DSC 211 Introduction to Optimization Winter 2024 Instructor: Jun-Kun Wang
Scribe: Yi Gu, Qiyue Gao January 18, 2024
Editors/TAs: Merlin Chang, Marialena Sfyraki

Lecture 4: Reduction

1 Review

Theorem 1. *The μ-strong convexity implies the μ-gradient dominant condition.*

Remark. *The condition number $\kappa := \frac{L}{\mu} \geq 1$.*

Proof. Let f be a μ-strongly convex and L-smooth function. For simplicity, let us think of μ as the strong convexity constant, since from Theorem 1 we have that μ-strong convexity implies the μ-gradient dominant condition. Additionally, by the second-order characterization of strong convexity we have that

$$y^\top \nabla^2 f(x) y \geq \mu \|y\|^2, \; \forall x \in \mathbf{C}, y \in \mathbb{R}^n, \mu > 0.$$

By the second-order characterization of L-smoothness we have that

$$y^\top \nabla^2 f(x) y \leq L \|y\|^2, \; \forall x \in \mathbf{C}, y \in \mathbb{R}^n.$$

If we choose the Euclidean norm $\|\cdot\|_2$ and a normalized vector y, that is $y \in \mathbb{R}^n$ such that $\|y\|_2 = 1$, then the above inequalities imply that

$$\lambda_{min}(\nabla^2 f(x)) \geq \mu,$$
$$\lambda_{max}(\nabla^2 f(x)) \leq L,$$

where λ_{min} and λ_{max} are the min and max eigenvalue of $\nabla^2 f(x)$, respectively. This suggests the following inequality

$$0 < \mu \leq \lambda_{min} \leq \lambda_{max} \leq L,$$

which entails that

$$\kappa := \frac{L}{\mu} \geq 1.$$

□

Remark. $\exp(-x)$ *is smooth in a bounded region $x \in [-c, c], c < \infty$.*

Proof.

$$\nabla^2 \exp(-x) = \exp(-x) \in [\exp(-c), \exp(c)] > 0,$$

which means there exists $L > 0$ such that the second-order characterization of smoothness is satisfied. □

Remark. $\exp(-x)$ *is differentiable but not smooth for* $x \in \mathbf{R}$*.*

Remark. *Many optimization people call* $\exp(-\frac{\mu}{L}k)$ *linear rate (consider logarithm), and* $\frac{LD^2}{k}$ *correspondingly sub-linear rate[**?**].*

Theorem 2. *Assume* $f(\cdot)$ *is* μ*-gradient dominant and* L*-smooth, then gradient descent with* $\eta = \frac{1}{L}$ *satisfies*

$$f(x_{k+1}) - \min_{x \in \mathbb{R}^d} f(x) \leq \left(1 - \frac{\mu}{L}\right)^k \left(f(x_1) - \min_{x \in \mathbb{R}^d} f(x)\right).$$

Theorem 3. *Assume* $f(\cdot)$ *is convex and* L*-smooth on* $\mathbb{R}^d$*, then gradient descent with* $\eta = \frac{1}{L}$ *satisfies*

$$f(x_{k+1}) - \min_{x \in \mathbb{R}^d} f(x) \leq \frac{2LD^2}{k}$$

where $D := \max_k \|x_k - x_*\|_2 \leq \|x_1 - x_*\|_2$*,* $x_* := \arg\min f(x)$*.*

Remark. *We also have the below inequality regarding to the linear rate:*

$$\left(1 - \frac{\mu}{L}\right)^k \leq \exp\left(-\frac{\mu}{L}k\right). \tag{1}$$

2 Reduction

Our goal is to solve the general unconstrained optimization problem

$$\min_x f(x)$$

We will not modify the underlying algorithm, and we will consider the following scenarios:

1. Given an algorithm with strong guarantees for **smooth and strongly convex function**, how to make it work for **smooth convex** functions with strong guarantees?

2. Given an algorithm with strong guarantees for **smooth and strongly convex function**, how to make it work for **non-smooth and strongly convex** functions with strong guarantees?

3. Given an algorithm with strong guarantees for **smooth and strongly convex function**, how to make it work for convex L_0-Lipschitz fucntions that are **neither strongly convex nor smooth**?

We will study a technique called Reduction.

Lemma 1 (To be proved in **HW2**). *Suppose $f(x)$ is L_f-smooth convex, $g(x)$ is L_g-smooth and μ_g-strongly convex. Then, the function defined by*

$$\tilde{f}(x) := f(x) + g(x)$$

is $\mu_{\tilde{f}}$-strongly convex and $L_{\tilde{f}}$-smooth, where $\mu_{\tilde{f}} := \mu_g$ and $L_{\tilde{f}} := L_f + L_g$.

Remark. *Suppose $f(\cdot)$ is L-smooth and convex, and let $g(x) := \frac{\lambda}{2}\|x - x_1\|_2^2$, for some $\lambda > 0$. If we define*

$$\tilde{f}(x) := f(x) + \frac{\lambda}{2}\|x - x_1\|_2^2 ,$$

then second order characterization we have that

$$\nabla^2 g(x) = \lambda I_d.$$

This implies that g is both λ-strongly convex and λ-smooth. Let $\mu_g\colon = \lambda$ and $L_g\colon = \lambda$. Then, by the previous lemma we have that the function $\tilde{f}(x) := f(x) + \frac{\lambda}{2}\|x - x_1\|_2^2$ is $\mu_{\tilde{f}}$-strongly convex and $L_{\tilde{f}}$-smooth, where $\mu_{\tilde{f}} := \mu_g = \lambda$ and $L_{\tilde{f}} := L + L_g = L + \lambda$.

Remark. *We denote $x_* \leftarrow \arg\min_x f(x)$ assuming such an $\arg\min$ exists, similarly $\tilde{x}_* \leftarrow \arg\min_x \tilde{f}(x)$. We are going to assume such existence for reduction.*

2.1 Scenario 1.

Given an algorithm with strong guarantees for **smooth and strongly convex function**, how to make it work for **smooth convex** functions with strong guarantees.

Solution. Suppose $f(\cdot)$ is an L-smooth convex function we want to optimize. We can construct a function $\tilde{f}(x) := f(x) + \frac{\lambda}{2}\|x - x_1\|_2^2$, for some $\lambda > 0$ and apply the algorithm to $\tilde{f}(x)$, where x_1 is the initial point.

Takeaway. *Transform the L-smooth convex function to strongly convex by adding a strongly convex function.*

We want to choose λ such that the function value converges to $f(x_*)$ under certain ϵ after k steps of the algorithm (i.e. $f(x_{k+1}) - f(x_*) \le \epsilon$). First, we have

$$\begin{aligned} f(x_{k+1}) - f(x_*) &= \left(\tilde{f}(x_{k+1}) - \frac{\lambda}{2}\|x_{k+1} - x_1\|_2^2\right) - \left(\tilde{f}(x_*) - \frac{\lambda}{2}\|x_* - x_1\|_2^2\right) \\ &= \tilde{f}(x_{k+1}) - \tilde{f}(x_*) + \frac{\lambda}{2}\left(\|x_* - x_1\|_2^2 - \|x_{k+1} - x_1\|_2^2\right). \end{aligned}$$

To find a suitable λ, we could let the second term $\frac{\lambda}{2}\left(\|x_* - x_1\|_2^2 - \|x_{k+1} - x_1\|_2^2\right) \le \frac{\epsilon}{2}$. Then, we have

$$\frac{\lambda}{2}\left(\|x_* - x_1\|_2^2 - \|x_{k+1} - x_1\|_2^2\right) \le \frac{\lambda}{2}\|x_* - x_1\|_2^2 \le \frac{\lambda}{2}D,$$

where $D = \|x_* - x_1\|_2^2$ is the squared distance between x_* and x_1 in terms of the l_2-norm. We want to impose $\frac{\lambda}{2}D \le \frac{\epsilon}{2}$ so we can choose

$$\lambda = \frac{\epsilon}{D}.$$

Suppose we are doing GD $(1 - \frac{\mu_{\tilde{f}}}{L_{\tilde{f}}})$ in the algorithm. Recall that by Lemma 1 we have

$$\begin{aligned} \mu_{\tilde{f}} &= \lambda = \frac{\epsilon}{D} \\ L_{\tilde{f}} &= L + \lambda = L + \frac{\epsilon}{D}. \end{aligned}$$

Now, we need the first term to have

$$\tilde{f}(x_{k+1}) - \tilde{f}(x_*) \le \frac{\epsilon}{2}.$$

We begin with

$$\begin{aligned} \tilde{f}(x_{k+1}) - \tilde{f}(x_*) &\le \tilde{f}(x_{k+1}) - \tilde{f}(\tilde{x}_*) && (\text{since } \tilde{f}(x_*) \ge \tilde{f}(\tilde{x}_*)) \\ &\le \left(1 - \frac{\mu_{\tilde{f}}}{L_{\tilde{f}}}\right)^k \left(\tilde{f}(x_1) - \tilde{f}(\tilde{x}_*)\right) && (\text{since we perform GD } (1 - \frac{\mu_{\tilde{f}}}{L_{\tilde{f}}})) \\ &= \left(1 - \frac{\epsilon}{LD + \epsilon}\right)^k \left(\tilde{f}(x_1) - \tilde{f}(\tilde{x}_*)\right). \end{aligned}$$

Here we can use the linear rate inequality (1) to further simplify and get a certain k:

$$\begin{aligned} \tilde{f}(x_{k+1}) - \tilde{f}(x_*) &\le \left(1 - \frac{\epsilon}{LD + \epsilon}\right)^k \left(\tilde{f}(x_1) - \tilde{f}(\tilde{x}_*)\right) \\ &\le \exp\left(-\frac{\epsilon}{LD + \epsilon}k\right)\left(\tilde{f}(x_1) - \tilde{f}(\tilde{x}_*)\right) \\ &\le \frac{\epsilon}{2}. \end{aligned}$$

We then have

$$k = \frac{LD+\epsilon}{\epsilon}\log\left(\frac{2(\tilde{f}(x_1)-\tilde{f}(\tilde{x}_*))}{\epsilon}\right) = \widetilde{O}\left(\frac{LD}{\epsilon}\right),$$

where $\widetilde{O}$ denotes the complexity with the log factor hidden.

2.2 Scenario 2

Given an algorithm with strong guarantees for **smooth and strongly convex function**, how to make it work for **non-smooth and strongly convex** functions with strong guarantees.

Example: **(Support vector machine)**:

$$l(x) := \sum_{i=1}^{n} max\{0, 1-y_i z_i^T x\} + \frac{\lambda}{2}\|x\|_2^2,$$

where the term $\sum_{i=1}^{n} max\{0, 1-y_i z_i^T x\}$ is non-smooth and the term $\frac{\lambda}{2}\|x\|_2^2$ is strongly convex. Note that $l(\theta) := max(0, 1-\theta)$ is referred to as "the Hinge loss".

Solution. Suppose $f(\cdot)$ is an μ-strongly convex function we want to optimize. Assume, additionally, that $f(\cdot)$ is L_0-Lipschitz, i.e.,

$$|f(x)-f(y)| \leq L_0\|x-y\|_2,\ L_0 > 0.$$

We can construct a function $\tilde{f}_\delta(x) := \mathbb{E}_{v\sim N(0,I_d)}\left[f(x+\delta v)\right]$ and apply the algorithm to $\tilde{f}_\delta(x)$.

Remark. *We add Gaussian perturbation to make $f(x)$ smooth.*

We have the following properties[2]:

1. $f(x) \leq \tilde{f}_\delta(x) \leq f(x) + L_0\delta\sqrt{d}$,
2. $\tilde{f}_\delta(x)$ has $\frac{L_0}{\delta}$-Lipschitz gradient, i.e.,

$$\|\nabla\tilde{f}_\delta(x) - \nabla\tilde{f}_\delta(y)\| \leq \frac{L_0}{\delta}\|x-y\|.$$

Given the above properties, we have

$$\begin{aligned}
& f(x_{k+1}) - f(x_*) \\
& = \tilde{f}_\delta(x_{k+1}) - \tilde{f}_\delta(x_*) + \Big(f(x_{k+1}) - \tilde{f}_\delta(x_{k+1})\Big) + \Big(\tilde{f}_\delta(x_*) - f(x_*)\Big) \\
& \leq \tilde{f}_\delta(x_{k+1}) - \tilde{f}_\delta(x_*) + \Big(\tilde{f}_\delta(x_*) - f(x_*)\Big) && \text{(by left inequality of property 1)} \\
& \leq \tilde{f}_\delta(x_{k+1}) - \tilde{f}_\delta(x_*) + L_0\delta\sqrt{d}. && \text{(by right inequality property 1)}
\end{aligned}$$

Similarly, we first need to choose δ and want to impose

$$L_0\delta\sqrt{d} \le \frac{\epsilon}{2} \;\Rightarrow\; \delta = \frac{\epsilon}{2L_0\sqrt{d}}.$$

We have $\mu_{\tilde{f}_\delta} = \mu$ and $L_{\tilde{f}_\delta} = \frac{L_0}{\delta} = \frac{2L_0^2\sqrt{d}}{\epsilon}$, thus

$$\begin{aligned}
\tilde{f}_\delta(x_{k+1}) - \tilde{f}_\delta(x_*) &\le \tilde{f}_\delta(x_{k+1}) - \tilde{f}_\delta(\tilde{x}_*) && (\text{since } \tilde{f}_\delta(x_*) \ge \tilde{f}_\delta(\tilde{x}_*)) \\
&\le \left(1 - \frac{\mu_{\tilde{f}_\delta}}{L_{\tilde{f}_\delta}}\right)^k \left(\tilde{f}_\delta(x_1) - \tilde{f}_\delta(\tilde{x}_*)\right) && (\text{since we perform GD on } \tilde{f}) \\
&= \left(1 - \frac{\mu\epsilon}{2L_0^2\sqrt{d}}\right)^k \left(\tilde{f}_\delta(x_1) - \tilde{f}_\delta(\tilde{x}_*)\right),
\end{aligned}$$

and we expect

$$\tilde{f}_\delta(x_k) - \tilde{f}_\delta(x_*) \le \frac{\epsilon}{2}.$$

Similarly, we can use the linear rate inequality (1) and find a certain k:

$$k = \frac{2L_0^2\sqrt{d}}{\mu\epsilon}\log\left(\frac{2(\tilde{f}(x_1) - \tilde{f}(x_*))}{\epsilon}\right) = \widetilde{O}\left(\frac{L_0^2\sqrt{d}}{\mu\epsilon}\right).$$

Proof of property 1. We have

$$\begin{aligned}
f(x) &= f(\mathbb{E}_{v\sim N(0,I_d)}[x + \delta v]) && (\text{by the linearity of expectation and since } \mathbb{E}_{v\sim N(0,I_d)}[v] = 0) \\
&\le \mathbb{E}_{v\sim N(0,I_d)}[f(x + \delta v)] && (\text{by Jensen's inequality and convexity of } f) \\
&= \tilde{f}_\delta(x).
\end{aligned}$$

□

Theorem 4 (Jensen's Inequality)**.** *Let $f : J \to \mathbb{R}$ be a convex function, where $J \subseteq \mathbb{R}^d$, and let X be an integrable random variable taking values in J. Then, $g(X)$ has an expectation and*

$$g\left(\mathbb{E}\left(X\right)\right) \le \mathbb{E}\left(g\left(X\right)\right).$$

2.3 Scenario 3

A convex function $f(\cdot)$ **neither strongly convex nor smooth**.

Solution. Suppose $f(\cdot)$ is an L_0-Lipschitz convex function we want to optimize. We can combine above two previous scenarios, applying the algorithm to

$$\hat{f}_\delta(x) = \mathbb{E}_{v\sim N(0,I_\mathrm{d})}[f(x+\delta v)] + \frac{\lambda}{2}\|x - x_1\|_2^2,$$

where we have

$$\begin{aligned} \mu_{\hat{f}} &= \lambda, \\ L_{\hat{f}} &= \frac{L_0}{\delta} + \lambda. \end{aligned}$$

Bibliographic notes

Part of the materials for the reduction techniques are from Chapter 2.4 of [1]. For proofs of the properties used with scenario 2, please refer to the paper by Duchi, Bartlett and Wainwright[2].

References

[1] Elad Hazan. Introduction to Online Convex Optimization MIT Press, 2023

[2] John C. Duchi, Peter L. Bartlett, Martin J. Wainwright. Randomized Smoothing for Stochastic Optimization. arXiv:1103.4296, 2012