Fantuan's Academia

Fantuan's Statistics Notes series Vol.4

Advanced Optimization Techniques

Author: Jingxuan Xu
Lecturer: Prof. Sinho Chewi

Fantuan's Math Notes

February 1, 2025

Fantuan's Math Notes

Contents

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

This note is a 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.

Fantuan's Math Notes

Fantuan's Math Notes

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 $(f(x), \nabla f(x), \nabla^2 f(x))$.

Fanyuan's Math Notes

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**[1] 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 1.2, there exists a multi-index α^* such that $x_\star \in X_{a^\star}$. 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 1.1. □

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

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

Fan Pu's Math Notes

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, $\bar{f(\cdot)}$ 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.
(1.3) $\implies$ (1.4): Rearranging Equation 1.3, 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) $\implies$ (1.5): 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

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

(1.5) $\implies$ (1.3): 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 1.5, since $x + sv - (x + stv) = s(1-t)v$, we finally arrive

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

Fantuan's Math Notes

$$\begin{aligned}
&\leqslant -t\int_0^1 \frac{1}{s(1-t)}\alpha s^2(1-t)^2\|v\|^2\,\mathrm{d}s && \text{(By equation 1.5)}\\
&= -t\int_0^1 \alpha s(1-t)\|v\|^2\,\mathrm{d}s\\
&= -\frac{\alpha}{2}t(1-t)\|v\|^2
\end{aligned}$$

which is exactly the Equation 1.3.

Assume f is twice continuously differentiable.
(1.5) $\Longrightarrow$ (1.6): Let $y = x + \epsilon v$ in Equation 1.5, 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) $\Longrightarrow$ (1.5): 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 1.5. □

They have each own interpretations: for $\alpha = 0$, Equation 1.3 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 1.5 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.

Pertuan's Math Notes

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 $\gamma(t) = x_\star + t\epsilon v$, $t \in [0, 1]$,

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

which shows that $\int_0^1 \langle \nabla f(x_\star + \epsilon t v), v \rangle \, \mathrm{d}t \geqslant 0$. 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

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

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.

Fan Pan's Math Notes

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

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

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

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

If f is twice continuously differentiable, then

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

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:

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

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 remains true even without assuming twice differentiability. We will see this later in Section 2.2.

Fantuan's Math Notes

Fantuan's Math Notes

Chapter 2

Gradient Method

2.1 Continuous Case: Gradient Flow

2.2 Discrete Case: Gradient Descent

Fantuan's Math Notes

Fantuan's Math Notes

Bibliography

[1] Nesterov, Y. et al. (2018). *Lectures on convex optimization*, volume 137. Springer.

Fantuan's Math Notes