ECE 273 Convex Optimization and Applications Instructor: Jun-Kun Wang
Scribe: Bo Wen, Christian Coggan April 11, 2024
Editor/TA: Marialena Sfyraki

Lecture 4: Gradient Descent of Smooth Function and Introduction to Constrained Optimization

1 Review of Lecture 3

1.1 Smoothness vs. Strong Convexity

We note the difference between the first-order and second-order definitions of L-smoothness and those of μ-strong convexity — the direction of the inequalities are flipped.

1.1.1 First Order

Definition 1 (**L-smoothness**). *A differentiable function is L-smooth w.r.t.* $\|\cdot\|$, *if* $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ *we have*

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

where $L > 0$.

Definition 2 (**μ-strong convexity**). *A differentiable function* $f : C \to \mathbb{R}$ *defined over a convex set* C *is* μ*-strongly convex w.r.t. a norm* $||\cdot||$ ***if and only if*** *for any* $\mathbf{x}, \mathbf{y} \in C$ *we have*

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

for some $\mu > 0$.

1.1.2 Second Order

Definition 3 (**L-smoothness**). *A twice differentiable function* $f(\cdot) : C \to \mathbb{R}$ *defined over a set* $C \subseteq \mathbb{R}^d$ *is smooth w.r.t. a norm* $\|\cdot\|_2$, ***if and only if***, $\forall \mathbf{x} \in C$

$$\mathbf{y}^\top \nabla^2 f(\mathbf{x}) \mathbf{y} \leq L \|\mathbf{y}\|_2^2$$

for any $\mathbf{y} \in \mathbb{R}^d$.

Definition 4 (**μ-strong convexity**). *A twice differentiable function $f : C \to \mathbb{R}$ defined over a convex set $C \subseteq \mathbb{R}^d$ is μ-strongly convex w.r.t. a norm $||\cdot||$* ***if and only if*** *for any $\mathbf{x} \in C$ we have*

$$\mathbf{y}^\top \nabla^2 f(\mathbf{x}) \mathbf{y} \geq \mu \|\mathbf{y}\|^2$$

for some $\mu > 0$ and any $\mathbf{y} \in \mathbb{R}^d$.

Remark: We see that L-smoothness and μ-strong convexity provide upper and lower bounds, respectively, for the "strength" of the curvature of f at each point in its domain.

Example 1 (Smoothness): $\frac{1}{2}x^2$

Example 2 (Smoothness): $\log(1 + \exp(-x))$

Example 3 (Non-smoothness): $\max\{0, 1 - x\}$

Example 4 (Non-smoothness): $exp(-x)$

1.2 Strong Convexity implies Gradient Dominance

Definition 5 (**Gradient Dominant** or **Polyak-Lojasiewicz (PL) Condition**). *We say a function $f : \mathbb{R}^d \to \mathbb{R}$ satisfies the "Gradient Dominance" condition, or equivalently satisfies the PL-condition if, $\forall \mathbf{x} \in \mathbb{R}^d$*

$$||\nabla f(\mathbf{x})||_2^2 \geq 2\mu \left(f(\mathbf{x}) - \min_{\mathbf{x}} f(\mathbf{x}) \right) \ , \quad \textit{for some} \ \ \mu > 0.$$

Example 1 (Gradient Dominance): $f(x) = x^2 + 2\sin^2(x)$ (non-convex)

Example 2 (Gradient Dominance): Any strongly convex function.

Theorem 1. *The μ-strong convexity implies the μ-Gradient Dominant condition, i.e.,*

$$||\nabla f(\mathbf{x})||_2^2 \geq 2\mu \left(f(\mathbf{x}) - \min_{\mathbf{x}} f(\mathbf{x}) \right) \ , \quad \textit{for some} \ \ \mu > 0.$$

Remark: It is significant to note that the parameterization is identical (same μ value) for the two definitions.

2 GD in Smooth and Gradient Dominant Functions

Theorem 2. *For a function $f : \mathbb{R}^d \to \mathbb{R}$ that is both μ-gradient dominant and L-smooth, performing gradient descent with step size $\eta = \frac{1}{L}$ satisfies*

$$f(x_{k+1}) - \min_x f(x) \leq \left(1 - \frac{\mu}{L}\right)^k \left(f(x_1) - \min_x f(x)\right)$$

Remark: Note that this is a linear convergence rate. An immediate corollary of these two theorems is that a μ-strongly convex and L-smooth function would also achieve linear convergence. In fact, strengthening μ-gradient dominance to μ-strong convexity does not improve the convergence rate of GD under this analysis.

2.1 Upper Bound on Step Size η

Recall that the Gradient Descent update rule is as follows:

$$x_{k+1} = x_k - \eta \nabla f(x_k).$$

Because the function is L-smooth, we have:

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 && \text{(by } L\text{-smoothness)} \\ &\leq f(x_k) - \eta \|\nabla f(x_k)\|^2 + \frac{L\eta^2}{2}\|\nabla f(x_k)\|^2 && \text{(using the update rule)} \\ &= f(x_k) - \left(\eta - \frac{L\eta^2}{2}\right)\|\nabla f(x_k)\|^2. \end{aligned}$$

In order to guarantee that $f(\cdot)$ is always decreasing, we need that $f(x_{k+1}) \leq f(x_k)$ for any x_k. By inspection of the above equation, this is guaranteed by the condition:

$$\begin{aligned} &\eta - \frac{L\eta^2}{2} \geq 0 \\ \Leftrightarrow\ &\eta \geq \frac{L\eta^2}{2} \\ \Leftrightarrow\ &\eta \leq \frac{2}{L}. \end{aligned}$$

Thus, for a function with smoothness constant L, a step size no greater than $\frac{2}{L}$ will guarantee that the function value is decreasing at every step.

3 GD of Smooth but Non-PL Function

Question: What happens if we relax the gradient-dominant (PL) condition?

Theorem 3. *For a function $f(\cdot)$ that is both L-smooth and convex, performing Gradient Descent with step size $\eta = \frac{1}{L}$ satisfies:*

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

where $D := \max_k \|x_k - x_*\| \leq \|x_1 - x_*\|$.

Remark: Since the optimality gap is bounded by $\frac{1}{K}$ rather than α^K for some $\alpha \in [0, 1]$, this is only a sublinear rate of convergence. So, relaxing the PL condition eliminates the guarantee of a linear convergence, even for convex and smooth functions.
Remark: Observe that we have

$$K = \tilde{\Theta}\left(\frac{1}{\epsilon}\right).$$

3.1 Convergence Guarantee (Reduction)

The key idea is to make the non-gradient dominant function/non-strongly convex to a strongly convex function and approximate the convergence condition of the original function using the condition of the new function.

Lemma 1. *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 $L_{\tilde{f}} := L_f + L_g$.

Given a L-smooth convex but not strongly convex function $f(\cdot)$, let

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

Since $\frac{\lambda}{2}\|x - x_1\|_2^2$ is λ-strongly convex and also λ-smooth, i.e.,

$$g(x) := \frac{\lambda}{2}\|x - x_1\|_2^2, \quad L_g = \mu_g = \lambda,$$

the lemma gives that $\tilde{f}(x)$ is a $L_{\tilde{f}}$ smooth and $\mu_{\tilde{f}}$-strongly convex function with

$$L_{\tilde{f}} = L_f + \lambda, \quad \mu_{\tilde{f}} = \lambda. \tag{2}$$

Then, providing x_k and $x_* = \arg\min_x f(x)$, we have

$$f(x_k) = \tilde{f}(x_k) - \frac{\lambda}{2}\|x_k - x_1\|_2^2, \tag{3}$$

$$f(x_*) = \tilde{f}(x_*) - \frac{\lambda}{2}\|x_* - x_1\|_2^2. \tag{4}$$

Subtracting (3) from (2)

$$f(x_k) - f(x_*) = \tilde{f}(x_k) - \tilde{f}(x_*) + \frac{\lambda}{2}\left(\|x_* - x_1\|_2^2 - \|x_k - x_1\|_2^2\right).$$

Suppose the convergence criterion is

$$f(x_k) - f(x_*) \leq \epsilon.$$

A convenient choice is to have

$$\tilde{f}(x_k) - \tilde{f}(x_*) \leq \frac{\epsilon}{2}, \tag{5}$$

and

$$\frac{\lambda}{2}\left(\|x_* - x_1\|_2^2 - \|x_k - x_1\|_2^2\right) \leq \frac{\lambda}{2}\left(\|x_* - x_1\|_2^2\right) \leq \frac{\epsilon}{2}. \tag{6}$$

Letting $D \equiv \|x_* - x_1\|_2^2$, this approximation gives

$$\lambda = \frac{\epsilon}{D}. \tag{7}$$

For (4), since $\tilde{x}_* = \mathrm{argmin}_x \tilde{f}(x)$, $\tilde{f}(x_k) - \tilde{f}(x_*)$ is bounded by

$$\tilde{f}(x_k) - \tilde{f}(x_*) \leq \tilde{f}(x_k) - \tilde{f}(\tilde{x}_*) \leq \frac{\epsilon}{2},$$

where we have used the fact that $\tilde{x}_* := \arg\min_x \tilde{f}(x)$ and therefore $\tilde{f}(x_*) \geq \tilde{f}(\tilde{x}_*)$. We can now simply determine how many iterations on $\tilde{f}$ will be required to achieve this $\frac{\epsilon}{2}$ bound. Since $\tilde{f}(x)$ is now strongly convex as well as smooth, we can achieve linear convergence as follows:

$$\tilde{f}(x_K) - \tilde{f}(\tilde{x}_*) \leq \left(1 - \frac{\mu_{\tilde{f}}}{L_{\tilde{f}}}\right)^{K-1}\left(\tilde{f}(x_1) - \tilde{f}(\tilde{x}_*)\right) \leq \frac{\epsilon}{2},$$

which gives

$$K \geq \frac{L_{\tilde{f}}}{\mu_{\tilde{f}}} \log\left(\frac{2\left(\tilde{f}(x_1) - \tilde{f}(\tilde{x}_*)\right)}{\epsilon}\right).$$

By (1), (6), and let $\Omega = \log\left(\frac{2\left(\tilde{f}(x_1)-\tilde{f}(\tilde{x}_*)\right)}{\epsilon}\right)$, we finally have:

$$\begin{aligned}
K &\geq \frac{L_{\tilde{f}}}{\mu_{\tilde{f}}}\Omega \\
&= \frac{L+\lambda}{\lambda}\Omega && \text{(by lemma)} \\
&= \frac{LD+\epsilon}{\epsilon}\Omega && (\text{since } \lambda = \frac{\epsilon}{D}) \\
&= \tilde{O}\left(\frac{LD+\epsilon}{\epsilon}\right) \\
&= \tilde{O}\left(\frac{LD}{\epsilon}\right).
\end{aligned}$$

Thus, $K = \tilde{O}\left(\frac{LD}{\epsilon}\right)$ is the number of iterations after which convergence is guaranteed.

Remark: The reduction method has advantages such as can be flexibly applied and relatively simple to prove. However, such a method is not the optimal analysis as approximation is used.

4 Constrained Optimization

4.1 Problem Definition

A constrained optimization problem is defined as

$$\min_{x\in C} f(x), \quad \textit{where } C \subset \mathbb{R}^d \textit{ is a convex set.}$$

Remark 1: Note that there need not exist a $x_* \in C$ such that $\nabla f(x_*) = 0$. Thus, the minimum is no longer required to be a stationary point.

Remark 2: Observe that C here is a strict subset of $\mathbb{R}^d$.

We are going to show the optimality properties of the optimal point of a convex constrained optimization problem. For that, we are going to consider the case that $f(\cdot)$ is not necessarily differentiable everywhere.

4.2 Subgradient

Definition 6 (**Subgradient**). *For a function $f(\cdot)$, we say g_x is a subgradient of $f(\cdot)$ at $\mathbf{x} \in$ dom f, if $\forall \mathbf{y}$ we have*

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

Fact: If $f(\cdot)$ is convex, a subgradient at any $\mathbf{x} \in$ dom f exists.

Remark: The subgradient is useful in cases when $f(x)$ is not differentiable everywhere.

Example (Subgradient): Consider $f(x) = |x|$. Then, we have that

$$\begin{aligned} &\text{for } x > 0: \quad \nabla f(x) = 1, \\ &\text{for } x < 0: \quad \nabla f(x) = -1. \end{aligned}$$

By definition of the subgradient we have that $\forall y$

$$f(y) \geq f(x) + \langle g_x, y - x \rangle.$$

The subgradient at $x = 0$ will satisfy

$$\begin{aligned} &|y| \geq 0 + g_x(y - 0), \ \forall y \\ \Leftrightarrow \ &|y| \geq g_x y, \ \forall y. \end{aligned}$$

We have that

$$\begin{aligned} &\text{for } y \geq 0: \quad \nabla y \geq g_x y \ \Leftrightarrow \ 1 \geq g_x, \\ &\text{for } y < 0: \quad \nabla - y \geq g_x y \ \Leftrightarrow \ -1 \leq g_x. \end{aligned}$$

Hence,

$$g_{x=0} \in [-1, 1].$$

Lemma 2. *When f is convex and differentiable, $g_x = \nabla f(\mathbf{x})$.*

Bibliographic notes

More prelimiaries of calculus and linear algebra can be found in Chapter 2 of [Duchi (2010)], Chapter 3 and Chapter 4.2 of [Sidford (2024)] and Chapter 1 of [Drusvyatskiy (2020)].

References

[Duchi (2010)] John Duchi. Introductory Lectures on Stochastic Optimization. 2010.

[Sidford (2024)] Aaron Sidford. Optimization Algorithms. 2024

[Drusvyatskiy (2020)] Dmitriy Drusvyatskiy. Convex Analysis and Nonsmooth Optimization. 2020.