\rangle}\langle x - x_n, p_n\rangle + \frac{1}{d^2} \\
\leqslant&\frac{d^2-1}{d^2} + \frac{1}{d^2} + \frac{2(d+1)}{d^2\sqrt{\langle p_n, \Sigma_n p_n\rangle}}\langle x - x_n, p_n\rangle + \frac{2(d+1)\langle p_n, x - x_n\rangle}{d^2\langle p_n, \Sigma_n p_n\rangle}\langle x - x_n, p_n\rangle \qquad (\langle x - x_n, \Sigma_n^{-1}(x-x_n)\rangle \leqslant 1) \\
=&1 + \frac{2(d+1)}{d^2\sqrt{\langle p_n, \Sigma_n p_n\rangle}}\langle x - x_n, p_n\rangle + \frac{2(d+1)\langle p_n, x - x_n\rangle}{d^2\langle p_n, \Sigma_n p_n\rangle}\langle x - x_n, p_n\rangle
\end{aligned}$$

Therefore, the only thing need to do is to show that the sum of the second & third term is less than or equal to 0. For simplicity of notation, we denote

$$k = \frac{\langle x - x_n, p_n\rangle}{\sqrt{\langle p_n, \Sigma_n p_n\rangle}}$$

The equation then becomes

$$\langle x - x_{n+1}, \Sigma_{n+1}^{-1}(x - x_{n+1})\rangle \leqslant 1 + \frac{2(d+1)}{d^2}(k + k^2)$$

Note that $k \leqslant 0$ since $\langle p_n, x - x_n\rangle \leqslant 0$ by assumption. Therefore, to make $k + k^2 \leqslant 0$, we only need $k \geqslant -1$. This is achieved by Cauchy-Schwartz inequality:

$$|\langle x - x_n, p_n\rangle| = |\langle \Sigma_n^{1/2}p_n, \Sigma_n^{-1/2}(x - x_n)\rangle| \leqslant \sqrt{\langle x - x_n, \Sigma_n^{-1}(x - x_n)\rangle}\sqrt{\langle p_n, \Sigma_n p_n\rangle} \leqslant \sqrt{\langle p_n, \Sigma_n p_n\rangle}$$

which completes the proof of $x \in \mathcal{C}_{n+1}$. This shows that $\mathcal{C}_{n+1} \supseteq \mathcal{C}_n \cap \{x \in \mathbb{R}^d : \langle p_n, x\rangle \leqslant \langle p_n, x_n\rangle\}$.

Now we show the ratio of volumes. We have

$$\frac{\text{vol}(\mathcal{C}_{n+1})}{\text{vol}(\mathcal{C}_n)} = \sqrt{\frac{\det(\Sigma_{n+1})}{\det(\Sigma_n)}}$$

We can write

$$\frac{2}{d+1}\frac{\Sigma_n p_n p_n^T \Sigma_n}{\langle p_n, \Sigma_n p_n\rangle} = \sqrt{\frac{2}{d+1}}\frac{\Sigma_n p_n}{\sqrt{\langle p_n, \Sigma_n p_n\rangle}}\left(\sqrt{\frac{2}{d+1}}\frac{\Sigma_n p_n}{\sqrt{\langle p_n, \Sigma_n p_n\rangle}}\right)^T = uu^T$$

where $u = \sqrt{2/(d+1)}\Sigma_n p_n/\sqrt{\langle p_n, \Sigma_n p_n\rangle}$. Therefore, using the *Matrix determinant lemma*, which says $\det(A + uv^T) =$

$(1 + v^T A^{-1} u)\det(A)$, we have

$$\det\left(\Sigma_n - \frac{2}{d+1}\frac{\Sigma_n p_n p_n^T \Sigma_n}{\langle p_n, \Sigma_n p_n\rangle}\right) = \det\left(\Sigma_n + (-u)u^T\right) = (1 - u^T \Sigma_n^{-1} u)\det(\Sigma_n)$$

Now we compute

$$u^T \Sigma_n^{-1} u = \frac{2}{d+1}\frac{p_n^T \Sigma_n \Sigma_n^{-1} \Sigma_n p_n}{\langle p_n, \Sigma_n p_n\rangle} = \frac{2}{d+1}$$

Therefore, we have

$$\begin{aligned}\det(\Sigma_{n+1}) &= \det\left(\frac{d^2}{d^2-1}\left(\Sigma_n - \frac{2}{d+1}\frac{\Sigma_n p_n p_n^T \Sigma_n}{\langle p_n, \Sigma_n p_n\rangle}\right)\right)\\ &= \left(\frac{d^2}{d^2-1}\right)^d \det\left(\Sigma_n - \frac{2}{d+1}\frac{\Sigma_n p_n p_n^T \Sigma_n}{\langle p_n, \Sigma_n p_n\rangle}\right)\\ &= \left(\frac{d^2}{d^2-1}\right)^d \left(1 - \frac{2}{d+1}\right)\det(\Sigma_n) = \left(\frac{d^2}{d^2-1}\right)^d \frac{d-1}{d+1}\det(\Sigma_n)\end{aligned}$$

Finally,

$$\frac{\mathrm{vol}(\mathcal{C}_{n+1})}{\mathrm{vol}(\mathcal{C}_n)} = \sqrt{\frac{\det(\Sigma_{n+1})}{\det(\Sigma_n)}} = \sqrt{\frac{d-1}{d+1}\left(\frac{d^2}{d^2-1}\right)^d}$$

The asymptotics follows from the following calculation,

$$\begin{aligned}\log\left(\frac{d-1}{d+1}\right) &= \log\left(\frac{1-1/d}{1+1/d}\right) = \log\left(1 - \frac{1}{d}\right) - \log\left(1 + \frac{1}{d}\right)\\ &= \left(-\frac{1}{d} - \frac{1}{2d^2} + O\left(\frac{1}{d^3}\right)\right) - \left(\frac{1}{d} - \frac{1}{2d^2} + O\left(\frac{1}{d^3}\right)\right) = -\frac{2}{d} + O(1/d^3)\end{aligned}$$

$$d\log\frac{d^2}{d^2-1} = d\log\left(1 + \frac{1}{d^2-1}\right) \geqslant d\log\left(1 + \frac{1}{d^2}\right) = d\left(\frac{1}{d^2} + O(1/d^4)\right) = \frac{1}{d} + O(1/d^3)$$

and therefore,

$$\log\sqrt{\frac{d-1}{d+1}\left(\frac{d^2}{d^2-1}\right)^d} = \frac{1}{2}\left(-\frac{1}{d} + O(1/d^3)\right) \Longrightarrow \sqrt{\frac{d-1}{d+1}\left(\frac{d^2}{d^2-1}\right)^d} \approx \exp\left(-\frac{1}{2d}\right) \approx 1 - \frac{1}{2d} + O(1/d^2) = 1 - \Omega\left(\frac{1}{d}\right)$$

and we complete the proof. □

The analysis of the ellipsoid method presents an additional difficulty: since the next $\mathcal{C}_{n+1}$ is only chosen to be a superset of $\mathcal{C}_n \cap \{x \in \mathbb{R}^d : \langle p_n, x\rangle \leqslant \langle p_n, x_n\rangle\}$, it is not guaranteed that $\mathcal{C} \subseteq \mathcal{C}_n$ for all n. In particular, the chosen point x_n may lie outside of $\mathcal{C}$.

Assume that we have access to a separation oracle for $\mathcal{C}$: given a point $x \notin \mathcal{C}$, the oracle outputs a non-zero vector $p \in \mathbb{R}^d$ such that $\sup_{\mathcal{C}}\langle p, \cdot\rangle \leqslant \langle p, x\rangle$. Modify the cutting plane method as follows:

if a chosen point x_n does not lie in $\mathcal{C}$, then let p_n be vector that separates x_n from $\mathcal{C}$, and intstead update $\mathcal{C}_{n+1}$ to be a superset of $\mathcal{C}_n \cap \{\langle p_n, x_n\rangle \leqslant \langle p_n, \cdot\rangle\}$. We also allow $\mathcal{C}_0 \supseteq \mathcal{C}$, so that x_0 is not necessarily feasible either.

Then, if the sets are chosen so that $\mathrm{vol}(\mathcal{C}_{n+1})/\mathrm{vol}(\mathcal{C}_n) \leqslant \lambda < 1$ for all n, then, we claim that, if $\mathrm{vol}(\mathcal{C}_N) < \mathrm{vol}(\mathcal{C})$, then there exists $n < N$ with $x_n \in \mathcal{C}$ such that

$$f(x_n) - f_\star \leqslant DL\lambda^{N/d} \left(\frac{\mathrm{vol}(\mathcal{C}_0)}{\mathrm{vol}(\mathcal{C})}\right)^{1/d}$$

Proof. □

Therefore, we can achieve $f(x_{N-1})-f_\star \leqslant \epsilon$ in $O(d^2 \log(DL/\epsilon))$ iterations. Compared to CoGM, which has $O(d \log(DL/\epsilon))$, the cost of obtaining an implementable version of the CoGM is a larger query complexity.

5.4 Lower Bounds for Non-Smooth Optimization

Theorem 5.4.1: Lower Bound for Convex, Non-smooth Minimization

For any $x_0 \in \mathbb{R}^d$, $d > N$, and $L, R > 0$, there exists a convex and L-Lipschitz function f over $B(x_\star, R)$ such that $x_0 \in B(x_\star, R)$ and for any gradient span algorithm,

$$f(x_N) - f_\star \gtrsim \frac{LR}{\sqrt{N}}$$

Proof. Assume $x_0 = 0$ without loss of generality. Define the function $f : \mathbb{R}^d \to \mathbb{R}$ by

$$f(x) = \gamma \max_{i\in[d]} x[i] + \frac{\alpha}{2}\|x\|^2$$

where $\alpha, \gamma > 0$ are to be chosen. This function is convex, and note that this function is Lipschitz with constant $\gamma + \alpha(\|x_\star\| + R)$. This is because

$$\forall\, x, y \in B(x_\star, R), \qquad \left|\max_{i\in[d]} x[i] - \max_{i\in[d]} y[i]\right| \leqslant \|x-y\|, \quad \nabla\left(\frac{\alpha}{2}\|x\|^2\right) = \alpha\|x\| \leqslant \alpha(\|x_\star\| + R)$$

Let $I_\star(x) := \{i \in [d] : x[i] = \max_{j\in[d]} x[j]\}$, then from Lemma 5.2.3,

$$\partial f(x) = \alpha x + \gamma\, \mathrm{conv}\{e_i : i \in I_\star(x)\}$$

One optimal point is $x_\star[k] = -\gamma/(\alpha d)$ for $k \in [d]$, by checking that $0 \in \partial f(x_\star)$. Thus, $\|x_\star\| = \gamma/(\alpha\sqrt{d})$ and the Lipschitz constant is at most $2\gamma + \alpha R$. We take a subgradient oracle which, given a point x, output $\alpha x + \gamma e_i \in \partial f(x)$, where $i = \min I_\star(x)$ is the first coordiante of x that achieves the maximum. Now we will show that, by induction, $x_n \in \mathcal{V}_n$ for all n, where

$$\mathcal{V}_n = \{x \in \mathbb{R}^d : x[k] = 0, \forall\, k = n+1, \cdots, d\}$$

The base case $x_0 = 0$ is true. Now suppose inductively that $x_k \in \mathcal{V}_k$ for all $k = 1, 2, \cdots, n$. Then, consider $n = k + 1$, we have new vector $\partial f(x) = \alpha x + \gamma e_i$ in the space. Since $x \in \mathcal{V}_n$, we have $\alpha x \in \mathcal{V}_n$. Moreover, if $x[k]$ has some nonnegative term for $k = 1, 2, \cdots, n$, we will have $i \leqslant n$, then $\alpha x + \gamma e_i \in \mathcal{V}_n$. Instead, if all $x[k]$ are negative, we have $i = n + 1$. Therefore, in any cases, $\partial f(x) \in \mathcal{V}_{n+1}$, which means that $x_{n+1} \in \mathcal{V}_{n+1}$ since we are considering gradient span algorithm. This completes the induction.

Since $d > N$, if follows that $f(x_N) \geqslant 0$. This is because, $x_N \in \mathcal{V}_n$, so there is some element of x that is equal to 0, which means that $\max_{i\in[d]} x_N[i] \geqslant 0$, so $f(x_N) \geqslant 0$. On the other hand,

$$f_\star = f(x_\star) = -\frac{\gamma^2}{\alpha d} + \frac{\gamma^2}{2\alpha d} = -\frac{\gamma^2}{2\alpha d}$$

Set $d = N+1$, $\gamma = L/4$, and $\alpha = \gamma/(R\sqrt{d})$ (to ensure that $\|x_0 - x_\star\| \leqslant R$), which leads to a Lipschitz constant of $L/2 + L/(4\sqrt{d}) \leqslant L$. It yields that

$$f(x_N) - f_\star \geqslant -f(x_\star) = \frac{\gamma^2}{2\alpha d} = \frac{\gamma R}{\sqrt{d}} \gtrsim \frac{LR}{\sqrt{N}}$$

which completes the proof. □

Note that this matches the guarantee of projected gradient descent in Therorem 5.2.1, so projected subgradient descent is *optimal* in the non-smooth setting. In other words, *without smoothness, there is no acceleration phenomenon.*

Theorem 5.4.2: Lower Bound for Strongly Convex, Non-Smooth Minimization

For any $x_0 \in \mathbb{R}^d$, $d > N$, and $\alpha, L > 0$, there exists $R > 0$ and an α-convex and L-Lipschitz function f over $B(x_\star, R)$ such that $x_0 \in B(x_\star, R)$ and for any gradient span algorithm,

$$f(x_N) - f_\star \gtrsim \frac{L^2}{\sqrt{N}}$$

Proof. □

Chapter 6

Structured Optimization

6.1 Conditional Gradient Descent aka Frank-Wolfe

In order to overcome the lower bounds in the black-box setting, we must take advantage of additional structure in the problem. The first method we study in this vein is the **Frank–Wolfe** or **conditional gradient descemt** method. Instead of assuming access to a projection oracle for the constraint set $\mathcal{C}$, it instead assumes access to a *linear optimization oracle* (LOO) over the set $\mathcal{C}$:

$$\text{Given } p \in \mathbb{R}^d, \text{ output } \operatorname*{argmin}_{x \in \mathcal{C}} \langle p, x \rangle \tag{LOO}$$

Here we assume that $\mathcal{C}$ is compact. The oracle equivantly maximizes the convex function $-\langle p, x \rangle$ over $\mathcal{C}$, so the argmin is attained at a vertex of $\mathcal{C}$. The following proposition is from functional analysis.

Definition 6.1.1: Vertex

A point $x \in \mathcal{C}$ is called an **extreme point** or a **vertex** of $\mathcal{C}$ if there do not exist $x_0, x_1 \in \mathcal{C}$ and $t \in (0, 1)$ such that $x = (1 - t)x_0 + tx_1$.

Proposition 6.1.2: Krein–Milman theorem

Every compact convex set is the convex hull of its extreme points.

For example, the set of vertices of the closed unit ball $\overline{B(0,1)}$ is the sphere $\partial B(0, 1)$. It follows that to implement LOO, it suffices to solve $\operatorname{argmin}_{\text{vertices of } \mathcal{C}} \langle p, \cdot \rangle$. The Frank-Wolfe method is

$$x_{n+1} = (1 - h_n)x_n + h_n \mathrm{LOO}(\nabla f(x_n)) \tag{FW}$$

Theorem 6.1.3: Convergence of Frank-Wolfe

Let f be convex and β-smooth over $\mathcal{C}$. Let $D = \operatorname{diam} \mathcal{C}$ and $h_n = 2/(n + 2)$. Then, for any $N \geqslant 1$, Frank-Wolfe satisfies

$$f(x_N) - f_\star \leqslant \frac{2\beta D^2}{N + 1}$$

Proof. Let $y_n = \mathrm{LOO}(\nabla f(x_n))$. Using β-smoothness,

$$\begin{aligned} f(x_{n+1}) - f(x_n) &\leqslant \langle \nabla f(x_n), x_{n+1} - x_n \rangle + \frac{\beta}{2}\|x_{n+1} - x_n\|^2 && (\beta\text{-smoothness}) \\ &\leqslant h_n \langle \nabla f(x_n), y_n - x_n \rangle + \frac{\beta}{2}\|h_n(y_n - x_n)\|^2 && (x_{n+1} - x_n = h_n(y_n - x_n)) \\ &\leqslant h_n \langle \nabla f(x_n), y_n - x_n \rangle + \frac{\beta D^2 h_n^2}{2} && (\text{Bounded by diameter}) \\ &\leqslant h_n \langle \nabla f(x_n), x_\star - x_n \rangle + \frac{\beta D^2 h_n^2}{2} && (\text{Definition of LOO}) \\ &\leqslant -h_n (f(x_n) - f_\star) + \frac{\beta D^2 h_n^2}{2} && (\text{Convexity}) \end{aligned}$$

where the second last inequality can change y_n to $x_\star$ since y_n optimizes the LOO oracle. Rearrange,

$$f(x_{n+1}) - f_\star \leqslant (1 - h_n)(f(x_n) - f_\star) + \frac{\beta D^2 h_n^2}{2}$$

For $h_n = 2/(n+2)$, we now prove the error bound by induction on n. Base case $n = 0$, we have $h_0 = 1$, and $f(x_1) - f_\star \leqslant \frac{\beta D^2}{2} < \beta D^2$. Suppose it holds at iteration n, then

$$f(x_{n+1}) - f_\star \leqslant \left(1 - \frac{2}{n+2}\right)(f(x_n) - f_\star) + \frac{\beta D^2}{2}\left(\frac{2}{n+2}\right)^2 \leqslant \frac{n}{n+2}\frac{2\beta D^2}{n+1} + \frac{2\beta D^2}{(n+2)^2} \leqslant \frac{2\beta D^2}{n+2}$$

which completes the proof. □

Besides fast convergence, Frank-Wolfe also have appealing property of **affine invariance**.

Proposition 6.1.4: Affine Invariance

Let $A \in \mathbb{R}^{d \times d}$ be an invertible matrix. Then, the iterates $\{\hat{x}_n\}_{n \in \mathbb{N}}$ of Frank-Wolfe applied to the problem of minimizing $\hat{x} \mapsto f(A\hat{x})$ over the set $A^{-1}\mathcal{C}$ are related to the iterates $\{x_n\}_{n \in \mathbb{N}}$ of Frank-Wolfe on the original problem of minimizing $x \mapsto f(x)$ via $x_n = A\hat{x}_n$.

Proof. Let $\hat{f}(\hat{x}) = f(A\hat{x})$. By chain rule, the gradient of $\hat{f}$ is

$$\nabla \hat{f}(\hat{x}) = A^T \nabla f(A\hat{x})$$

the LOO for $\hat{f}$ can be written as

$$\hat{y}_n = \underset{y \in A^{-1}\mathcal{C}}{\mathrm{argmin}} \langle \nabla \hat{f}(\hat{x}_n), y \rangle = \underset{y : Ay \in \mathcal{C}}{\mathrm{argmin}} \langle A^T \nabla f(A\hat{x}_n), y \rangle = \underset{y : Ay \in \mathcal{C}}{\mathrm{argmin}} \langle \nabla f(A\hat{x}_n), Ay \rangle$$

Let $y_n := A\hat{y}_n$, then $y_n = \underset{y \in \mathcal{C}}{\mathrm{argmin}} \langle \nabla f(A\hat{x}_n), y \rangle$. Combining these information, we write the iteration for $\hat{x}_n$,

$$\hat{x}_{n+1} = (1 - h_n)\hat{x}_n + h_n \hat{y}_n \quad \Longrightarrow \quad A\hat{x}_{n+1} = (1 - h_n)A\hat{x}_n + h_n y_n$$

Since the iteration for $\{x_n\}$ is $x_{n+1} = (1 - h_n)x_n + h_n y_n$, we conclude that $x_n = A\hat{x}_n$. □

Besides positing different oracle access than projected gradient methods, the Frank-Wolfe method has the appealing property of producing *sparse solutions.* Recall the theorem from convex geometry.

Theorem 6.1.5: Carathéodory's Theorem

Let $\mathcal{C} \subseteq \mathbb{R}^d$ be compact convex set and let $x \in \mathcal{C}$. Then, x can be writte nas a convex combination of $d+1$ vertices of $\mathcal{C}$.

Note that the choice of $d+1$ vertices depends on x itself. However, the size of representation grows with the dimension. Indeed, if we want an approximate representation, we can achieve *dimension-free.*

Theorem 6.1.6: Approximate Carathéodory

Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a compact convex set with diameter D. Let $0 < \epsilon < 1$, and let $x \in \mathcal{C}$. Then, there exists vertices $y_1, \cdots, y_N \in \mathcal{C}$ with

$$\left\| x - \frac{1}{N} \sum_{i=1}^{N} y_i \right\| \leqslant \epsilon D, \quad N \leqslant \frac{1}{\epsilon^2}$$

Proof. By Theorem 6.1.5, there exists vertices $\bar{y}_1, \cdots, \bar{y}_{d+1} \in \mathcal{C}$ and a probability distribution λ over $1, \cdots, d+1$ such that $x = \sum_{j=1}^{d+1} \lambda_j \bar{y}_j$. Now consider the distribution $\mu = \sum_{j=1}^{d+1} \lambda_j \delta_{\bar{y}_j}$, and sample points $Y_1, \cdots, Y_N \overset{\text{i.i.d.}}{\sim} \mu$. Note that each Y_i is a vertex of $\mathcal{C}$. Then, since the mean of μ is x, we can compute the variance

$$\mathbb{E}\left[\left\| x - \frac{1}{N} \sum_{i=1}^{N} Y_i \right\|^2 \right] = \frac{1}{N^2} \mathbb{E}\left[\left\| \sum_{i=1}^{N} (Y_i - x) \right\|^2 \right] = \frac{1}{N^2} \sum_{i=1}^{N} \mathbb{E}\|Y_i - x\|^2 \leqslant \frac{1}{N^2} N D^2 = \frac{D^2}{N}$$

where the second equality follows from independence. Therefore, there exists a realization $Y_1, \cdots, Y_N$ that satisfies

$$\left\| x - \frac{1}{N} \sum_{i=1}^{N} Y_i \right\|^2 \leqslant \frac{D^2}{N}$$

choose N to make RHS at most $\epsilon^2 D^2$, we have $N = \frac{1}{\epsilon^2}$. □

Now comes the punchline: Franke–Wolfe renders the approximate Carathéodory's theorem constructive. Indeed, suppose that the LOO always outputs a vertex. After $N-1$ iterations of Frank-Wolfe starting from a vertex, the iterate x_{N-1} is a convex combination of at most N vertices. The theorem 6.1.3 can therefore be seen as a generalization of the approximate Carathéodory principle: the iterate of Frank-Wolfe is a sparse combination of vertices which is approximately optimal.

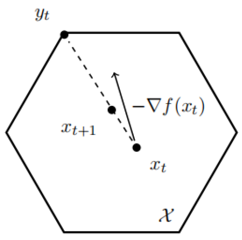


Figure 6.1: Illustration of Conditional Gradient Descent

6.2 Proximal Gradient Descent

Can we solve non-smooth problems at the same rate as smooth problems? The black-box lower bounds say no in general, but if the non-smooth part is "simple" in the sense that it admits an implementable proximal oracle, the answer becomes yes.

Definition 6.2.1: Proximal Oracle

Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$. The **proximal oracle** for f is the mapping $\operatorname{prox}_f : \mathbb{R}^d \to \mathbb{R}^d$ given by

$$\operatorname{prox}_f(y) := \operatorname*{argmin}_{x \in \mathbb{R}^d} \left\{ f(x) + \frac{1}{2}\|y - x\|^2 \right\}$$

If f is a regular convex function, then the optimization problem defining the proximal oracle is strongly convex, so it admits a unique minimizer. Note also that

$$\operatorname{prox}_{hf}(y) = \operatorname*{argmin}_{x \in \mathbb{R}^d} \left\{ hf(x) + \frac{1}{2}\|y - x\|^2 \right\} = \operatorname*{argmin}_{x \in \mathbb{R}^d} \left\{ f(x) + \frac{1}{2h}\|y - x\|^2 \right\}$$

where $h > 0$ plays the role of a step size.

Definition 6.2.2: Moreau-Yosida Envelope

Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$. The **Moreau-Yosida envelope** of f with parameter $h > 0$ is the mapping $f_h : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ given by

$$f_h(y) := \inf_{x \in \mathbb{R}^d} \left\{ f(x) + \frac{1}{2h}\|y - x\|^2 \right\}$$

6.2.1 Algorithms and Examples

The simplest algorithm is the **proximal point method**

$$x_{n+1} = \operatorname{prox}_{hf}(x_n) \tag{PPM}$$

Assume for the moment that f is smooth and the next point x_{n+1} can be obtained from the first-order optimality condition for prox_{hf}, This leads to

$$0 = \nabla f(x_{n+1}) + \frac{1}{h}(x_{n+1} - x_n) \quad \Longrightarrow \quad x_{n+1} = x_n - h\nabla f(x_{n+1})$$

This is an *implicit discretization*, where gradient descent is an *explicit discretization*. The advantage of an explicit method is easy of implementation. The advantage of an implicit method is stability.

The most powerful results using the proximal oracle, are for the problem of **composite optimization**. Here, the goal is to minimize a sum of functions:

$$\text{minimize} \quad F = f + g$$

We assume that f is smooth and that g is non-smooth.

Example 1: LASSO as composite optimization

$$f : \theta \mapsto \frac{1}{2n}\sum_{i=1}^{n}(Y_i - \langle \theta, X_i\rangle)^2, \quad g : \theta \mapsto \lambda\|\theta\|_1$$

Since the non-smooth part is particularly simple, so we can compute its proximal oracle in closed form. First note that it is coordinate-wise decomposable:

$$\begin{aligned}\operatorname{prox}_{\lambda\|\cdot\|_1}(y) &= \operatorname*{argmin}_{x\in\mathbb{R}^d}\left\{\lambda\|x\|_1 + \frac{1}{2}\|y-x\|^2\right\}\\ &= \operatorname*{argmin}_{x\in\mathbb{R}^d}\left\{\lambda\sum_{i=1}^{d}|x[i]| + \frac{1}{2}\sum_{i=1}^{d}(y[i]-x[i])^2\right\}\\ &= \sum_{i=1}^{d}\left(\operatorname*{argmin}_{x[i]\in\mathbb{R}}\left(\lambda|x[i]| + \frac{1}{2}(y[i]-x[i])^2\right)\right)e_i\end{aligned}$$

Therefore, it suffices to solve the problem in dimension one. Indeed, we can solve the closed form solution of this proximal oracle. Indeed, in one dimension, we have

$$\operatorname{prox}_{\lambda|\cdot|}(y) = \operatorname*{argmin}_{x\in\mathbb{R}}\left(\lambda|x| + \frac{1}{2}(y-x)^2\right)$$

Let $h(x) = \lambda|x| + \frac{1}{2}(y-x)^2$.

- When $x > 0$, h is differentiable, $h'(x) = \lambda + x - y$. If $y > \lambda$, then it is minimized at $x = y - \lambda$ on $(0, \infty)$. If $y \leqslant \lambda$, then $h'(x) \geqslant 0$ on $(0,\infty)$, thus it is then minimized at $x = 0$ since h is continuous.

- When $x < 0$, h is differentiable, $h'(x) = -\lambda + x - y$. If $y < -\lambda$, then it is minimized at $x = y + \lambda$ on $(-\infty, 0)$. If $y \geqslant -\lambda$, then $h'(x) \leqslant 0$ on $(-\infty, 0)$, thus it is then minimized at $x = 0$ since h is continuous.

Combining all these, we conclude

- If $-\lambda \leqslant y \leqslant \lambda$, h is minimized at $x = 0$.

- If $y > \lambda$, it is minimized at $x = y - \lambda$ on $(0,\infty)$, with minimal value $h(y-\lambda) = \lambda y - \frac{1}{2}\lambda^2$. It is minimized at $x = 0$ on $(-\infty, 0]$, with minimal value $h(0) = \frac{1}{2}y^2$. Notice that $\lambda y - \frac{1}{2}\lambda^2 < \frac{1}{2}y^2$, since after rearranging, this becomes $(y-\lambda)^2 > 0$, without equality since $\lambda \neq y$. The global minimum is obtained at $x = y - \lambda$.

- If $y < -\lambda$, it is minimized at $x = y + \lambda$ on $(-\infty, 0)$, with minimal value $h(y+\lambda) = -\lambda y - \frac{1}{2}\lambda^2$. It is minimized at $x = 0$ on $[0,\infty)$, with minimal value $h(0) = \frac{1}{2}y^2$. Similarly, the global minimal is obtained at $x = y + \lambda$.

Therefore, we conclude that

$$\operatorname{prox}_{\lambda|\cdot|}(y) = \begin{cases} y - \lambda, & y > \lambda \\ 0, & -\lambda \leqslant y \leqslant \lambda \\ y + \lambda, & y < -\lambda \end{cases} = (|y| - \lambda)_+ \operatorname{sgn}(y) = \operatorname{thresh}_\lambda(y)$$

The operator thresh_λ, known as the *soft thresholding operator*, reduces the magnitude of its input by λ, or to 0 if the original magnitude is less than λ. The proximal operator for $\lambda\|\cdot\|_1$ simply applies thresh_λ to each coordinate.

Example 2: Constrained Optimization as Composite Optimization Consider the problem of minimizing a smooth function f over a closed convex set $\mathcal{C}$. We can also treat this as composite optimization with $g = \chi_{\mathcal{C}}$. In this case, the proximal oracle for g is

$$\text{prox}_{h\chi_{\mathcal{C}}}(y) = \operatorname*{argmin}_{x\in\mathbb{R}^d}\left\{\chi_{\mathcal{C}}(x) + \frac{1}{2h}\|y-x\|^2\right\} = \operatorname*{argmin}_{x\in\mathcal{C}}\left\{\frac{1}{2h}\|y-x\|^2\right\} = \pi_{\mathcal{C}}(y)$$

So, the proximal oracle for $\chi_{\mathcal{C}}$ is the projection oracle for $\mathcal{C}$.

The above examples motivate the assumption that we have access to the proximal oracle for the non-smooth part g. Further examples of computable proximal oracles can be found on the website proximity-operator.net.

The algorithm we consider in this context is known as **proximal gradient descent**.

$$x_{n+1} = \operatorname*{argmin}_{x\in\mathbb{R}^d}\left\{f(x_n) + \langle\nabla f(x_n), x - x_n\rangle + g(x) + \frac{1}{2h}\|x-x_n\|^2\right\} \tag{PGD}$$

This algorithm take the objective function $F = f+g$ and linearize only the smooth part $f(x) \approx f(x_n)+\langle\nabla f(x_n), x-x_n\rangle$, leave the non-smooth part unchanged, and add a proximal term to make it strongly convex. The update can be rewritten as follows. By completing the square,

$$x_{n+1} = \operatorname*{argmin}_{x\in\mathbb{R}^d}\left\{g(x) + \frac{1}{2h}\|x - x_n + h\nabla f(x_n)\|^2\right\} = \text{prox}_{hg}(x_n - h\nabla f(x_n))$$

This corresponds to taking an explicit step of f, followed by an implicit step on g.

Example 3: LASSO continued For the LASSO problem, the iteration reads

$$x_{n+1} = \text{thresh}_{\lambda h}(x_n - h\nabla f(x_n))$$

In the literature, this is known as the *iterative shrinking-thresholding algorithm* (ISTA). For constrained optimization, proximal gradient descent is projected gradient descent.

Remark: Even for non-convex f, as long as $x^+ = \text{prox}_{hf}(x)$ is well-defined, we can have some sense of convergence. Since by definition, x^+ minimizes the objective $f(x') + \frac{1}{2h}\|x'-x\|^2$ in x', we can conclude that, for all $y \in \mathbb{R}^d$, we have

$$f(x^+) + \frac{1}{2h}\|x^+ - x\|^2 \leqslant f(y) + \frac{1}{2h}\|y-x\|^2$$

Let $y = x_\star$, we have

$$f(x^+) - f_\star \leqslant \frac{1}{2h}\left(\|x_\star - x\|^2 - \|x^+ - x\|^2\right) \leqslant \frac{1}{2h}\|x - x_\star\|^2$$

Thus, if we can implement PPM for arbitrary large step sizes $h > 0$, we can solve non-convex optimization.

6.2.2 Convergence Analysis

We study the convergence of proximal gradient descent, since it includes proximal point method as a special case ($f = 0$).

Theorem 6.2.3: Convergence of PGD

Let f be α_f-convex and β_f-smooth, and let g be α_g convex. Let the step size h satisfy $h \leqslant 1/\beta_f$. Let x^+ denote the next iterate of PGD started from x, and let $y \in \mathbb{R}^d$. Then,

$$(1 + \alpha_g h)\|y - x^+\|^2 \leqslant (1 - \alpha_f h)\|y - x\|^2 - 2h(F(x^+) - F(y))$$

In particular, if we set $y = x_\star$ and iterate, it yields

$$F(x_N) - F_\star \leqslant \frac{\alpha_f + \alpha_g}{2(\lambda_h^{-N} - 1)}\|x_0 - x_\star\|^2$$

where $\lambda_h = (1 - \alpha_f h)/(1 + \alpha_g h)$.

Proof. Let ψ_x denote the objective function in the definition of PGD. i.e.,

$$\psi_{x_n}(x) = f(x_n) + \langle \nabla f(x_n), x - x_n\rangle + g(x) + \frac{1}{2h}\|x - x_n\|^2$$

Note that $f(x_n)$ is a constant, $\langle \nabla f(x_n), x - x_n\rangle$ is a linear term, so they do not contribute to convexity. Thus, ψ_x is $(\alpha_g + 1/h)$-strongly convex with minimizer x^+. By quadratic growth inequality,

$$\psi_x(y) \geqslant \psi_x(x^+) + \frac{\alpha_g + 1/h}{2}\|y - x^+\|^2$$

On one hand, by α_f-convexity,

$$\begin{aligned}
\psi_x(y) &= f(x) + \langle \nabla f(x), y - x\rangle + g(y) + \frac{1}{2h}\|y - x\|^2 \\
&\leqslant f(y) - \langle \nabla f(x), y - x\rangle - \frac{\alpha_f}{2}\|y - x\|^2 + \langle \nabla f(x), y - x\rangle + g(y) + \frac{1}{2h}\|y - x\|^2 \\
&= F(y) + \frac{1/h - \alpha_f}{2}\|y - x\|^2
\end{aligned}$$

where we use $f(y) \geqslant f(x) + \langle \nabla f(x), y - x\rangle + \frac{\alpha_f}{2}\|y - x\|^2$. On the other hand, by β_f-smoothness,

$$\begin{aligned}
\psi_x(x^+) &= f(x) + \langle \nabla f(x), x^+ - x\rangle + g(x^+) + \frac{1}{2h}\|x^+ - x\|^2 \\
&\geqslant f(x^+) - \langle \nabla f(x), x^+ - x\rangle - \frac{\beta_f}{2}\|x^+ - x\|^2 + \langle \nabla f(x), x^+ - x\rangle + g(x^+) + \frac{1}{2h}\|x^+ - x\|^2 \\
&= F(x^+) + \frac{1/h - \beta_f}{2}\|x^+ - x\|^2 \geqslant F(x^+)
\end{aligned}$$

where we use $f(x^+) \leqslant f(x) + \langle \nabla f(x), x^+ - x\rangle + \frac{\beta_f}{2}\|x^+ - x\|^2$, and in the final inequality, $h \leqslant 1/\beta_f$. Combining these three and rearranging, we can have

$$\begin{aligned}
& F(y) + \frac{1/h - \alpha_f}{2}\|y - x\|^2 \geqslant F(x^+) + \frac{\alpha_g + 1/h}{2}\|y - x^+\|^2 \\
\implies \quad & (1 + \alpha_g h)\|y - x^+\|^2 \leqslant (1 - \alpha_f h)\|y - x\|^2 - 2h(F(x^+) - F(y))
\end{aligned}$$

Note that by taking $y = x$, it yields the descent property

$$F(x^+) - F(x) \leqslant -\frac{1 + \alpha_g h}{2h}\|x^+ - x\|^2 \leqslant 0$$

Instead, we set $y = x_\star$, and use discrete Grönwall to iterate, with $A = \frac{1-\alpha_f h}{1+\alpha_g h}$ and $B_n = -\frac{2h}{1+\alpha_g h}(F(x_{n+1}) - F(x_\star))$, we have

$$\begin{aligned}
& 0 \leqslant \|x_N - x_\star\|^2 \leqslant \left(\frac{1 - \alpha_f h}{1 + \alpha_g h}\right)^N \|x_0 - x_\star\|^2 - \sum_{n=1}^{N}\left(\frac{1 - \alpha_f h}{1 + \alpha_g h}\right)^{N-n} \frac{2h}{1 + \alpha_g h}(F(x_n) - F_\star) \\
\implies \quad & \frac{2h}{1 + \alpha_g h}(F(x_N) - F_\star)\sum_{n=1}^{N}\left(\frac{1 - \alpha_f h}{1 + \alpha_g h}\right)^{-n} \leqslant \|x_0 - x_\star\|^2 \qquad \text{(Descent Property)} \\
\implies \quad & \frac{1 + \alpha_g h}{\alpha_f h + \alpha_g h}\left(\left(\frac{1 - \alpha_f h}{1 + \alpha_g h}\right)^{-N} - 1\right)\frac{2h}{1 + \alpha_g h}(F(x_N) - F_\star) \leqslant \|x_0 - x_\star\|^2 \\
\implies \quad & F(x_N) - F_\star \leqslant \frac{\alpha_f + \alpha_g}{2(\lambda_h^{-N} - 1)}\|x_0 - x_\star\|^2
\end{aligned}$$

where $\lambda_h = (1 - \alpha_f h)/(1 + \alpha_g h)$. This completes the proof. □

The key feature here is that it essentially recovers the *smooth rate* for gradient descent despite the presence of non-smoothness in the objective. Thus, for the LASSO problem, we can solve it as quickly as if it were a smooth problem via ISTA.

Moreover, the one-step inequality above is the PGD analogue of the one-step inequality which holds for gradient desecent, and in turn, is the only property of gradient descent which plays a role in the proof of Nesterov acceleration. This naturally leads to an accelerated algorithm for composite optimization. Starting with $x_{-1} = x_0$, consider

$$x_{n+1} = x_n + \theta_n(x_n - x_{n-1}) - \mathrm{PGD}_{F,1/\beta}(x_n + \theta_n(x_n - x_{n-1})) \tag{APGD}$$

where $\mathrm{PGD}_{F,1/\beta}$ denotes one-step of proximal gradient descent on $F = f + g$ with step size $h = 1/\beta$.

Theorem 6.2.4: Convergence of APGD

Let f be convex and β-smooth, and let g be convex. Define the sequence $\lambda_0 = 0$ and $\lambda_{n+1} = \frac{1}{2}\left(1 + \sqrt{1 + 4\lambda_n^2}\right)$ for $n \in \mathbb{N}$. Set $\theta_n = (\lambda_n - 1)/\lambda_{n+1}$. Then, APGD satisfies

$$F(x_N) - F_\star \leqslant \frac{2\beta\|x_0 - x_\star\|^2}{N^2}$$

When applied to LASSO, this algorithm is known as fast ISTA or FISTA. Rates in the strongly convex setting can be obtained from the reduction on Lemma 3.1.1.

Chapter 7

Optimization in Non-Euclidean Spaces

7.1 Fenchel Duality

Definition 7.1.1: Convex Conjugate

Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be proper $(\operatorname{dom} f \neq \emptyset)$. The **convex conjugate** or **Fenchel-Legendre conjugate** of f is the function $f^* : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ defined by

$$f^*(y) := \sup_{x \in \mathbb{R}^d} \left(\langle x, y \rangle - f(x) \right)$$

For any proper function f, the conjugate f^* is always *convex and lower semicontinuous*, since it is a supremum of affine functions. Conversely, if f is a regular convex function (proper, convex, and lower-semicontinuous), then $f = f^{**}$. This will be proved later.

Example:

1. If $f(x) = \frac{1}{2}\langle x, Ax \rangle$ where $A \succ 0$, then $f^*(y) = \frac{1}{2}\langle y, A^{-1}y \rangle$.

2. If $f(x) = |x|^p/p$ for $p > 1$ and $x \in \mathbb{R}$, then $f^*(y) = |y|^q/q$ where $1/p + 1/q = 1$.

3. Let $|||\cdot|||$ denote a norm over $\mathbb{R}^d$, and let $|||\cdot|||_*$ denote the dual norm: $|||y|||_* := \sup_y \{\langle x, y \rangle : x \in \mathbb{R}^d, |||x||| \leqslant 1\}$. If $f(x) = |||x|||$, then $f^*(y) = \chi_{\mathcal{C}}(y)$ where $\mathcal{C} := \{y \in \mathbb{R} : |||y|||_* \leqslant 1\}$ is the closed unit ball in the dual norm.

Proof. 1. The convex conjugate can be written as

$$f^*(y) = \sup_{x \in \mathbb{R}^d} \left(\langle x, y \rangle - \frac{1}{2}\langle x, Ax \rangle \right) = \sup_{x \in \mathbb{R}^d} \left\langle x, y - \frac{1}{2}Ax \right\rangle$$

Inside is a quadratic function $x^T y - \frac{1}{2}x^T Ax$, and it is maximized at $x = A^{-1}y$. Thus, the convex conjugate is

$$f^*(y) = \left\langle A^{-1}y, y - \frac{1}{2}AA^{-1}y \right\rangle = \frac{1}{2}\langle y, A^{-1}y \rangle$$

2. The convex conjugate can be written as

$$f^*(y) = \sup_{x\in\mathbb{R}^d}\left(xy - \frac{|x|^p}{p}\right)$$

We need to find the maximum point of $h(x) = xy - |x|^p/p$. When $y \geqslant 0$, the optimal point would be achieve at $x \geqslant 0$ (on $x < 0$ $h(x)$ is always negative). $h'(x) = y - x^{p-1}$. So it is maximized at $x = y^{1/(p-1)}$. The convex conjugate is then

$$f^*(y) = y^{\frac{p}{p-1}} - \frac{y^{\frac{p}{p-1}}}{p} = \frac{p-1}{p}y^{\frac{p}{p-1}} = \frac{y^q}{q}$$

Similarly, when $y < 0$, the optimal point would be at $x < 0$. $h'(x) = y + (-x)^{p-1}$. So it is maximized at $x = -(-y)^{1/(p-1)}$. The convex conjugate is then

$$f^*(y) = -(-y)^{\frac{1}{p-1}}(-(-y)) - \frac{(-y)^{\frac{p}{p-1}}}{p} = \frac{p-1}{p}(-y)^{\frac{p}{p-1}} = \frac{(-y)^q}{q}$$

Therefore, we conclude that $f^*(y) = |y|^q/q$.

3. The convex conjugate can be written as

$$f^*(y) = \sup_{x\in\mathbb{R}^d}\left(\langle x, y\rangle - |\!|\!|x|\!|\!|\right)$$

When $|\!|\!|y|\!|\!|_* \leqslant 1$, for $|\!|\!|x|\!|\!| \leqslant 1$, we have $\langle x, y\rangle \leqslant |\!|\!|x|\!|\!||\!|\!|y|\!|\!|_*$ by the definition of dual norm. Indeed, this can be generalized to $x \in \mathbb{R}^d$ since we can scale x by any constant and with the property of norm, $\langle \alpha x, y\rangle \leqslant |\alpha||\!|\!|x|\!|\!||\!|\!|y|\!|\!|_* = |\!|\!|\alpha x|\!|\!||\!|\!|y|\!|\!|_*$. Therefore, we have

$$\langle x, y\rangle \leqslant |\!|\!|x|\!|\!||\!|\!|y|\!|\!|_* \leqslant |\!|\!|x|\!|\!| \quad\Longrightarrow\quad \langle x, y\rangle - |\!|\!|x|\!|\!| \leqslant 0$$

Since when $x = 0$, we have $\langle x, y\rangle - |\!|\!|x|\!|\!| = 0$, the objective is maximized at $x = 0$, so $f^\star(y) = 0$.

When $|\!|\!|y|\!|\!|_* > 1$, by the definition of dual norm, for any $\epsilon > 0$, there exists $\tilde{x}$ with $|\!|\!|x|\!|\!| = 1$ such that

$$\langle \tilde{x}, y\rangle \geqslant |\!|\!|y|\!|\!|_* - \epsilon$$

Now for any $t > 0$ set $x = t\tilde{x}$, and we have

$$\langle x, y\rangle - |\!|\!|x|\!|\!| = t\langle \tilde{x}, y\rangle - t|\!|\!|\tilde{x}|\!|\!| \geqslant t\left(|\!|\!|y|\!|\!|_* - \epsilon - |\!|\!|\tilde{x}|\!|\!|\right) \geqslant t(|\!|\!|y|\!|\!|_* - \epsilon - 1)$$

Since $|\!|\!|y|\!|\!|_* > 1$, we can choose ϵ small enough so that $|\!|\!|y|\!|\!|_* - \epsilon - 1 > 0$. Then, send $t \to \infty$ yields that

$$\langle x, y\rangle - |\!|\!|x|\!|\!| \geqslant t(|\!|\!|y|\!|\!|_* - \epsilon - 1) \to \infty$$

Therefore, we conclude that $f^*(y) = \chi_{\mathcal{C}}(y)$.

□

Before formally establishing further properties of this duality, we explain the origin of this concept in classical mechanics.

7.1.1* Connection with Classical Mechanics

Newton's law of motion states that the trajectory $(x_t)_{t\geqslant 0}$ of a particle of mass m obeys the differential equation $m\ddot{x}_t = F(x_t)$, where F is the force. The force is typically given as the gradient of a potential: $F = -\nabla\phi$.

In 1662, Pierre de Fermat proposed an explanation for the law of refraction via his principle of least action: light takes the path which minimizes the total travel time. Is there such a principle for classical mechanics as well? In 1760, Joseph-Louis Lagrange found such a variational principle: let $L(x,v) = \frac{1}{2}m\|v\|^2 - \phi(x)$ denote the *Lagrangian*, the difference of kinetic energy and potential energy. The action functional is

$$\mathcal{A}\left((x_t)_{t\in[0,T]}\right) := \int_0^T L(x_t,\dot{x}_t)\,\mathrm{d}t$$

Lagrangian mechanics states that if a particle starts at x_0 at time 0, and ends at x_T at time T, then the path it takes in between is a stationary point of the action functional subject to the endpoint constraints.

We solve for the path using calculus of variations. Let $x_{[0,T]} = (x_t)_{t\in[0,T]}$ be a shorthand for the path. If $x_{[0,T]}$ is a stationary point, it means that for any perturbation $\delta x_{[0,T]}$, the difference $\mathcal{A}\left(x_{[0,T]} + \delta x_{[0,T]}\right) - \mathcal{A}\left(x_{[0,T]}\right)$ should vanish to first order in $\delta x_{[0,T]}$. The endpoint constraints require that $\delta x_0 = \delta x_T = 0$. Thus,

$$\begin{aligned}\mathcal{A}\left(x_{[0,T]} + \delta x_{[0,T]}\right) - \mathcal{A}\left(x_{[0,T]}\right) &= \int_0^T \left(L(x_t+\delta x_t, \dot{x}_t+\delta\dot{x}_t) - L(x_t,\dot{x}_t)\right)\,\mathrm{d}t \\ &= \int_0^T \left(\langle\nabla_x L(x_t,\dot{x}_t), \delta x_t\rangle + \langle\nabla_v L(x_t,\dot{x}_t), \delta\dot{x}_t\rangle\right)\,\mathrm{d}t + o(\|\delta x\|) \\ &= \int_0^T \langle\nabla_x L(x_t,\dot{x}_t) - \partial_t\nabla_v L(x_t,\dot{x}_t), \delta x_t\rangle\,\mathrm{d}t + o(\|\delta x\|) \qquad \text{(Integration by part)}\end{aligned}$$

The stationary point therefore satisfies the *Euler-Lagrange equation*

$$\partial_t\nabla_v L(x_t,\dot{x}_t) = \nabla_x L(x_t,\dot{x}_t)$$

For $L(x,v) = \frac{1}{2}m\|v\|^2 - \phi(x)$, it recovers Newton's equation.

We now introduce the Legendre transform. Define the *Hamiltonian* H to be the convex conjugate of L with respect to the v variavle, i.e.,

$$H(x,p) := \sup_{v\in\mathbb{R}^d}\left(\langle p,v\rangle - L(x,v)\right)$$

The first order condition reveals that

$$\nabla_v\left(\langle p,v\rangle - L(x,v)\right) = p - \nabla_v L(x,v) = 0 \quad\Longrightarrow\quad p = \nabla_v L(x,v)$$

Instead of working with the variable (x,v), we now work with the variables (x,p). We will argue that a regular convex function f satisfies $f = f^{**}$. Assuming $v\mapsto L(x,v)$ is regular convex, it yields the dual representation

$$L(x,v) = \sup_{p\in\mathbb{R}^d}\left(\langle p,v\rangle - H(x,p)\right)$$

and the first order condition for this problem yields the inverse transformation

$$v = \nabla_p H(x, p)$$

Thus, if we define $p_t = \nabla_v L(x_t, \dot{x}_t)$, we can reformulate the Euler-Lagrange equation as follows. First, $\dot{x}_t = v_t = \nabla_p H(x_t, p_t)$. Also, by *envelop theorem*, with mild condition, if $f(x) = \sup_y g(x, y)$, then $\nabla f(x) = \nabla_x g(x, y^\star(x))$, where $y^\star(x)$ achieves the supremum. In our case,

$$\nabla_x H(x, p) = \nabla_x \sup_{v \in \mathbb{R}^d} \{\langle p, v\rangle - L(x.v)\} = -\nabla_x L(x, v^\star(p)), \quad v^\star(p) = \nabla_p H(x, p)$$

Thus, $\nabla_x H(x, p) = -\nabla_x L(x, v)$ by envelop theorem. So, $\dot{p}_t = \partial_t \nabla_v L(x_t, \dot{x}_t) = \nabla_x L(x_t, \dot{x}_t) = -\nabla_x H(x_t, p_t)$, where the middle equality follows from Euler-Lagrangian. Insummary,

$$\dot{x}_t = \nabla_p H(x_t, p_t), \quad \dot{p}_t = -\nabla_x H(x_t, p_t)$$

These are known as *Hamilton's equation*, and it is easy to verify that they conserve the Hamiltonian: $\partial_t H(x_t, p_t) = 0$. For our running example, $p = mv$ is the *momentum*, and $H(x, p) = \langle p, \frac{p}{m}\rangle - \frac{1}{2}m \left\|\frac{p}{m}\right\|^2 + \phi(x) = \frac{1}{2m}\|p\|^2 + \phi(x)$ is the total energy. Hamilton's equations read

$$m\dot{x}_t = p_t, \quad p_t = -\nabla \phi(x_t)$$

Specialize now to the case where the Lagrangian only depends on v, (i.e., $\phi = 0$, particle move in straight lines). Define the following function of space and time:

$$u(t, x) = \inf \left\{ \int_0^t L(\dot{x}_s) \,\mathrm{d}s + f(x_0) \;\middle|\; x : [0, t] \to \mathbb{R}^d, x_t = x \right\}$$

In word, we minimized the action functional up to time t, subject to the constraint that we hit x at time t. We also add an initial cost $f(x_0)$. The function u aresembles the notion of the *value function* or *cost-to-go function* in dynamic programming, and indeed it satisfies a dynamic programming principle: for $0 \leqslant s < t$,

$$u(t, y) = \inf_{x \in \mathbb{R}^d} \left\{ (t - s) L\left(\frac{y - x}{t - s}\right) + u(s, x) \right\} \tag{7.1}$$

The heuristic derivation of this identity is as folows: consider a potential candidate x for the value of the path at time s. Given x, the best possible value of $\int_0^s L(\dot{x}_r) \,\mathrm{d}r + f(x_0)$ is $u(s, x)$. For the remaining part, it just gives by straight line

$$\int_s^t L(\dot{x}_r) \,\mathrm{d}r = (t - s) \frac{1}{t - s} \int_s^t L(\dot{x}_r) \,\mathrm{d}r \geqslant (t - s) L\left(\frac{1}{t - s} \int_s^t \dot{x}_r \,\mathrm{d}r\right) = (t - s) L\left(\frac{y - x}{t - s}\right)$$

where the middle inequality holds by Jensen's inequality, the final equality holds since $x_t = y$ by endpoint constraint. The lower bound is achieved if $\dot{x}_r$ is constant for $r \in [s, t]$, i.e., $x_{[s,t]}$ is a straight line.

In particular, since $u(0, \cdot) = f$, we see that

$$u(t, y) = \inf_{x \in \mathbb{R}^d} \left\{ tL\left(\frac{y - x}{t}\right) + f(x) \right\}$$

Definition 7.1.2: Hopf-Lax Semigroup

The **Hopf-Lax semigroup** $(Q_t)_{t\geqslant 0}$ is a family of operators which maps functions to functions, such that

$$Q_t f(y) = \inf_{x\in\mathbb{R}^d} \left\{ tL\left(\frac{y-x}{t}\right) + f(x) \right\}$$

This is a *semigroup of operators*, not the common definition of semigroup. The dynamic programming principle 7.1 shows that $Q_t f = Q_{t-s}(Q_s f)$. Thus, we have the properties $Q_0 = id$, $Q_{s+t} = Q_s Q_t = Q_t Q_s$ for all $s, t \geqslant 0$, which are the defining properties of a semigroup. In convex analysis, the corresponding operation is known as the infimal convolution.

Definition 7.1.3: Infimal Convolution

Let $f, g : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$. The **infimal convolution** of f and g, denoted $f \,\Box\, g$, is the function defined by

$$(f \,\Box\, g)(y) = \inf_{x\in\mathbb{R}^d} \{f(x) + g(y-x)\}$$

In this notation, $Q_t f = tL(\cdot/t) \,\Box\, f$. The operation of convex conjugation turns addition into infimal convolution and vice versa.

Proposition 7.1.4: Convex Conjugation and Infimal Convolution

Let f, g be regular convex functions. Then,

$$(f \,\Box\, g)^* = f^* + g^*$$

Conversely, if $\operatorname{int}\operatorname{dom} f \cap \operatorname{int}\operatorname{dom} g \neq \emptyset$, then

$$(f+g)^* = f^* \,\Box\, g^*$$

Proof. For the first statement, note that

$$\begin{aligned}(f \,\Box\, g)^*(y) &= \sup_{x\in\mathbb{R}^d} \{\langle x, y\rangle - (f \,\Box\, g)(x)\} = \sup_{x\in\mathbb{R}^d} \{\langle x, y\rangle - \inf_{z\in\mathbb{R}^d} \{f(z) + g(x-z)\}\} \\ &= \sup_{x,z\in\mathbb{R}^d} \{\langle z, y\rangle - f(z) + \langle x-z, y\rangle - g(x-z)\} = f^*(y) = g^*(y)\end{aligned}$$

The first statement also implies that $(f^* \,\Box\, g^*)^* = f^{**} + g^{**} = f + g$ (which will be proved later). By applying convex conjugate to both sides, we have the second statement if $f^* \,\Box\, g^*$ equals its double conjugate. For this, we need to know that $f^* \,\Box\, g^*$ is regular convex, which follows from the condition on the domains, see Rockafeller[12] Theorem 16.4. □

There is an analogy with the Fourier transform, which transforms convolution into multiplication. Recall that for $f, g : \mathbb{R}^d \to \mathbb{C}$, the Fourier transform is given by $\mathcal{F}(f(\xi)) = \int f(x)\exp(-2\pi i\langle \xi, x\rangle)\,\mathrm{d}x$, the convolution is given by $(f * g)(y) = \inf f(x)g(y-x)\,\mathrm{d}x$, and we have the key property $\mathcal{F}(f * g) = \mathcal{F}(f)\mathcal{F}(g)$.
To see a connection more precisely, note that we usually work with the algebra $(+, \cdot)$. Now introduce a new structure, consisting $(\min, +)$. The identity element is $+\infty$, and $+$ distributes over min, i.e., $x + \min(y, z) = \min(x+y, x+z)$.

However, we also lose some properties, e.g., not every element has an inverse for the min operation. This is called a *min-plus algebra* despite it is not technically an algebra. More acurately, it is the *tropical semiring.*
If we think of integrals as continuous summations, then convolution is a sum of products. Infimal convlution is a min of sums. Hence, infimal convolution is the tropical analogue of convolution. Some further analogies are listed below.

$(+, \times)$	$(\min, +)$
convolution	infimal convolution
Fourier transform	convex conjugate
Gaussians	convex quadratics
diffusion processes	gradient flow
heat equation	Hamilton-Jacobi equation
heat semigroup	Hopf-Lax semigroup

We conclude this discussion by using this perspective to show that the Hopf-Lax semigroup solves the following PDE, known as the *Hamilton-Jacobi equation*

$$\partial_t u + H(\nabla_x u) = 0$$

The proof is patterned on the following derivation of the solution to the heat equation $\partial_t u = \Delta u$ with initial condition $u(0, \cdot) = f$; here $\Delta u = \sum_{i=1}^d \partial_i^2 u$ is the Laplacian. If we take the Fourier transform of both sides of the equation, then $\partial_t \mathcal{F}(u) = \mathcal{F}(\Delta u) = -4\pi^2 \|\cdot\|^2 \mathcal{F}(u)$, where the last equality follows from differentiating the Fourier transform under the integral. This implies that $\partial_t \log \mathcal{F}(u) = -4\pi^2 \|\cdot\|^2$, or $\mathcal{F}(u(t, \cdot)) = \mathcal{F}(f \exp(-4\pi^2 t \|\cdot\|^2))$. Using the fact that the inverse Fourier transform transforms multiplication into convolution, one can then show that $u(t, \cdot) = f * \mathcal{F}(\exp(-4\pi^2 t \|\cdot\|^2)) = f * \mathrm{Normal}(0, 4tI)$.
In the same way, we start with Hamilton-Jacobi equation and take the convex conjugate of both sides. Using the shorthand notation $f_t = u(t, \cdot)$, and since $f_t^*(p) = \sup_{v \in \mathbb{R}^d} \{\langle p, v \rangle - f_t(v)\}$ with the supremum attained at $v = \nabla f_t^*(p)$, we have

$$\partial_t f_t^*(p) = -\partial_t f_t(\nabla f_t^*(p)) = H(\nabla f_t(\nabla f_t^*(p))) = H(p)$$

Hence $f_t^* = tH + f$ and $f_t = (tH)^* \,\Box\, f = tL(\cdot/t) \,\Box\, f$. Thus, the solution is given by the Hopf-Lax semigroup as claimed.

When $L = H = \frac{1}{2}\|\cdot\|^2$, the Hamilton-Jacobi equation becomes $\partial_t u + \frac{1}{2}\|\nabla_x u\|^2 = 0$ and the Hopf-Lax semigroup $Q_t f$ coincides with the Moreau-Yosida envelop, since

$$Q_t f(y) = \inf_{x \in \mathbb{R}^d} \left\{ tL\left(\frac{y-x}{t}\right) + f(x) \right\}, \quad L(v) = \frac{1}{2}\|v\|^2$$

and thus

$$Q_t f(y) = \inf_{x \in \mathbb{R}^d} \left(f(x) + \frac{1}{2t}\|y - x\|^2 \right) = \text{ Moreau-Yosida envelop}$$

This yields an unexpected connection between the Hamilton-Jacobi equation and the PPM.

7.1.2 Duality Correspondences

Theorem 7.1.5: Fenchol-Young Inequality

Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be regular and convex. Then,

$$f(x) + f^*(p) \geqslant \langle p, x\rangle, \quad \text{for all } p, x \in \mathbb{R}^d$$

Moreover, equality holds if and only if $p \in \partial f(x)$, if and only if $x \in \partial f^*(p)$.

Proof. From the definition of f^*, we have

$$f^*(p) \geqslant \langle x, p\rangle - f(x)$$

which is exactly the inequality. If equality holds, then for any $p', x' \in \mathbb{R}^d$,

$$f(x') \geqslant \langle p, x'\rangle - f^*(p) = \langle p, x'\rangle + f(x) - \langle p, x\rangle = f(x) + \langle p, x' - x\rangle$$

$$f^*(p) \geqslant \langle p', x\rangle - f(x) = \langle p', x\rangle + f^*(p) - \langle p, x\rangle = f^*(p) + \langle x, p' - p\rangle$$

i.e., $p \in \partial f(x)$ and $x \in \partial f^*(p)$. Conversely, if $p \in \partial f(x)$, then

$$f^\star(p) = \sup_{x' \in \mathbb{R}^d} \{\langle p, x'\rangle - f(x')\} \leqslant \langle p, x\rangle - f(x)$$

which is only possible if this is an equality. □

Theorem 7.1.6: Double Conjugation

let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$. Then, $f \geqslant f^{**}$. Moreover, if f is regular and convex, then equality holds: $f = f^{**}$.

Proof. For the first statement,

$$f^{**}(z) = \sup_{y \in \mathbb{R}^d} \left\{ \langle y, z\rangle - \sup_{x \in \mathbb{R}^d} \{\langle x, y\rangle - f(x)\} \right\} = \sup_{y \in \mathbb{R}^d} \inf_{x \in \mathbb{R}^d} \{\langle y, z - x\rangle + f(x)\} \leqslant f(z) \tag{7.2}$$

where the last inequality holds by choosing $x = z$.

Now assume that f is regular and convex. If $z \in \operatorname{int} \operatorname{dom} f$, then by Theorem 5.1.14, there exists $p \in \partial f(z)$, so that $f(x) \geqslant f(z) + \langle p, x - z\rangle$ for all $x \in \mathbb{R}^d$. By taking $y = p$,

$$f^{**}(z) \geqslant \inf_{x \in \mathbb{R}^d} \{\langle p, z - x\rangle + f(x)\} \geqslant f(z)$$

which proves the equality for such z. We omit the proof for $z \notin \operatorname{int} \operatorname{dom} f$, see, e.g., Rockafeller[12] Theorem 12.2. □

This result implies that in general, if $f_\star$ is the largest convex and lower semicontinuous function which is smaller than f, then $f_\star = f^{**}$. Indeed, $f_\star \geqslant f^{**}$ by definition, whereas $f \geqslant f_\star$ implies $f^{**} \geqslant (f_\star)^{**} = f_\star$.

The proof above also shows that whenever $\partial f(x) \neq \emptyset$, then $f(x) = f^{**}(x)$. In particular, if $x_\star$ is a minimizer of

f, then $0 \in \partial f(x_\star)$ and $f(x_\star) = f^{**}(x_\star)$; Moreover, by taking $y = 0$ in 7.2 we see that $\inf f = \inf f^{**}$. Thus, we can start with a non-convex function f and 'convexify' it by replacing it with f^{**} while preserving the optimal value, although this is seldom useful in practice.

Properties of f are often reflected as 'dual' properties for f^*. For example, if f is regular convex, then the following holds: (Rockafeller[12])

- f is Lipschitz if and only if $\operatorname{dom} f^*$ is bounded.
- $\operatorname{epi} f$ contains no non-vertical half-lines if and only if $\operatorname{dom} f^* = \mathbb{R}^d$.
- f has no lines along which it is affine if and only if $\operatorname{int}\operatorname{dom} f^* \neq \emptyset$.
- f has bounded level sets if and only if $0 \in \operatorname{int}\operatorname{dom} f^*$.
- f is differentiable at x with $\nabla f(x) = p$ if and only if $(p, f^*(p))$ is an exposed point of $\operatorname{epi} f^*$. (An *exposed point* of a convex set is a point at which some linear function attains its strict maximum over the convex set.)

For our purposes, we are most interested in conditions under which ∇f is a well-defined bijection from an open convex set $\mathcal{C}$ to its image $\nabla f(\mathcal{C})$, with inverse given by $(\nabla f)^{-1} = \nabla f^*$. In this case, the correspondence between f and f^* is known as the *Legendre transformation* and we informally discussed it in the previous section.

Definition 7.1.7: Essentially Smooth/Essentially Strictly Convex

Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be regular convex.

- We say that f is **essentially smooth** if f is differentiable on $\mathcal{C} = \operatorname{int}\operatorname{dom} f$ and $\lim_{n\to\infty} \|\nabla f(x_n)\| \to \infty$ whenever $\{x_n\}_{n\in\mathbb{N}}$ is a sequence in $\mathcal{C}$ converging to $\partial\mathcal{C}$.
- We say that f is **essentially strictly convex** if f is strictly convex on every convex subset of $\operatorname{dom}\partial f = \{x \in \mathbb{R}^d : \partial f(x) \neq \emptyset\}$.

Lemma 7.1.8

A regular convex function f is essentially smooth if and only if f^* is essentially strictly convex.

Theorem 7.1.9

Let f be regular, strictly convex, and essentially smooth over $\mathcal{C} = \operatorname{int}\operatorname{dom} f$. Then, $f^\star$ is regular, strictly convex, and essentially smooth over $\mathcal{C}^\star = \operatorname{int}\operatorname{dom} f^*$. Moreover, $\nabla f : \mathcal{C} \to \mathcal{C}^\star$ is a continuous bijection with $(\nabla f)^{-1} = \nabla f^\star$.

Definition 7.1.10: Legendre Type

We say that a function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is of **Legendre type** if it satisfies the assumption of Theorem 7.1.9.

All these are in Rockafeller[12] chapter 26.

To summarize, the condition that f is regular convex ensures duality at the level of $f = f^{**}$. The condition that

f is of Legendre type ensures duality at the level of $(\nabla f)^{-1} = \nabla f^*$. Note also that if f, f^* are sufficiently smooth, then by differentiating the equality $\nabla f(\nabla f^*) = \mathrm{Id}$, by chain rule one obtain the identity

$$\nabla^2 f \circ \nabla f^\star = (\nabla^2 f^*)^{-1}$$

In particular, $\nabla^2 f \succcurlyeq \alpha I$ is equivalent to $(\nabla^2 f^*)^{-1} \preccurlyeq \alpha^{-1} I$. i.e., there is a duality between the properties of strong convexity and smoothness. Let's prove this without assuming differentiability.

Lemma 7.1.11: Convexity-Smoothness Duality

Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be regular and α-convex for some $\alpha > 0$. Then, f^* is α^{-1}-smooth.

Proof. By the duality correspondence, $\mathrm{dom}\, f^\star = \mathbb{R}^d$ and f^* is differentiable everywhere. For two points $y, y' \in \mathbb{R}^d$, let $x, x' \in \mathbb{R}^d$ achieve the suprema in the definitions of $f^*(y)$, $f^*(y')$ respectively. By Fenchel-Yong inequality's equality condition, we have $x = \nabla f^*(y)$ and $x' = \nabla f^*(y')$. Then, by strong convexity of $f - \langle \cdot, y \rangle$,

$$f(x') - \langle x', y \rangle \geqslant f(x) - \langle x, y \rangle + \frac{\alpha}{2}\|x' - x\|^2$$

Adding this to the analogous inequality with x and x' swapped, i.e.,

$$f(x) - \langle x, y' \rangle \geqslant f(x') - \langle x', y' \rangle + \frac{\alpha}{2}\|x' - x\|^2$$

We can have

$$\alpha\|\nabla f^*(y') - \nabla f^*(y)\|^2 = \alpha\|x' - x\|^2 \leqslant \langle x, y \rangle + \langle x', y' \rangle - \langle x', y \rangle - \langle x, y' \rangle = \langle x' - x, y' - y \rangle$$

Use Cauchy-Schwarz, we have

$$\alpha\|\nabla f^*(y') - \nabla f^*(y)\|^2 \leqslant \|x' - x\|\|y' - y\| = \|\nabla f^*(y') - \nabla f^*(y)\|\|y' - y\|$$

Rearranging, we have $\|\nabla f^*(y') - \nabla f^*(y)\| \leqslant \alpha^{-1}\|y' - y\|$ which implies

$$\langle \nabla f^*(y') - \nabla f^*(y), y' - y \rangle \leqslant \|\nabla f^*(y') - \nabla f^*(y)\|\|y' - y\| \leqslant \alpha^{-1}\|y' - y\|^2$$

So f^* is α^{-1}-smooth. □

To end this section, we do a refined analysis on proximal point method. First we need a corollary from above.

Corollary 7.1.12: Contractivity of the Proximal Operator

Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be α-convex. Then, prox_f is $1/(1+\alpha)$-Lipschitz.

Proof. We can write

$$\mathrm{prox}_f(y) = \operatorname*{argmin}_{x \in \mathbb{R}^d} \left\{ f(x) + \frac{1}{2}\|y - x\|^2 \right\} = -\operatorname*{argmax}_{x \in \mathbb{R}^d} \left\{ \langle x, y \rangle - f(x) - \frac{1}{2}\|x\|^2 \right\}$$

by expanding the square. This shows that $-\mathrm{prox}_f$ is the gradient of the convex conjugate of the function $f + \frac{1}{2}\|\cdot\|^2$ by the Fenchel-Young equality condition, which is $(1+\alpha)$-convex. □

For a closed convex set $\mathcal{C}$, $\pi_{\mathcal{C}} = \mathrm{prox}_{\chi_{\mathcal{C}}}$, so this corollary recovers the non-expansivity of projection. It also shows that proximal point method with an α-convex function contracts with rate $1/(1+\alpha h)$.

To avoid technical difficulties, assume that f is convex and differentiable everywhere. We have shown before that if f is differentiable, it holds that $x^+ = x - h\nabla f(x^+)$. Expanding the square, we obtain

$$\begin{aligned}
\|x^+ - y\|^2 &= \|x - h\nabla f(x^+) - y\|^2 \\
&= \|x-y\|^2 - 2h\langle x - y, \nabla f(x^+)\rangle + h^2\|\nabla f(x^+)\|^2 \\
&= \|x-y\|^2 - 2h\langle x^+ + h\nabla f(x^+) - y, \nabla f(x^+)\rangle + h^2\|\nabla f(x^+)\|^2 \\
&= \|x-y\|^2 - 2h\langle x^+ - y, \nabla f(x^+)\rangle - h^2\|\nabla f(x^+)\|^2 \\
&\leqslant \|x-y\|^2 - 2h(f(x^+) - f(y)) - h^2\|\nabla f(x^+)\|^2 && \text{(Convexity)}
\end{aligned}$$

Let $y = x_\star$, we will get

$$\|x^+ - x_\star\|^2 \leqslant \|x - x_\star\|^2 - 2h(f(x^+) - f_\star) - h^2\|\nabla f(x^+)\|^2 \tag{7.3}$$

Define the Lyapunov function

$$\mathcal{L}_n = n^2h^2\|\nabla f(x_n)\|^2 + 2nh(f(x_n) - f_\star) + \|x_n - x_\star\|^2$$

where $\{x_n\}_{n\in\mathbb{N}}$ are the iterates of proximal point method. By Corollary 7.1.12, proximal operator is contractive with rate $1/(1+\alpha h)$. In our case, $\alpha = 0$. The Lyapunov function at step $n+1$ is

$$\begin{aligned}
\mathcal{L}_{n+1} &= (n+1)^2h^2\|\nabla f(x_{n+1})\|^2 + 2(n+1)h(f(x_{n+1}) - f_\star) + \|x_{n+1} - x_\star\|^2 \\
&\leqslant (n+1)^2h^2\|\nabla f(x_{n+1})\|^2 + 2(n+1)h(f(x_{n+1}) - f_\star) + \|x_n - x_\star\|^2 - 2h(f(x_{n+1}) - f_\star) - h^2\|\nabla f(x_{n+1})\|^2 \\
& && \text{(Equation 7.3)} \\
&= (n^2+2n)h^2\|\nabla f(x_{n+1})\|^2 + 2nh(f(x_{n+1}) - f_\star) + \|x_n - x_\star\|^2 \\
&= (n^2+2n)h^2\|\nabla f(x_{n+1})\|^2 + 2nh(f(x_{n+1}) - f(x_n) + f(x_n) - f_\star) + \|x_n - x_\star\|^2 \\
&\leqslant (n^2+2n)h^2\|\nabla f(x_{n+1})\|^2 + 2nh(f(x_n) - f_\star) + 2nh\langle \nabla f(x_{n+1}), x_{n+1} - x_n\rangle + \|x_n - x_\star\|^2 && \text{(Convexity)} \\
&= (n^2+2n)h^2\|\nabla f(x_{n+1})\|^2 + 2nh(f(x_n) - f_\star) + 2nh\langle \nabla f(x_{n+1}), -h\nabla f(x_{n+1})\rangle + \|x_n - x_\star\|^2 \\
& && (x^+ = x - h\nabla f(x^+)) \\
&= (n^2+2n)h^2\|\nabla f(x_{n+1})\|^2 + 2nh(f(x_n) - f_\star) - 2nh^2\|\nabla f(x_{n+1})\|^2 + \|x_n - x_\star\|^2 \\
&= n^2h^2\|\nabla f(x_{n+1})\|^2 + 2nh(f(x_n) - f_\star) + \|x_n - x_\star\|^2
\end{aligned}$$

Now to change the $\nabla f(x_{n+1})$ to $\nabla f(x_n)$, we use the nonexpansivity of proximal operator, where

$$\left\|\mathrm{prox}_f(x_n) - \mathrm{prox}_f(x_{n-1})\right\| = \|x_{n+1} - x_n\| \leqslant \|x_n - x_{n-1}\|$$

by definition of PPM. Then, we use $x^+ = x - h\nabla f(x^+)$ to obtain

$$\|x_{n+1} - x_n\| \leqslant \|x_n - x_{n-1}\| \quad \Longrightarrow \quad \|\nabla f(x_{n+1})\| \leqslant \|\nabla f(x_n)\|$$

Combining these, we have

$$\mathcal{L}_{n+1} \leqslant n^2h^2\|\nabla f(x_{n+1})\|^2 + 2nh(f(x_n) - f_\star) + \|x_n - x_\star\|^2 \leqslant n^2h^2\|\nabla f(x_n)\|^2 + 2nh(f(x_n) - f_\star) + \|x_n - x_\star\|^2 = \mathcal{L}_n$$

Therefore, the Lyapunov function is decreasing. Use $\mathcal{L}_N \leqslant \mathcal{L}_0$, we have

$$N^2h^2\|\nabla f(x_N)\|^2 \leqslant \mathcal{L}_N \leqslant \mathcal{L}_0 = \|x_0 - x_\star\|^2 \quad \Longrightarrow \quad \|\nabla f(x_N)\| \leqslant \frac{\|x_0 - x_\star\|}{Nh}$$

By convexity, we can have

$$f_\star - f(x_N) \geqslant \langle \nabla f(x_N), x_\star - x_N\rangle \quad \Longrightarrow \quad 2Nh(f(x_N) - f_\star) \leqslant 2Nh\langle \nabla f(x_N), x_N - x_\star\rangle$$

Using Cauchy-Schwarz inequality and $ab \leqslant \frac{a^2}{2} + \frac{b^2}{2}$, we have

$$2Nh(f(x_N) - f_\star) \leqslant 2Nh\|\nabla f(x_N)\|\|x_N - x_\star\| \leqslant N^2h^2\|\nabla f(x_N)\|^2 + \|x_N - x_\star\|^2$$

This leads to

$$4Nh(f(x_N) - f_\star) \leqslant N^2h^2\|\nabla f(x_N)\|^2 + 2Nh(f(x_N) - f_\star) + \|x_N - x_\star\|^2 = \mathcal{L}_N \leqslant \mathcal{L}_0 = \|x_0 - x_\star\|^2$$

Therefore, we have the final results:

$$\|\nabla f(x_N)\| \leqslant \frac{\|x_0 - x_\star\|}{Nh}, \qquad f(x_N) - f_\star \leqslant \frac{\|x_0 - x_\star\|^2}{4Nh}$$

If $h \searrow 0$ while $Nh \to t$, it recovers the sharp bound of GF in Theorem 2.1.5.

If we instead, assume that f is differentiable everywhere but not necessarily convex, and satisfies PŁ inquality with constant $\alpha > 0$. Then, we can also get a sharp rate of convergence of proximal point method in this setting, which turns out to be non-trivial.

For any $x \in \mathbb{R}^d$, let $(Q_t)_{t\geqslant 0}$ denote the Hopf-Lax semigroup and $x_t = \operatorname{prox}_{tf}(x)$. The following derivation is from Chen et al.[4]. Let

$$f_{t,x}(z) := f(z) + \frac{1}{2t}\|z - x\|^2, \quad x_t = \operatorname*{argmin}_{z\in\mathbb{R}^d}\left\{f(z) + \frac{1}{2t}\|z - x\|^2\right\} = \operatorname*{argmin}_{z\in\mathbb{R}^d} f_{t,x}, \quad Q_tf(x) = f_{t,x}(x_t)$$

Then, $x_t = \operatorname{prox}_{tf}(x)$ and $x \mapsto f_{t,x}(x_t)$ is the Moreau-Yosida envelop of f. Using the first-order optimality condition,

$$0 = \nabla f(x_t) + \frac{1}{t}(x_t - x) \quad \Longrightarrow \quad x_t = x - t\nabla f(x_t)$$

Since the Moreau-Yosida envelop solves the Hamilton-Jacobi equation as discussed before, $\partial_t u + \frac{1}{2}\|\nabla_x u\|^2 = 0$, substitute in, we have

$$\begin{aligned}
&\partial_t f_{t,x}(x_t) + \frac{1}{2}\left\|\nabla_x f_{t,x}(x_t)\right\|^2 = 0\\
\implies\quad &\partial_t f_{t,x}(x_t) + \frac{1}{2}\left\|-\frac{1}{t}(x_t - x)\right\|^2 = 0\\
\implies\quad &\partial_t f_{t,x}(x_t) = \partial_t Q_t f(x) = -\frac{1}{2t^2}\|x_t - x\|^2
\end{aligned}$$

It can be written as

$$\begin{aligned}
\partial_t f_{t,x}(x_t) &= -\frac{1}{2t^2}\|x_t - x\|^2\\
&= -\frac{\alpha}{2t(1+\alpha t)}\|x_t - x\|^2 - \frac{1}{2t^2(1+\alpha t)}\|x_t - x\|^2\\
&= -\frac{\alpha}{2t(1+\alpha t)}\|x_t - x\|^2 - \frac{1}{2(1+\alpha t)}\|\nabla f(x_t)\|^2 && \text{(1st-order optimality)}\\
&\leqslant -\frac{\alpha}{2t(1+\alpha t)}\|x_t - x\|^2 - \frac{\alpha}{1+\alpha t}(f(x_t) - f_\star) && \text{(PŁ inequality)}\\
&= -\frac{\alpha}{1+\alpha t}\left(f(x_t) + \frac{1}{2t}\|x_t - x\|^2 - f_\star\right) = -\frac{\alpha}{1+\alpha t}(f_{t,x}(x_t) - f_\star)
\end{aligned}$$

Therefore, we have

$$\partial_t(Q_t f(x) - f_\star) \leqslant -\frac{\alpha}{1+\alpha t}(Q_t f(x) - f_\star) \tag{7.4}$$

Here we use a variant of Grönwall's inequality, with A_t can be dependent on t.

Theorem 7.1.13: Grönwall's Inequality II

Suppose that $u : [0,T] \to \mathbb{R}$ is a continuously differentiable curve that satisfies the differential inequality

$$\dot{u}(t) \leqslant A(t)u(t), \quad t \in [0,T]$$

Then, it holds that

$$u(t) \leqslant u(0)\exp\left(\int_0^t A(s)\,\mathrm{d}s\right), \quad t \in [0,T]$$

Proof. Suppose $u'(t) \leqslant A(t)u(t)$ for $t \in [0,T]$, then define the function

$$v(t) = \exp\left(\int_0^t A(s)\,\mathrm{d}s\right)$$

Note that it satisfies $v'(t) = A(t)v(t)$ with $v(0) = 1$ and $v(t) > 0$ for all t. We have

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{u(t)}{v(t)} = \frac{u'(t)v(t) - u(t)v'(t)}{v(t)^2} = \frac{u'(t)v(t) - u(t)A(t)v(t)}{v(t)^2} = \frac{u'(t) - A(t)u(t)}{v(t)} \leqslant 0$$

by assumption. Thus, $u(t)/v(t)$ is decreasing with t. This shows that

$$\frac{u(t)}{v(t)} \leqslant \frac{u(0)}{v(0)} \quad\implies\quad u(t) \leqslant u(0)\exp\left(\int_0^t A(s)\,\mathrm{d}s\right)$$

□

Using this on our equation 7.4, where $A(t) = -\frac{\alpha}{1+\alpha t}$, we have

$$\begin{aligned} Q_t f(x) - f_\star &\leqslant (Q_0 f(x) - f_\star) \exp\left(-\int_0^t \frac{\alpha}{1+\alpha s}\, \mathrm{d}s \right) \\ &= (Q_0 f(x) - f_\star) \exp\left(-\left[\log(1+\alpha s)\right]_{s=0}^t \right) = (Q_0 f(x) - f_\star) \frac{1}{1+\alpha t} \end{aligned}$$

To figure out $Q_0 f(t)$, note that when $t = 0$, $f_{t,x}(z) = \infty$ for all $z \neq x$. Therefore, $Q_0 f(x) = \inf f_{t,x}(z) = f_{t,z}(x) = f(x)$. Therefore,

$$Q_t f(x) - f_\star \leqslant \frac{1}{1+\alpha t}(f(x) - f_\star)$$

i.e.,

$$f(x_t) + \frac{1}{2t}\|x_t - x\|^2 - f_\star \leqslant \frac{1}{1+\alpha t}(f(x) - f_\star)$$

This means that

$$\begin{aligned} \frac{1}{1+\alpha t}(f(x) - f_\star) &\geqslant f(x_t) + \frac{1}{2t}\| -t\nabla f(x_t)\|^2 - f_\star && \text{(1st order optimality)} \\ &= f(x_t) + \frac{t}{2}\|\nabla f(x_t)\|^2 - f_\star \\ &\geqslant f(x_t) - f_\star + \alpha t(f(x_t) - f_\star) \\ &= (1+\alpha t)(f(x_t) - f_\star) \end{aligned}$$

which indicates the final result:

$$f(x_t) - f_\star \leqslant \frac{f(x) - f_\star}{(1+\alpha t)^2}$$

7.2 Mirror Methods

Until now, we have identified points x and gradient $\nabla f(x)$ as part of the same space $\mathbb{R}^d$. However, this is just because the self-dual nature of Euclidean norm. Suppose now, that $(\mathcal{X}, |||\cdot|||)$ is a general finite-dimensional normed vector space and $f : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$. The dual space is $(\mathcal{X}^\star, |||\cdot|||_\star)$, where $\mathcal{X}^\star$ is the space of linear functionals $\ell : \mathcal{X} \to \mathbb{R}$, equipped with the dual norm $|||\ell|||_\star := \sup\{|\ell(x)| : |||x||| \leqslant 1\}$. The derivative of f at x is defined to be the linearization at x: if there exists an element $\ell \in \mathcal{X}^\star$ such that

$$|f(x+v) - f(x) - \ell(v)| = o(|||v|||) \quad \text{as } v \to 0$$

we say that f is *differentiable* at x, and we write $Df(x)$ for the functional ℓ. Note that in this formalism, the derivative $Df(x)$ is an element of the dual space.

Remark: Above, we wrote $Df(x)$ instead of $\nabla f(x)$ to emphasize that in this context, we should no longer think of $Df(x)$ as belonging to the orignial space $\mathcal{X}$. However, when $\mathcal{X} = \mathbb{R}^d$, it is still convenient to identify $Df(x)$ as a vector in $\mathbb{R}$, and we therefore continue to use the notation $\nabla f(x)$. This is fine as long as we remember:

- It does not make sense to add a point $x \in \mathcal{X}$ to a gradient $\nabla f(x) \in \mathcal{X}^\star$.
- The size of $\nabla f(x)$ should be measured in the dual norm $|||\cdot|||_\star$.

We can not do, e.g. 'expand the square' as we did in previous proofs using non-Euclidean norms. Thus, to develop analogues of previous algorithms in different norms directly is not preferable. Instead, we use Fenchel-Legendre duality:

Throughout this section, $\phi : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is a convex function of Legendre type. We refer to it as the mirror map.

The idea is to use the auxiliary function ϕ to map the iterate x_n into the dual space via $x_n^\star = \nabla\phi(x_n)$. Now we are in the dual space, it makes sense to take a gradient step: $x_{n+1}^\star = x_n^\star - h\nabla f(x_n)$. Then, we use $\nabla\phi^\star$ to return: $x_{n+1} = \nabla\phi^\star(x_{n+1})$. The goal of this section is to formalize this idea.

7.2.1 Bregman Divergences

Definition 7.2.1: Bregman Divergence

Given a function $\phi : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ of Legendre type over $\mathcal{C}_\phi$, the corresponding **Bregman divergence** associated with ϕ is the map $D_\phi : \mathbb{R}^d \times \mathcal{C}_\phi \to \mathbb{R} \cup \{\infty\}$ defined by

$$D_\phi(x, y) := \phi(x) - \phi(y) - \langle \nabla\phi(y), x - y\rangle$$

In words, $D_\phi(\cdot, y)$ is defined by subtracting from ϕ its linearization at y. We can observe the following properties:

- $D_\phi \geqslant 0$: from convexity of ϕ.
- D_ϕ is convex w.r.t. its first argument.
- If ϕ is twice continuously differentiable, then

$$D_\phi(x, y) \sim \frac{1}{2}\langle x - y, \nabla^2\phi(y)(x - y)\rangle, \quad \text{as } x \to y$$

Therefore, D_ϕ behaves as a squared distance between x and y, but is not techinically a distance.

Example 1: $\phi(x) = \frac{1}{2}\|x\|^2$. Then, $\nabla\phi$ is the identity mapping and

$$D_\phi(x, y) = \frac{1}{2}\|x\|^2 - \frac{1}{2}\|y\|^2 - \langle y, x - y\rangle = \frac{1}{2}\|x - y\|^2$$

So, our study of mirror methods contain the preceding Euclidean methods.

Example 2: $\phi(x) = \sum_{i=1}^d \{x[i] \log x[i] - x[i]\}$ for $x \in \mathbb{R}_+^d$. Then, $\nabla\phi(x) = \log x$, where log is applied coordinatewise. The associated Bregman divergence is the Kullback-Leibler divergence

$$D_\phi(x, y) = \sum_{i=1}^d \left\{ x[i] \log \frac{x[i]}{y[i]} - x[i] + y[i] \right\}$$

Example 3: $\phi(X) = \operatorname{tr}(X \log X - X)$ for $X \succ 0$, called the von Neumann entropy. The associated Bregman divergence is the quantum relative entropy

$$D_\phi(X, Y) = \operatorname{tr}(X(\log X - \log Y) - X + Y)$$

Some properties of Bregman divergence is proved below.

Proposition 7.2.2: Properties of Bregman Divergence

1. For all $x, x' \in \mathcal{C}_\phi$, we have $D_\phi(x, x') = D_{\phi^\star}(\nabla\phi(x'), \nabla\phi(x))$.

2. Let $\mathcal{C} \subseteq \mathcal{C}_\phi$ be a closed convex set and let $\Pi_{\mathcal{C}}^\phi : \mathcal{C}_\phi \to \mathcal{C}$ denote the Bregman projection operator:

$$\Pi_{\mathcal{C}}^\phi(x) := \underset{\mathcal{C}\cap\mathcal{C}_\phi}{\operatorname{argmin}} D_\phi(\cdot, x)$$

   Then, $\langle \nabla\phi(\Pi_{\mathcal{C}}^\phi(x)) - \nabla\phi(x), \Pi_{\mathcal{C}}^\phi(x) - z\rangle \leqslant 0$ for all $z \in \mathcal{C}$, and the Pythagorean inequality holds:

$$D_\phi(z, x) \geqslant D_\phi(z, \Pi_{\mathcal{C}}^\phi(x)) + D_\phi(\Pi_{\mathcal{C}}^\phi(x), x)$$

3. Let X be a random variable with $\mathbb{E}|\phi(X)| < \infty$. For any $v \in \mathcal{C}_\phi$, we have

$$\mathbb{E}\left[D_\phi(X, v)\right] - \mathbb{E}\left[D_\phi(X, \mathbb{E}[X])\right] = D_\phi(\mathbb{E}[X], v)$$

   Consequently, the Bregman barycenter coincides with the usual mean:

$$\underset{v\in\mathcal{C}_\phi}{\operatorname{argmin}} \mathbb{E}\left[D_\phi(X, v)\right] = \mathbb{E}[X]$$

Proof. 1. By Fenchel-Young inequality, $\phi(x) + \phi^\star(y) = \langle x, y\rangle$, where $y = \nabla\phi(x)$ and in particular, $x = \nabla\phi^\star(y)$. If we denote $y' = \nabla\phi(x')$, and $x' = \nabla\phi^\star(y')$, we have

$$\begin{aligned}
D_{\phi^\star}(\nabla\phi(x'), \nabla\phi(x)) &= D_{\phi^\star}(y', y) \\
&= \phi^\star(y') - \phi^\star(y) - \langle \nabla\phi^\star(y), y' - y\rangle \\
&= \langle x', y'\rangle - \phi(x') - \langle x, y\rangle + \phi(x) - \langle x, y' - y\rangle \\
&= \phi(x) - \phi(x') - \langle y', x - x'\rangle \\
&= \phi(x) - \phi(x') - \langle \nabla\phi(x'), x - x'\rangle = D_\phi(x, x')
\end{aligned}$$

2. As in the proof of Lemma 1.3.5, the necessary condition for optimality of $\Pi_{\mathcal{C}}^\phi(x)$ is that

$$\begin{aligned}
\left\langle \nabla_1 D_\phi\left(\Pi_{\mathcal{C}}^\phi(x), x\right), z - \Pi_{\mathcal{C}}^\phi(x)\right\rangle &= \left\langle \nabla_u\left(\phi(u) - \phi(x) - \langle\nabla\phi(x), u - x\rangle\right)|_{u=\Pi_{\mathcal{C}}^\phi(x)}, z - \Pi_{\mathcal{C}}^\phi(x)\right\rangle \\
&= \left\langle \nabla\phi\left(\Pi_{\mathcal{C}}^\phi(x)\right) - \nabla\phi(x), z - \Pi_{\mathcal{C}}^\phi(x)\right\rangle \geqslant 0, \quad \forall\, z \in \mathcal{C}
\end{aligned}$$

where ∇_1 denotes gradient w.r.t. the first component. Take negative signs on both sides, we have exactly $\langle \nabla\phi(\Pi_{\mathcal{C}}^\phi(x)) - \nabla\phi(x), \Pi_{\mathcal{C}}^\phi(x) - z\rangle \leqslant 0$ for all $z \in \mathcal{C}$. Now we start to prove the Pythagorean inequality. First,

$$D_\phi(z, x) = \phi(z) - \phi(x) - \langle \nabla\phi(x), z - x\rangle$$

$$D_\phi\left(z, \Pi_{\mathcal{C}}^\phi(x)\right) = \phi(z) - \phi\left(\Pi_{\mathcal{C}}^\phi(x)\right) - \left\langle \nabla\phi\left(\Pi_{\mathcal{C}}^\phi(x)\right), z - \Pi_{\mathcal{C}}^\phi(x)\right\rangle$$

Therefore, we can write

$$
\begin{aligned}
D_\phi(z,x) - D_\phi\left(z, \Pi_{\mathcal{C}}^{\phi}(x)\right) &= \phi(z) - \phi(x) - \langle \nabla\phi(x), z - x\rangle - \phi(z) + \phi\left(\Pi_{\mathcal{C}}^{\phi}(x)\right) + \left\langle \nabla\phi\left(\Pi_{\mathcal{C}}^{\phi}(x)\right), z - \Pi_{\mathcal{C}}^{\phi}(x)\right\rangle \\
&= \phi\left(\Pi_{\mathcal{C}}^{\phi}(x)\right) - \phi(x) - \left\langle \nabla\phi\left(\Pi_{\mathcal{C}}^{\phi}(x)\right), \Pi_{\mathcal{C}}^{\phi}(x) - z\right\rangle - \langle \nabla\phi(x), z - x\rangle
\end{aligned}
$$

Note that we can write

$$
D_\phi\left(\Pi_{\mathcal{C}}^{\phi}(x), x\right) = \phi\left(\Pi_{\mathcal{C}}^{\phi}(x)\right) - \phi(x) - \left\langle \nabla\phi(x), \Pi_{\mathcal{C}}^{\phi}(x) - x\right\rangle
$$

Therefore, we have

$$
\begin{aligned}
D_\phi(z,x) - D_\phi\left(z, \Pi_{\mathcal{C}}^{\phi}(x)\right) - D_\phi\left(\Pi_{\mathcal{C}}^{\phi}(x), x\right) &= \left\langle \nabla\phi(x), \Pi_{\mathcal{C}}^{\phi}(x) - x\right\rangle - \left\langle \nabla\phi\left(\Pi_{\mathcal{C}}^{\phi}(x)\right), \Pi_{\mathcal{C}}^{\phi}(x) - z\right\rangle - \langle \nabla\phi(x), z - x\rangle \\
&= \left\langle \nabla\phi(x) - \nabla\phi\left(\Pi_{\mathcal{C}}^{\phi}(x)\right), \Pi_{\mathcal{C}}^{\phi}(x) - z\right\rangle \geqslant 0
\end{aligned}
$$

where we use the inequality we proved above. Therefore,

$$
D_\phi(z,x) \geqslant D_\phi\left(z, \Pi_{\mathcal{C}}^{\phi}(x)\right) + D_\phi\left(\Pi_{\mathcal{C}}^{\phi}(x), x\right)
$$

3. We have

$$
\mathbb{E}[D_\phi(X,v)] = \mathbb{E}\left[\phi(X) - \phi(v) - \langle \nabla\phi(v), X - v\rangle\right] = \mathbb{E}[\phi(X)] - \phi(v) - \langle \nabla\phi(v), \mathbb{E}[X] - v\rangle
$$

Similarly,

$$
\mathbb{E}[D_\phi(X,\mathbb{E}[X])] = \mathbb{E}[\phi(X)] - \phi\left(\mathbb{E}(X)\right) - \langle \nabla\phi\left(\mathbb{E}[X]\right), \mathbb{E}[X] - \mathbb{E}[X]\rangle = \mathbb{E}[\phi(X)] - \phi\left(\mathbb{E}(X)\right)
$$

Substract these two, we have

$$
\mathbb{E}[D_\phi(X,v)] - \mathbb{E}[D_\phi(X,\mathbb{E}[X])] = \phi(\mathbb{E}[X]) - \phi(v) - \langle \nabla\phi(v), \mathbb{E}[X] - v\rangle = D_\phi(\mathbb{E}[X], v)
$$

With this equation, since RHS is nonnegative, where it equals to zero if and only if $v = \mathbb{E}[X]$. Since $\mathbb{E}[D_\phi(X,\mathbb{E}[X])]$ is constant, we have

$$
\underset{v\in\mathcal{C}_\phi}{\operatorname{argmin}} \mathbb{E}[D_\phi(X,v)] = \mathbb{E}[X]
$$

which completes the proof. □

Proposition 7.2.3: Convexity-Smoothness Duality in Different Norms

If ϕ is α-convex relative to a norm $|||\cdot|||$, then $\phi^\star$ is α^{-1}-smooth relative to the dual norm $|||\cdot|||_\star$.

Proof. Since ϕ is Legendre type, its gradient $\nabla\phi$ is one-to-one and onto the interior of the domain of $\phi^\star$. In particular, for any s and t, if we set $x = \nabla\phi^\star(s)$ and $y = \nabla\phi^\star(t)$, it follows that $s = \nabla\phi(x)$ and $t = \nabla\phi(y)$. Using the α-convexity, we have

$$
\phi(y) \geqslant \phi(x) + \langle \nabla\phi(x), y - x\rangle + \frac{\alpha}{2}|||y - x|||^2 = \phi(x) + \langle s, y - x\rangle + \frac{\alpha}{2}|||y - x|||^2
$$

Similarly, exchanging the role of x and y, we have

$$\phi(x) \geqslant \phi(y) - \langle \nabla\phi(y), y - x\rangle + \frac{\alpha}{2}|\!|\!|y - x|\!|\!|^2 = \phi(y) - \langle t, y - x\rangle + \frac{\alpha}{2}|\!|\!|y - x|\!|\!|^2$$

Adding these two, we obtain

$$\alpha|\!|\!|y - x|\!|\!|^2 \leqslant \langle t - s, y - x\rangle$$

By Hölder's inequality,

$$\alpha|\!|\!|y - x|\!|\!|^2 \leqslant \langle t - s, y - x\rangle \leqslant |\!|\!|t - s|\!|\!|_\star |\!|\!|y - x|\!|\!| \quad \Longrightarrow \quad \alpha|\!|\!|y - x|\!|\!| \leqslant |\!|\!|t - s|\!|\!|_\star$$

which is equivalent to $|\!|\!|\nabla\phi^\star(t) - \nabla\phi^\star(s)|\!|\!| \leqslant (1/\alpha)|\!|\!|t - s|\!|\!|_\star$, and this shows that $\phi^\star$ is α^{-1}-smooth. □

7.2.2 Relative Convexity and Smoothness

Definition 7.2.4: Relative Convex/Smooth

Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be differentiable on $\operatorname{int}\operatorname{dom} f \subseteq \mathcal{C}_\phi$.

- f is **α-convex relative to ϕ** if $D_f \geqslant \alpha D_\phi$.
- f is **β-smooth relative to ϕ** if $D_f \leqslant \beta D_\phi$.

These are equivalent reformulations of these definitions. See Lu[9] for details.

Proposition 7.2.5: Properties of Relative Convexity

For any $\alpha \geqslant 0$, TFAE:

- f is α-convex relative to ϕ.
- $f - \alpha\phi$ is convex.
- $\langle \nabla f(y) - \nabla f(x), y - x\rangle \geqslant \alpha\langle \nabla\phi(y) - \nabla\phi(x), y - x\rangle$ for all $x, y \in \operatorname{int}\operatorname{dom} f$.

If f is twice continuously differentiable on $\operatorname{int}\operatorname{dom} f$, then

- $\nabla^2 f \succcurlyeq \alpha\nabla^2\phi$ on $\operatorname{int}\operatorname{dom} f$.

Proposition 7.2.6: Properties of Relative Smoothness

For any $\beta \geqslant 0$, TFAE:

- f is β-smooth relative to ϕ.
- $\beta\phi - f$ is convex.
- $\langle \nabla f(y) - \nabla f(x), y - x\rangle \leqslant \beta\langle \nabla\phi(y) - \nabla\phi(x), y - x\rangle$ for all $x, y \in \operatorname{int}\operatorname{dom} f$.

If f is twice continuously differentiable on $\operatorname{int}\operatorname{dom} f$, then

- $\nabla^2 f \preccurlyeq \beta\nabla^2\phi$ on $\operatorname{int}\operatorname{dom} f$.

For the case of $\phi = \frac{1}{2}\|\cdot\|^2$, we recover the usual notions of convexity and smoothness. These relative definitions satisfy similar properties as convexity/smoothness, e.g., if f_1, f_2 are α_1- and α_2-convex relative to ϕ and $\lambda_1, \lambda_2?0$, then $\lambda_1 f_1 + \lambda_2 f_2$ is $(\lambda_1\alpha_1 + \lambda_2\alpha_2)$-convex relative to ϕ. Also, we have a growth bound.

Lemma 7.2.7: Relative Growth

Suppose f is α-convex relative to ϕ for some $\alpha > 0$, and that f is minimized at an interior point $x_\star$ of its domain. Then, for all $x \in \mathbb{R}^d$,

$$f(x) - f_\star \geqslant \alpha D_\phi(x, x_\star)$$

Proof. The LHS is $D_f(x, x_\star) = f(x) - f_\star - \langle \nabla f(x_\star), x - x_\star\rangle = f(x) - f_\star$. □

7.2.3 Algorithm and Convergence Analysis

We first consider the continuous-time picture. Since we add the gradient of f in the dual space, the dynamics we consider evolve according to

$$\partial_t \nabla\phi(x_t) = -\nabla f(x_t) \tag{7.5}$$

By chain rule, we have equivalent formulation in the primal space:

$$\dot{x}_t = -(\nabla^2\phi(x_t))^{-1}\nabla f(x_t) \tag{7.6}$$

This can be interpreted as a preconditioned gradient flow. Despite equivance of these two equations, the discretization of 7.6 is usually called natural gradient descent and it is related to the subject of information geometry[3]. However, **mirror descent** is obtained from discretization of 7.5. The key distinguishing feature of mirror methods from preconditioned gradient flow is *the existence of the* ***global*** *process measure given by the Bregman divergence D_ϕ*. In contrast, precondition method are purely local in nature.

We consider the **mirror proximal gradient descent** method

$$x_{n+1} = \operatorname*{argmin}_{x\in\mathbb{R}^d}\left\{f(x_n) + \langle\nabla f(x_n), x - x_n\rangle + g(x) + \frac{1}{h}D_\phi(x, x_n)\right\} \tag{MPGD}$$

This iteration contains the following algorithms:

- When $g = 0$, since $\nabla_1 D_\phi(x, x_n) = \nabla\phi(x) - \nabla\phi(x_n)$, where ∇_1 denotes gradient w.r.t. the first component, the first-order optimality condition reads $h\nabla f(x_n) + \nabla\phi(x_{n+1}) - \nabla\phi(x_n) = 0$, which is

  $$\nabla\phi(x_{n+1}) = \nabla\phi(x_n) - h\nabla f(x_n)$$

  This is the **mirror descent**.

- When $f = 0$, we have

  $$x_{n+1} = \operatorname*{argmin}_{x\in\mathbb{R}^d}\left\{g(x) + \frac{1}{h}D_\phi(x, x_n)\right\} =: \mathrm{prox}^\phi_{hg}(x_n)$$

  which is the **mirror proximal point method**.

- When $g = \chi_{\mathcal{C}}$, where $\mathcal{C} \subseteq \mathcal{C}_\phi$ is a closed convex set, we have

$$\begin{aligned} x_{n+1} &= \operatorname*{argmin}_{x \in \mathcal{C}} \left\{ \langle \nabla f(x_n), x - x_n \rangle + \frac{1}{h} \left(\phi(x) - \langle \nabla \phi(x_n), x - x_n \rangle \right) \right\} \\ &= \operatorname*{argmin}_{x \in \mathcal{C}} \left\{ \phi(x) - \langle \nabla \phi(x_n) - h \nabla f(x_n), x - x_n \rangle \right\} \\ &= \Pi_{\mathcal{C}}^{\phi} (\nabla \phi^\star (\nabla \phi(x_n) - h \nabla f(x_n))) \end{aligned}$$

which is the mirror analogue of projected gradient descent. The last equality holds since, if we denote $y = \nabla \phi^\star(\nabla \phi(x_n) - h\nabla f(x_n))$, then $\nabla \phi(y) = \nabla \phi(x_n) - h \nabla f(x_n)$. Then the Bregman divergence $D_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla \phi(y), x - y \rangle = \phi(x) - \langle \nabla \phi(y), x - x_n \rangle - \phi(y) + \langle \nabla \phi(y), y - x_n \rangle$, which has only a constant difference with $\phi(x) - \langle \nabla \phi(x_n) - h \nabla f(x_n), x - x_n \rangle$.

Remark: More generally, We can write MPGD in the form

$$x_{n+1} = \operatorname{prox}_{hg}^{\phi} (\nabla \phi^\star (\nabla \phi(x_n) - h \nabla f(x_n)))$$

with similar argument as above.

Example: Consider the mirror map $\phi : x \mapsto -\sum_{i=1}^d \log x[i]$, defined over $\mathbb{R}_+^d$. We will compute the Bregman proximal operator for $\| \cdot \|_1$, which is $\operatorname{prox}_{h \| \cdot \|_1}^{\phi}$. Note that

$$\nabla \phi(x) = \left(-\frac{1}{x[1]}, -\frac{1}{x[2]}, \cdots, -\frac{1}{x[d]} \right), \quad \Longrightarrow \quad D_\phi(x, x_n) = \sum_{i=1}^d \left(\log \frac{x_n[i]}{x[i]} + \frac{x[i]}{x_n[i]} - 1 \right)$$

Therefore, the proximal operator minimizes $h \|x\|_1 + D_\phi(x, x_n)$ w.r.t. x. Take derivative, we have $h - 1/x[i] + 1/x_n[i] = 0$, so the proximal operator becomes

$$\operatorname{prox}_{h \| \cdot \|_1}^{\phi} (x_n) = \left(\frac{x_n[1]}{1 + h x_n[1]}, \cdots, \frac{x_n[d]}{1 + h x_n[d]} \right)$$

Theorem 7.2.8: Convergence of MPGD

Let f be α_f-convex and β_f-smooth, and let g be α_g-convex, all relative to ϕ. Let the step size h satisfy $h \leqslant 1/\beta_f$, let x^+ denote the next iteration of MPGD started from x, and let $y \in \mathbb{R}^d$. Then,

$$(1 + \alpha_g h) D_\phi(y, x^+) \leqslant (1 - \alpha_f h) D_\phi(y, x) - h(F(x^+) - F(y))$$

In particular, if we set $y = x_\star$ and iterate, it yields

$$F(x_N) - F_\star \leqslant \frac{\alpha_f + \alpha_g}{\lambda_h^{-N} - 1} D_\phi(x_\star, x_0)$$

where $\lambda_h := (1 - \alpha_f h)/(1 + \alpha_g h)$.

Proof. This proof is patterned upon the proof of Theorem 6.2.3. Let ψ_x denote the objective function in MPGD starting from x (rather than x_n), i.e.,

$$\psi_x(y) = f(x) + \langle \nabla f(x), y - x \rangle + g(y) + \frac{1}{h} D_\phi(y, x)$$

Then, $f(x)$ is constant, second term is linear, so ψ_x is $(\alpha_g + 1/h)$-convex relative to ϕ with minimizer at x^+, so by the growth inequality in Lemma 7.2.7:

$$\psi_x(y) \geqslant \psi_x(x^+) + \left(\alpha_g + \frac{1}{h}\right) D_\phi(y, x^+)$$

On one hand, by α_f-convexity,

$$\begin{aligned}\psi_x(y) &= f(x) + \langle \nabla f(x), y - x \rangle + g(y) + \frac{1}{h} D_\phi(y, x) \\ &= f(y) - D_f(y, x) + g(y) + \frac{1}{h} D_\phi(y, x) \leqslant F(y) + \left(\frac{1}{h} - \alpha_f\right) D_\phi(y, x)\end{aligned}$$

where the last step uses convexity $D_f(y, x) \geqslant \alpha_f D_\phi(y, x)$. On the other hand, by β_f-smoothness,

$$\begin{aligned}\psi_x(x^+) &= f(x) + \langle \nabla f(x), x^+ - x \rangle + g(x^+) + \frac{1}{h} D_\phi(x^+, x) \\ &= f(x^+) - D_f(x^+, x) + g(x^+) + \frac{1}{h} D_\phi(x^+, x) \geqslant F(x^+) + \left(\frac{1}{h} - \beta_f\right) D_\phi(x^+, x) \geqslant F(x^+)\end{aligned}$$

where the last step uses $h \leqslant 1/\beta_f$. Combine these three equations, we have

$$\begin{aligned}& F(y) + \left(\frac{1}{h} - \alpha_f\right) D_\phi(y, x) \geqslant F(x^+) + \left(\alpha_g + \frac{1}{h}\right) D_\phi(y, x^+) \\ \implies \quad & (1 + \alpha_g h) D_\phi(y, x^+) \leqslant (1 - \alpha_f h) D_\phi(y, x) - h(F(x^+) - F(y))\end{aligned}$$

which proves the one-step bound.

Note that by taking $y = x$, it yields the descent property

$$F(x^+) - F(x) \leqslant -\frac{1 + \alpha_g h}{2h} D_\phi(x, x^+) \leqslant 0$$

Instead, we set $y = x_\star$, and use discrete Grönwall to iterate, with $A = \frac{1 - \alpha_f h}{1 + \alpha_g h}$ and $B_n = -\frac{h}{1 + \alpha_g h}(F(x_{n+1}) - F(x_\star))$, we have

$$\begin{aligned}& 0 \leqslant D_\phi(x_\star, x_N) \leqslant \left(\frac{1 - \alpha_f h}{1 + \alpha_g h}\right)^N D_\phi(x_\star, x_0) - \sum_{n=1}^{N} \left(\frac{1 - \alpha_f h}{1 + \alpha_g h}\right)^{N-n} \frac{h}{1 + \alpha_g h}(F(x_n) - F_\star) \\ \implies \quad & \frac{h}{1 + \alpha_g h}(F(x_N) - F_\star) \sum_{n=1}^{N} \left(\frac{1 - \alpha_f h}{1 + \alpha_g h}\right)^{-n} \leqslant D_\phi(x_\star, x_0) \qquad \text{(Descent Property)} \\ \implies \quad & \frac{1 + \alpha_g h}{\alpha_f h + \alpha_g h} \left(\left(\frac{1 - \alpha_f h}{1 + \alpha_g h}\right)^{-N} - 1\right) \frac{h}{1 + \alpha_g h}(F(x_N) - F_\star) \leqslant D_\phi(x_\star, x_0) \\ \implies \quad & F(x_N) - F_\star \leqslant \frac{\alpha_f + \alpha_g}{\lambda_h^{-N} - 1} D_\phi(x_\star, x_0)\end{aligned}$$

where $\lambda_h = (1 - \alpha_f h)/(1 + \alpha_g h)$. This completes the proof. □

Although this result is the analogue of the smooth convergence rate for gradient descent, since $\nabla\phi$ necessarily blows up at the boundary $\partial\mathcal{C}_\phi$, so can ∇f. Therefore, this theorem actually covers examples in which f is not at all smooth in the usual sense.

Definition 7.2.9: Convexity/Smoothness Relative to a Norm

A function f is α-convex (resp. β-smooth) **relative to a norm** $|||\cdot|||$ if for all $x, y \in \operatorname{int}\operatorname{dom} f$,

$$D_f(x,y) \geqslant \frac{\alpha}{2}|||y-x|||^2 \quad \left(\text{resp. } D_f(x,y) \leqslant \frac{\beta}{2}|||y-x|||^2\right)$$

Suppose that ϕ is strongly convex relative to a norm $|||\cdot|||$. Then, to check that f is smooth relative to ϕ, it suffices to check that f is smooth relative to $|||\cdot|||$, so the norm can act as a useful intermediary. Moreover, whereas the Bregman structure is crucial for carrying out the iterative analysis of MPGD, the norm structure is often convenient too, e.g., for the use of tools such as Cauchy–Schwarz. To illustrate this, we consider the non-smooth case. Here, we assume that f is Lipschitz w.r.t. $|||\cdot|||$:

$$|f(x) - f(y)| \leqslant L|||x-y|||, \quad \forall\, x, y \in \mathcal{C}_\phi$$

We again consider MPGD, except that $\nabla f(x_n)$ should be interpreted as a subgradient. We leave the notation unchanged because it should not cause confusion. The Lipschitz condition is then equivalent to the subgradient bound

$$|||\nabla f(x)|||_\star \leqslant L \quad \forall\, x \in \mathcal{C}_\phi$$

Theorem 7.2.10: Convergence of MPGD, non-smooth case

Let f and g be convex, and let f be L-Lipschitz w.r.t. a norm $|||\cdot|||$. Let ϕ be α_ϕ-convex relative to $|||\cdot|||$. Then, for MPGD, it holds that

$$F\left(\frac{1}{N}\sum_{n=1}^{N} x_n\right) - F_\star \leqslant \frac{1}{N}\sum_{n=1}^{N}(F(x_n) - F_\star) \leqslant \frac{D_\phi(x_\star, x_0)}{Nh} + \frac{2L^2 h}{\alpha_\phi}$$

In particular, if $R_\phi^2 \geqslant D_\phi(x_\star, x_0)$ and we choose step size $h^2 = \alpha_\phi R_\phi^2/(2L^2 N)$, we have

$$F\left(\frac{1}{N}\sum_{n=1}^{N} x_n\right) - F_\star \leqslant LR_\phi\sqrt{\frac{8}{\alpha_\phi N}}$$

Proof. Following the proof of Theorem 7.2.8, we still have

$$\psi_x(x^+) + \frac{1}{h}D_\phi(x_\star, x^+) \leqslant \psi_x(x_\star) \leqslant F(x_\star) + \frac{1}{h}D_\phi(x_\star, x)$$

For the lower bound, we no longer have smoothness. Instead, by Cauchy-Schwarz,

$$D_f(x^+, x) = f(x^+) - f(x) - \langle \nabla f(x), x^+ - x\rangle \leqslant L|||x^+ - x||| + |||\nabla f(x)|||_\star |||x^+ - x||| \leqslant 2L|||x^+ - x|||$$

Thus, by α_ϕ-convexity of ϕ,

$$\begin{aligned}
\psi_x(x^+) &= f(x) + \langle \nabla f(x), x^+ - x \rangle + g(x^+) + \frac{1}{h} D_\phi(x^+, x) \\
&= F(x^+) - D_f(x^+, x) + \frac{1}{h} D_\phi(x^+, x) \\
&\geqslant F(x^+) - 2L|||x^+ - x||| + \frac{\alpha_\phi}{2h} |||x^+ - x|||^2 \\
&\geqslant F(x^+) - \frac{2L^2 h}{\alpha_\phi}
\end{aligned}$$

where the last inequality is achieved by minimizing w.r.t. $|||x^+ - x|||$, which leads to $|||x^+ - x||| = \frac{2hL}{\alpha_\phi}$. This leads to the one-step bound

$$D_\phi(x_\star, x^+) \leqslant D_\phi(x_\star, x) - h(F(x^+) - F_\star) + \frac{2L^2 h^2}{\alpha_\phi}$$

Iterate this ineuqality using discrete Grönwall, with $A = 1$ and $B_n = -h(F(x_{n+1}) - F_\star) + 2L^2 h^2/\alpha_\phi$, we have

$$\begin{aligned}
0 \leqslant D_\phi(x_\star, x_N) &\leqslant D_\phi(x_\star, x_0) + \sum_{n=1}^{N} \left(\frac{2L^2 h^2}{a_\phi} - h(F(x_n) - F_\star) \right) \\
\implies \quad \frac{1}{N} \sum_{n=1}^{N} (F(x_n) - F_\star) &\leqslant \frac{D_\phi(x_\star, x_0)}{Nh} + \frac{2L^2 h}{a_\phi}
\end{aligned}$$

Since F is convex, by Jensen's inequality, we have the first inequality holds. Take $R_\phi^2 \geqslant D_\phi(x_\star, x_0)$ and $h^2 = a_\phi R_\phi^2/(2L^2 N)$, we have

$$\frac{D_\phi(x_\star, x_0)}{Nh} + \frac{2L^2 h}{\alpha_\phi} \leqslant LR_\phi \sqrt{\frac{2}{a_\phi N}} + \frac{2L^2}{a_\phi} \frac{R_\phi}{L} \sqrt{\frac{a_\phi}{2N}} = LR_\phi \sqrt{\frac{8}{a_\phi N}}$$

which is the second inequality. □

Example: Why we would bother optimize w.r.t. non-Euclidean norms? Suppose that $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is convex and we wish to minimize it over the simplex $\Delta_d := \{x \in \mathbb{R}_+^d : \sum_{i=1}^d x[i] = 1\}$. If f is Lipschitz, then we can apply projected subgradient descent and obtain an ϵ-approximate solution in $O(L^2 R^2/\epsilon^2)$ iterations. Here, L is the Lipschitz constant, and $R \leqslant 2$ is the radius.

For example, if we have d actions and the loss of the ith action is $\ell[i]$, where the losses are bounded: $|\ell[i]| \leqslant 1$. If we choose an action randomly according to a probability distribution $x \in \Delta_d$, the expected loss is $\langle \ell, x \rangle$. We then want to minimize $f(x) := \langle \ell, x \rangle$ over Δ_d. (Trivially, the solution is given by the distribution which puts all of its mass on $\operatorname{argmin}_{i \in [d]} \ell[i]$. This problem is simply meant to illustrate the pitfalls of Euclidean norm.) The Lipschitz constant of f is $\|\ell\|$, which could be as large as $\sqrt{d}$ in the worst case. The resulting complexity estimate of $O(d/\epsilon^2)$ is poor in high dimension.

Implicit in this discussion, however, is that we are measuring the Lipschitz constant and the radius with respect to the usual Euclidean norm. In this setting, however, it may make more sense to use the ℓ_1 norm for radius, in which case the Lipschitz constant is $\|\ell\|_\infty \leqslant 1$.

In this example, we use the entropic mirror map $\phi(x) = \sum_{i=1}^{d}\{x[i]\log x[i] - x[i]\}$.

Pinsker's inequality:
ϕ is 1-convex relative to the ℓ_1-norm $\|\cdot\|_1$ over the probability simplex Δ_d.

To minimize $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ over Δ_d, we apply MPGD with $g = \chi_{\Delta_d}$. Then, $\nabla\phi(x) = \log x$, and $\nabla\phi^\star(x) = \exp(x)$, where log and exp are applied pointwise, so

$$\nabla\phi^\star(\nabla\phi(x_n) - h\nabla f(x_n)) = \exp\left(\log x_n - h\nabla f(x_n)\right) = x_n \odot \exp\left(-h\nabla f(x_n)\right)$$

where exp is applied pointwise, and $\odot$ is the Hadamard (pointwise) product.

$\Pi^{\phi}_{\Delta_d}(x) = x/\|x\|_1$ simply normalizes the vector.

Proof. Note that

$$\Pi^{\phi}_{\Delta_d}(z) = \underset{x\in\Delta_d}{\operatorname{argmin}} D_\phi(x,z) = \underset{x\in\Delta_d}{\operatorname{argmin}}\left\{\sum_{i=1}^{d} x[i]\log\frac{x[i]}{z[i]}\right\}$$

after deleting all the constants. Introducing the Lagrangian multiplier

$$\mathcal{L}(x,\lambda) = \sum_{i=1}^{d} x[i]\log\frac{x[i]}{z[i]} - \lambda\left(\sum_{i=1}^{d} x[i] - 1\right)$$

Differentiate w.r.t. each $x[i]$:

$$\frac{\partial\mathcal{L}}{\partial x[i]} = \log\frac{x[i]}{z[i]} + 1 - \lambda = 0 \quad\Longrightarrow\quad x[i] = z[i]\exp\left(\lambda - 1\right)$$

Since $\sum_{i=1}^{d} x[i] - 1 = 0$, we have $\exp(\lambda-1)\sum_{i=1}^{d} z[i] = 1$, which shows that $\exp(\lambda - 1) = 1/\|z\|_1$, which leads to $x[i] = \frac{z[i]}{\|z\|_1}$. □

Hence, the algorithm reads

$$x_{n+1} = \Pi^{\phi}_{\cdot_\ulcorner}(\nabla\phi^\star(\nabla\phi(x_n) - h\nabla f(x_n))) = \frac{x_n \odot \exp\left(-h\nabla f(x_n)\right)}{\|x_n \odot \exp\left(-h\nabla f(x_n)\right)\|_1}$$

Consider initializing at the unifom distribution $x_0 = \mathbf{1}_d/d$. Then, for any $x_\star \in \Delta_d$,

$$D_\phi(x_\star, x_0) = \mathrm{KL}(x_\star\|x_0) = \log d - \sum_{i=1}^{d} x_0[i]\log\frac{1}{x_0[i]} \leqslant \log d$$

Consequently, we can take $R_\phi = \sqrt{\log d}$, and with Theorem 7.2.10, we have

$$f\left(\frac{1}{N}\sum_{i=1}^{N} x_n\right) - f_\star \leqslant L_1\sqrt{\frac{8\log d}{N}}$$

where L_1 is the Lipschitz constant of f in the ℓ_1 norm, which is smaller than 1 as mentioned before. This estimate is far better than the one described using the Euclidean norm, and we only pay an overhead which is logarithmic in the dimension.

7.3 Online Algorithm and Multiplicative Weights

Chapter 8

Alternating Minimization

In this section, we study the method of alternating minimization. The goal is to minimize a function f by decomposing the optimization variable x into D variables $x^1, \cdots, x^D$. In this decomposition, the individual variables do not have to be 1-dimensional, so we let $x^i \in \mathbb{R}^{d_i}$. The method is defined as

$$x_{n+1}^i := \underset{x^i \in \mathbb{R}^{d_i}}{\operatorname{argmin}} f(x_{n+1}^1, \cdots, x_{n+1}^{i-1}, x^i, x_n^{i+1}, \cdots, x_n^D)$$

That is, we iterate through the variables cyclically and minimize f over the ith variable x^i, holding the other variables fixed. The decomposition is chosen so that it is cheap to compute the minimizer over each individual variable.

Motivation: Low-Rank Matrix Recovery

Suppose that we want to recover an unknown matrix $X_\star \in \mathbb{R}^{p_1 \times p_2}$ which is observed through noisy observations $y_i \approx \langle A_i, X_\star \rangle$, where the matrices $A_i \in \mathbb{R}^{p_1 \times p_2}$ are known. If we further posit that $X_\star$ is low-rank, say of rank at most r, then we aim to solve

$$\underset{X \in \mathbb{R}^{p_1 \times p_2}}{\text{minimize}} \sum_{i=1}^{n} (y_i - \langle A_i, X \rangle)^2, \quad \text{subject to} \quad \operatorname{rank} X \leqslant r$$

The rank constraint is difficult to deal with, so we instead factorize the matrix as $X = UV^T$ where $U \in \mathbb{R}^{p_1 \times r}$ and $V \in \mathbb{R}^{p_2 \times r}$. This factorization is known as the *Burer-Monteiro factorization.* The problem becomes

$$\underset{U \in \mathbb{R}^{p_1 \times r}, V \in \mathbb{R}^{p_2 \times r}}{\text{minimize}} \sum_{i=1}^{n} (y_i - \langle A_i, UV^T \rangle)^2$$

This is a non-convex problem, but at least it is now amenable to gradient-based methods. Alternatively, we can apply alternating minimization. In words, we minimize over U while holding V fixed, and then minimize over V while holding U fixed, and so on. Each iteration corresponds to solving an unconstrained least-squares problem and admits a closed-form solution.

8.1 Special Case: Alternating Projections

We can use alternating minimization to find a point in the intersection of two closed convex sets $\mathcal{C}_1$ and $\mathcal{C}_2$. In this case, we take

$$f(x, y) = \chi_{\mathcal{C}_1}(x) + \chi_{\mathcal{C}_2}(y) + \|y - x\|^2$$

If there exists $x_\star \in \mathcal{C}_1 \cap \mathcal{C}_2$, then $(x_\star, x_\star)$ is a minimizer for f, and the alternating minimization algorithm reads:

$$x_{n+1} := \underset{x \in \mathbb{R}^d}{\operatorname{argmin}} f(x, y_n) = \Pi_{\mathcal{C}_1}(y_n)$$

$$y_{n+1} := \underset{y \in \mathbb{R}^d}{\operatorname{argmin}} f(x_{n+1}, y) = \Pi_{\mathcal{C}_2}(x_{n+1})$$

Thus, we alternate projecting onto $\mathcal{C}_1$ and onto $\mathcal{C}_2$. This method is quite useful when proejctions onto $\mathcal{C}_1$ and $\mathcal{C}_2$ individually are cheap, but the projection on the intersection is expensive. The method easily generalizes to the intersection of more than two convex sets.

We consider a generalization to **Alternating Bregman Projections**,

$$x_{n+1} := \Pi^{\phi}_{\mathcal{C}_1}(y_n), \quad y_{n+1} := \Pi^{\phi}_{\mathcal{C}_2}(x_{n+1}) \tag{ABP}$$

We assume that $\mathcal{C}_1 \cap \mathcal{C}_2 \neq \emptyset$ and that $\mathcal{C}_1, \mathcal{C}_2 \subseteq \mathcal{C}_\phi$.

Lemma 8.1.1: Monotonicity of Iterating Bregman Divergence

For any $x_\star \in \mathcal{C}_1 \cap \mathcal{C}_2$, the iterates of ABP satisfy

$$\sum_{n=1}^{\infty} \{D_\phi(x_n, y_{n-1}) + D_\phi(y_n, x_n)\} \leqslant D_\phi(x_\star, y_0)$$

Also, monotonicity holds

$$D_\phi(x_\star, y_0) \geqslant D_\phi(x_\star, x_1) \geqslant D_\phi(x_\star, y_1) \geqslant \cdots$$

Proof. By the Pythagorean inequality in Proposition 7.2.2,

$$D_\phi(x_\star, y_n) \geqslant D_\phi\left(x_\star, \Pi^{\phi}_{\mathcal{C}_1}(y_n)\right) + D_\phi\left(\Pi^{\phi}_{\mathcal{C}_1}(y_n), y_n\right) = D_\phi(x_\star, x_{n+1}) + D_\phi(x_{n+1}, y_n)$$

From another projection,

$$D_\phi(x_\star, x_{n+1}) \geqslant D_\phi\left(x_\star, \Pi^{\phi}_{\mathcal{C}_2}(x_{n+1})\right) + D_\phi\left(\Pi^{\phi}_{\mathcal{C}_2}(x_{n+1}), x_{n+1}\right) = D_\phi(x_\star, y_{n+1}) + D_\phi(y_{n+1}, x_{n+1})$$

Add these together, we have

$$D_\phi(x_\star, y_n) \geqslant D_\phi(x_\star, y_{n+1}) + D_\phi(y_{n+1}, x_{n+1}) + D_\phi(x_{n+1}, y_n)$$

Rearrange, we have

$$D_\phi(x_\star, y_n) - D_\phi(x_\star, y_{n+1}) \geqslant D_\phi(y_{n+1}, x_{n+1}) + D)_\phi(x_{n+1}, y_n)$$

Since $D_\phi \geqslant 0$, we sum these inequalities and from 0 to $N-1$, and take $N \to \infty$, we have

$$\begin{aligned} &\sum_{n=0}^{N-1} \{D_\phi(x_{n+1}, y_n) + D_\phi(y_{n+1}, x_{n+1})\} \leqslant \sum_{n=0}^{N-1} \{D_\phi(x_\star, y_n) - D_\phi(x_\star, y_{n+1})\} = D_\phi(x_\star, y_0) - D_\phi(x_\star, y_N) \\ \implies &\sum_{n=1}^{\infty} \{D_\phi(x_n, y_{n-1}) + D_\phi(y_n, x_n)\} \leqslant D_\phi(x_\star, y_0) \end{aligned}$$

which is exactly the first inequality in the Lemma. Moreover, from Pythagorean inequality itself, we can get

$$D_\phi(x_\star, y_n) \geqslant D_\phi(x_\star, x_{n+1}) + D_\phi(x_{n+1}, y_n) \geqslant D_\phi(x_\star, x_{n+1}), \quad D_\phi(x_\star, x_{n+1}) \geqslant D_\phi(x_\star, y_{n+1}) + D_\phi(y_{n+1}, x_{n+1}) \geqslant D_\phi(x_\star, y_{n+1})$$

which implies the monotonicity. □

We can use the preceding lemma to prove a convergence result for ABP. The following corollary relies on two additional technical assumptions for ϕ which must be checked, but note that they hold for the Euclidean case $\phi = \frac{\|\cdot\|^2}{2}$.

Corollary 8.1.2: Convergence of ABP

Assume that the following conditions hold:

1. For any $x \in \mathcal{C}_\phi$, the sublevel sets of $D_\phi(x, \cdot)$ are compact.

2. If $\{z_n\}_{n\in\mathbb{N}}$, $\{z'_n\}_{n\in\mathbb{N}} \subseteq \mathcal{C}_\phi$ are such that $D_\phi(z_n, z'_n) \to 0$, then $z_n - z'_n \to 0$.

Then, the iterates of ABP satisfy $x_n \to x_\star$ and $y_n \to y_\star$ for some $x_\star \in \mathcal{C}_1 \cap \mathcal{C}_2$.

Proof. The first assumption ensures that there is a convergent subsequence $\{x_{n_k}\}_{k\in\mathbb{N}}$ that converges to some $x_\star \in \mathcal{C}_\phi$. This is because by Lemma 8.1.1, for any fixed $x_\star \in \mathcal{C}_1 \cap \mathcal{C}_2$, we have $D_\phi(x_\star, x_n) \leqslant D_\phi(x_\star, y_0)$ for all n. In other words, every iterate x_n lies in the sublevel set $\{z \in \mathcal{C}_\phi : D_\phi(x_\star, z) \leqslant D_\phi(x_\star, y_0)\}$. Since the sublevel sets are compact, the sequence x_n has at least one convergent subsequence. Since $x_n \in \mathcal{C}_1$ for all n and $\mathcal{C}_1$ is closed, then $x_\star \in \mathcal{C}_1$. Moreover, by Lemma 8.1.1, $D_\phi(\Pi^\phi_{\mathcal{C}_2}(x_n), x_n) = D_\phi(y_n, x_n) \to 0$, so the second property shows that $\Pi^\phi_{\mathcal{C}_2}(x_n) - x_n \to 0$. Since $\mathcal{C}_2$ is closed, $x_\star \in \mathcal{C}_2$ as well.

To upgrade the subsequential convergence to full convergence, we observe that $D_\phi(x_\star, x_{n_k}) \to 0$, whence the monotonicity statement implies $D_\phi(x_\star, x_n) \to 0$ and $D_\phi(x_\star, y_n) \to 0$. By the second assumption, $x_n \to x_\star$ and $y_n \to x_\star$. □

Furthermore, Lemma 8.1.1 implies

$$\min_{n=1,2,\cdots,N} \{D_\phi(x_n, y_{n-1}) + D_\phi(y_n, x_n)\} \leqslant \frac{D_\phi(x_\star, y_0)}{N}$$

This does not, however, imply a rate of convergence for x_n to $x_\star$. For example, if $\mathcal{C}_1$ and $\mathcal{C}_2$ are two lines that meet each other at a very small angle, then the successive projections can lie very close to each other even though they are both very far from the common point of intersection.

8.2 Convergence Analysis

We use the shorthand x^S to denote the components in S, $x^S := \{x^i\}_{i\in S}$, where we abbreviate consecutive indices $\{i, \cdots, j\}$ as $i : j$. We perform an analysis in the smooth case. However, similarly to how gradient-based methods do not suffer from non-smoothness provided that one has access to a proximal oracle, it turns out that coordinate-based methods do not suffer from non-smoothness provided that the non-smooth part respects the coordinate decomposition. Hence, we consider the slightly more general problem of minimizing

$$F : \mathbb{R}^{d_1\times\cdots d_D} \to \mathbb{R}, \quad F(x^{1:D}) := f(x^{1:D}) + \sum_{i=1}^{D} g_i(x^i)$$

where f is convex and smooth, and each g_i is convex. For shorthand, we write $g := \bigoplus_{i=1}^{D} g_i$, that is, $g(x^{1:D}) := \sum_{i=1}^{D} g_i(x^i)$. The algorithm reads,

$$x_{n+1}^i \in \operatorname*{argmin}_{x^i\in\mathbb{R}^{d_i}} \{f(x_{n+1}^{1:i-1}, x^i, x_n^{i+1:D}) + g_i(x^i)\} \tag{AM}$$

Theorem 8.2.1: Convergence of AM

Let f be convex and β-smooth, and each g_i is convex. Then, alternating minimization achieves $F(x_N^{1:D}) - F_\star \leqslant \epsilon$ if

$$N \geqslant \left(\log_{1/2} \frac{F(x_0^{1:D}) - F_\star}{4\beta D^2 R^2}\right)_+ + \frac{8\beta D^2 R^2}{\epsilon}$$

where $R := \sup_{n\in\mathbb{N}} \|x_n^{1:D} - x_\star^{1:D}\|$.

Proof. By 2.4 applying, we have

$$f(x_n^{1:D}) \geqslant f(x_{n+1}^1, x_n^{2:D}) + \langle \nabla_1 f(x_{n+1}^1, x_n^{2:D}), x_n^1 - x_{n+1}^1\rangle + \frac{1}{2\beta}\|\nabla f(x_{n+1}^1, x_n^{2:D}) - \nabla f(x_n^{1:D})\|^2$$

where it is ∇_1 in the inner product since all other terms has $x_n^i - x_n^i$ and equals to 0. On the other hand, since $\nabla_1 f(x_{n+1}^1, x_n^{2:D}) \in -\partial g_1(x_{n+1}^1)$, which is the first-order subgradient optimality condition, we have

$$g_1(x_n^1) + \langle \nabla_1 f(x_{n+1}^1, x_n^{2:D}), x_n^1 - x_{n+1}^1\rangle \geqslant g_1(x_{n+1}^1)$$

by the definition of subgradient. □

Bibliography

[1] Altschuler, J. and Parrilo, P. (2023). Acceleration by stepsize hedging: Multi-step descent and the silver stepsize schedule. *Journal of the ACM*.

[2] Altschuler, J. M. and Parrilo, P. A. (2024). Acceleration by stepsize hedging: Silver stepsize schedule for smooth convex optimization. *Mathematical Programming*, pages 1–14.

[3] Amari, S.-i. and Nagaoka, H. (2000). *Methods of information geometry*, volume 191. American Mathematical Soc.

[4] Chen, Y., Chewi, S., Salim, A., and Wibisono, A. (2022). Improved analysis for a proximal algorithm for sampling. In *Conference on Learning Theory*, pages 2984–3014. PMLR.

[5] Karimi, H., Nutini, J., and Schmidt, M. (2016). Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*, pages 795–811. Springer.

[6] Lessard, L., Recht, B., and Packard, A. (2016). Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95.

[7] Liang, J., Mitra, S., and Wibisono, A. (2024). On independent samples along the langevin diffusion and the unadjusted langevin algorithm. *arXiv preprint arXiv:2402.17067*.

[8] Lojasiewicz, S. (1963). Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, 117:87–89.

[9] Lu, H., Freund, R. M., and Nesterov, Y. (2018). Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354.

[10] Nemirovskij, A. S. and Yudin, D. B. (1983). Problem complexity and method efficiency in optimization.

[11] Nesterov, Y. et al. (2018). *Lectures on convex optimization*, volume 137. Springer.

[12] Rockafellar, R. T. (1997). *Convex analysis*, volume 28. Princeton university press.