ECE 273 Convex Optimization and Applications — Instructor: Jun-Kun Wang
Scribe: your name — April 2, 2024
Editor/TA: Marialena Sfyraki

Lecture 1: Mathematical Background and Gradient Flow

1 Review: Calculus

We begin by reviewing some results in Calculus that will be used in this course.

Definition 1. *(**Derivative**) For a function $g(\cdot) : \mathbb{R} \to \mathbb{R}$ and $x \in \mathbb{R}$, consider*

$$L = \lim_{\delta \to 0} \frac{g(x+\delta) - g(x)}{\delta}.$$

The function $g(\cdot)$ is said to be "differentiable" if this limit exits for all $x \in \mathbb{R}$. In that case, L is called the "derivative" of $g(\cdot)$. We denote the derivative as $g'(x)$, $\dot{g}(x)$, or $\frac{dg(x)}{dx}$.

Definition 2. *(**Gradient**) For a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^d$, the gradient is*

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{bmatrix},$$

where

$$\frac{\partial f}{\partial x_1} = \lim_{\delta \to 0} \frac{f(x_1 + \delta; x_2; \ldots; x_d) - f(x_1; x_2; \ldots; x_d)}{\delta}.$$

Remark: The gradient of f is a function from $\mathbb{R}^d$ to $\mathbb{R}^d$, and can be pictured as a vector field (or vector-valued function), which gives the direction and the rate of the fastest increase.

Definition 3. *(**Hessian**) For a twice continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^d$, the Hessian matrix of $f(\cdot)$ at $\mathbf{x}$ is defined by*

$$\nabla^2 f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \frac{\partial^2 f}{\partial x_d \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_d^2} \end{bmatrix} \in \mathbb{R}^{d \times d}$$

Remark: The Hessian is a symmetric matrix.

Example: Let $f : \mathbb{R}^d \to \mathbb{R}$ be defined by $f(\mathbf{x}) = x_1^2 x_2$. Then

$$\nabla f(\mathbf{x}) = \begin{bmatrix} 2x_1x_2 \\ x_1^2 \end{bmatrix} \in \mathbb{R}^2,$$

and

$$\nabla^2 f(\mathbf{x}) = \begin{bmatrix} 2x_2 & 2x_1 \\ 2x_1 & 0 \end{bmatrix} \in \mathbb{R}^{2\times 2}.$$

Theorem 1. ***(Fundamental Theorem of Calculus):*** *Let $f : [a,b] \to \mathbb{R}$ be a continuously differentiable function. Then,*

$$f(b) - f(a) = \int_a^b f'(\theta)\, d\theta.$$

Theorem 2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. Define*

$$\mathbf{x}_\alpha = (1-\alpha)\mathbf{x} + \alpha\mathbf{y},$$

for some $\alpha \in [0,1]$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. Then,

$$f(\mathbf{y}) - f(\mathbf{x}) = \int_0^1 \langle \nabla f(\mathbf{x}_\alpha),\ \mathbf{y} - \mathbf{x} \rangle d\alpha$$

Additionally, if f is twice differentiable, then

$$\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}) = \int_0^1 \nabla^2 f(\mathbf{x}_\alpha)(\mathbf{y} - \mathbf{x}) d\alpha,$$

where $\nabla^2 f(\mathbf{x}_\alpha) \in \mathbb{R}^{d\times d}$ and $(\mathbf{y} - \mathbf{x}) \in \mathbb{R}^d$.

Theorem 3. *(**Chain Rule**): Let $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ be differentiable functions, and let $x \in \mathbb{R}$. Then, the composite function $h : \mathbb{R} \to \mathbb{R}$ given by $h(x) = f\big(g(x)\big)$ is differentiable on $\mathbb{R}$ and its derivative is given by*

$$h'(x) = f'\big(g(x)\big) \cdot g'(x)$$

Remark: This rule can be extended to functions of several variables. In general, if $y = g(z)$ and $z = h(x)$, the chain rule is expressed as:

$$\frac{dy}{dx} = \frac{dy}{dz} \cdot \frac{dz}{dx}$$

This formula shows how the rate of change of a composite function is influenced by the rates of change of its components.

2 Norm

Consider a fixed vector $\mathbf{x} \in \mathbb{R}^d$. We define

l_2-Norm:

$$||\mathbf{x}||_2 = \sqrt{\sum_{i=1}^{d} x_i^2}$$

l_1-Norm:

$$||\mathbf{x}||_1 = \sum_{i=1}^{d} |x_i|$$

l_∞-Norm:

$$||\mathbf{x}||_\infty = \max_i \{|x_i|\}$$

Definition 4. *(**Cauchy-Schwartz Inequality**): For every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ we have*

$$\langle \mathbf{x}, \mathbf{y} \rangle \ \leq ||\mathbf{x}||_2 ||\mathbf{y}||_2,$$

where $\langle \cdot, \cdot \rangle$ is the inner-product.

3 Rates of Convergence

A solid and sound comparison of numerical methods relies on precise rates of progress in the iterates. For example, we might measure the progress an algorithm via the optimality gap.

Definition 5. *(**Optimality Gap**): Given a function f such that $f : \mathbb{R}^d \to \mathbb{R}$, the optimality gap is the difference between the value of f at $\mathbf{x}_t \in \mathbb{R}^d$ for some $t \in \mathbb{R}$ and the optimal value, i.e.*

$$f(\mathbf{x}_t) - \min_{\mathbf{x}} f(\mathbf{x}).$$

Fix a sequence of real numbers $a_k > 0$ with $a_k \to 0$

- **Sublinear rate:** We say a_k converges sublinearly if there exist constants $c > 0, q > 0$ satisfying

$$a_k \leq \frac{c}{k^q} \text{ for all } k. \tag{1}$$

Larger q and smaller c indicates a faster convergence reate.

From (1), we decude that the number of iterations k such that $a_k \leq \epsilon$ is

$$k \geq \left(\frac{c}{\epsilon}\right)^{1/q}. \tag{2}$$

Note that the importance of the value of c should not be discounted; the convergence rate depends strongly on this value.

- **Linear rate:**

  We say a_k converges linearly if there exist constants $c > 0, q \in (0, 1]$ satisfying

  $$a_k \leq c(1-q)^k \text{ for all } k. \tag{3}$$

  In this case, we call $1 - q$ the linear rate of convergence.

  From (3), we decude that the number of iterations k such that $a_k \leq \epsilon$ for a target accuracy ϵ is

  $$c(1-q)^k \leq \epsilon \iff k \geq \frac{-1}{\log(1-q)} \log\left(\frac{c}{\epsilon}\right). \tag{4}$$

  Taking into account the inequality $\log(1-q) \leq -q,$ for $q \in [0, 1]$, we deduce that $a_k \leq \epsilon$ for every

  $$k \geq \frac{1}{q} \log\left(\frac{c}{\epsilon}\right). \tag{5}$$

  The dependency on q is strong, while the dependency on c is very weak (as c is inside the log).

4 Gradient Descent and Gradient Flow

A formal specification of the Gradient Descent (GD) algorithm follows.

Algorithm 1 GRADIENT DESCENT

Input: an initial point $\mathbf{x}_0 \in \textbf{dom } f$ and step size η.
for $k = 1$ to K **do**
  $\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k - \eta \nabla f(\mathbf{x_k})$
end for
Return $\mathbf{x}_{k+1}$.

Remark: The parameter η is called the *step size* or *learning rate*.

In order to better understand gradient descent, let's consider the curve that at each instant proceeds in the direction of steepest descent of f. For this method, let's consider a function $f : X \to \mathbb{R}$, the method of gradient flow starts at some initial point $x_0 \in X$ and seek to find the optimum of f by following the integral curve defined by the following differential equations.

Definition 6. *(**Gradient Flow**): Let $f : \mathbb{R}^d \to \mathbb{R}$ be a smooth function. Gradient flow is a smooth curve $\mathbf{x} : \mathbb{R} \to \mathbb{R}^d$ such that*

$$\frac{d\mathbf{x}(t)}{dt} = -\nabla f\left(\mathbf{x}(t)\right)$$

4.1 Insights into the Algorithm

Gradient Flow is Gradient Descent as $\eta \to 0$. More specifically, consider

$$\lim_{\eta \to 0} \frac{\mathbf{x}_{k+1} - \mathbf{x}_k}{\eta} = \lim_{\eta \to 0} -\nabla f(\mathbf{x}_k)$$
$$\Leftrightarrow \frac{d\mathbf{x}}{dt} = -\nabla f(\mathbf{x})$$

Consider applying Gradient Flow to $\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$, that is

$$\frac{d\mathbf{x}(t)}{dt} = -\nabla f\left(\mathbf{x}(t)\right).$$

Then,

$$\begin{aligned}\frac{df}{dt} &= \sum_i^d \frac{\partial f}{\partial x_i} \frac{\partial x_i}{\partial t} \\ &= \left\langle \nabla f(\mathbf{x}), \frac{d\mathbf{x}(t)}{dt} \right\rangle \\ &= \langle \nabla f(\mathbf{x}), -\nabla f(\mathbf{x}) \rangle \\ &= -||\nabla f(\mathbf{x})||_2^2 \\ &\leq 0\end{aligned}$$

Thus, as long as $\nabla f(\mathbf{x}) \neq \mathbf{0}$, the function is always decreasing. This does not necessarily imply that it finds the optimal point.

4.2 Gradient Dominant Condition

Definition 7. *(**Gradient Dominant** or **Polyak-Lojasiewicz (PL) Condition**): We say a function $f : \mathbb{R}^d \to \mathbb{R}$ satisfies the "Gradient Dominance" condition if*

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

We say that f is μ-gradient dominant.

Definition 8. ***(Stationary Point)***: *Given a differentiable function f such that $f : \mathbb{R}^d \to \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^d$, a stationary point is a point such that*

$$\nabla f(\mathbf{x}) = \mathbf{0} \in \mathbb{R}^d.$$

Remark: For any function satisfying the P.L. condition, every stationary point is a global optimum point.

Example 1: All strongly convex functions

Example 2: $f(x) = x^2 + 2\sin^2(x)$

Consequence: Suppose that f is additionally μ-gradient dominant. Then, taking the derivative of an optimality gap we get

$$\begin{aligned} \frac{d(f(\mathbf{x}_t) - f_*)}{dt} &= \frac{df(\mathbf{x}_t)}{dt} && \text{, as } f_* \text{ is a constant} \\ &= -||\nabla f(\mathbf{x}_t)||_2^2 && \text{, by Gradient Flow} \\ &\leq -2\mu \left(f(\mathbf{x}_t) - \min_{\mathbf{x}} f(\mathbf{x}) \right) && \text{, since } f \text{ is } \mu\text{-gradient dominant} \end{aligned} \tag{6}$$

Inequality (1) implies that

$$f(\mathbf{x}_t) - \min_{\mathbf{x}} f(\mathbf{x}) \leq e^{-2\mu t} \left(f(\mathbf{x}_0) - \min_{\mathbf{x}} f(\mathbf{x}) \right) \tag{7}$$

for μ-gradient dominant functions, where $\mathbf{x}_0$ is the initial point.

Why does (1) imply (2)? Let

$$\theta_t := f(\mathbf{x}_t) - f_*.$$

Then, inequality (1) can be expressed as

$$\begin{aligned} &\frac{d\theta_t}{dt} \leq -2\mu\theta_t \\ &\Leftrightarrow \frac{d\theta_t}{\theta_t} \leq -2\mu dt \\ &\Rightarrow \int_{\theta_0}^{\theta_t} \frac{d\theta_t}{\theta_t} \leq \int_0^t -2\mu dt \\ &\Leftrightarrow \log(\theta_t) - \log(\theta_0) \leq -2\mu t && \text{, since } \frac{d}{dx}\log x = \frac{1}{x}. \end{aligned}$$

Therefore,

$$\frac{\theta_t}{\theta_0} \leq exp(-2\mu t)$$
$$\Leftrightarrow \theta_t \leq \theta_0 exp(-2\mu t)$$

Plugging back in, we get

$$f(\mathbf{x}_t) - \min_{\mathbf{x}} f(\mathbf{x}) \leq \exp\left(-2\mu t\right)\left(f(\mathbf{x}_0) - \min_{\mathbf{x}} f(\mathbf{x})\right)$$

Bibliographic notes

More prelimiaries of calculus and linear algebra can be found in Chapter 1 of [Drusvyatskiy (2020)] and Chapter 2 of [Vishnoi (2021)].

References

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

[Vishnoi (2021)] Nisheeth K. Vishnoi. Algorithms for Convex Optimization. Cambridge University Press, 2021