DSC 211 Introduction to Optimization Winter 2024 Instructor: Jun-Kun Wang
Scribe: Berry Ma, Jake Ryan January 11, 2024
Editors/TAs: Merlin Chang, Marialena Sfyraki

Lecture 2: 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

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

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 $h(x) = f\left(g\left(x\right)\right)$ is differentiable on $\mathbb{R}$ and its derivative is given by*

$$h'(x) = f'\left(g\left(x\right)\right) \cdot g'(x)$$

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 Gradient Descent and Gradient Flow

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

Algorithm 1 Gradient Descent

input a starting point $\mathbf{x}_0 \in$ **dom** f and step size η.

0: **for** $k = 0, 1, \ldots$ **do**

1: $\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k - \eta \nabla f(\mathbf{x_k})$

2: **End (for)**

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[3].

Definition 5. *(**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*

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

3.1 Insights into the Algorithm

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

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

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.

3.2 Gradient Dominant Condition

Definition 6. *(**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 7. *(**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: $f(x) = x^2 + 4\sin^2(x)$

Definition 8. *(**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 minimum value of f over all possible $\mathbf{x} \in \mathbb{R}^d$, i.e.*

$$f(\mathbf{x}_t) - \min_{\mathbf{x}} f(\mathbf{x}) = f(\mathbf{x}_t) - f_*,$$

where

$$f_* := \min_{\mathbf{x}} f(\mathbf{x}).$$

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

$$\begin{aligned} \frac{d(f(\mathbf{x}_t) - f_*)}{dt} &= \frac{df(\mathbf{x}_t)}{dt} && \text{, as } f_* \text{ is a constant} \\ &= -||\nabla f(\mathbf{x}_t)||_2^2 && \text{, by Gradient Flow} \\ &\leq -2\mu \left(f(\mathbf{x}_t) - \min_{\mathbf{x}} f(\mathbf{x}) \right) && \text{, since } f \text{ is } \mu\text{-gradient dominant} \end{aligned} \tag{1}$$

Inequality (1) implies that

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

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

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

Therefore,

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

Plugging back in, we get

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

4 Convex Sets and Functions

Definition 9. *(**Convex Sets**): A set $C \subseteq \mathbb{R}^d$ is called convex if for any $\mathbf{x}, \mathbf{y} \in C$ and any $\alpha \in [0, 1]$, we have*

$$\alpha\mathbf{x} + (1-\alpha)\mathbf{y} \in C.$$

Remark: This property ensures that the line segment between any two points in the set lies entirely within the set. Geometrically, it implies that the set does not have any "holes" or "gaps".

Example 1: The vector space $\mathbb{R}^d$.

Example 2: Hyper-planes

$$\{\mathbf{x} \in \mathbb{R}^d : \langle \mathbf{x}, \mathbf{a} \rangle = c\},\ \mathbf{a} \in \mathbb{R}^d,\ c \in \mathbb{R}$$

.

Example 3: Half-spaces

$$\{\mathbf{x} \in \mathbb{R}^d : \langle \mathbf{x}, \mathbf{a} \rangle \leq c\},\ \mathbf{a} \in \mathbb{R}^d,\ c \in \mathbb{R}$$

Example 4: l_p-Norm-balls with $p \geq 1$

$$\begin{aligned} &||\mathbf{x}||_p \leq r \\ \Leftrightarrow &(\sum_i |x_i|^p)^{\frac{1}{p}} \leq r \end{aligned}$$

Definition 10. *(**Zero Order Characterization of Convex Functions**): A function $f : C \to \mathbb{R}$ defined over a convex set C is called convex if, for any $\mathbf{x}, \mathbf{y} \in C$ and any $\alpha \in [0, 1]$, the following inequality holds*

$$f\left(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}\right) \leq \alpha f(\mathbf{x}) + (1-\alpha)f(\mathbf{y}).$$

Remark: In other words, a function is convex if the line segment between any two points on its graph lies above the graph itself. Geometrically, this means that the function does not have any "hills" between its points.

Remark: The opposite applies for concave functions, i.e.

$$f\left(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}\right) \geq \alpha f(\mathbf{x}) + (1-\alpha)f(\mathbf{y}).$$

For linear functions, the equality holds (they are both convex and concave)

$$f\left(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}\right) = \alpha f(\mathbf{x}) + (1-\alpha)f(\mathbf{y}).$$

Theorem 4. *(**First Order Characterization of Convex Functions**): A differentiable function $f : C \to \mathbb{R}$ defined over a convex set C is called convex **if and only if**, for any $\mathbf{x}, \mathbf{y} \in C$*

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

Theorem 5. *(**Second Order Characterization of Convex Functions**): A twice-differentiable function $f : C \to \mathbb{R}$ defined over a convex set C is convex **if and only if**, for any $\mathbf{x} \in C$, the Hessian matrix evaluated at $\mathbf{x}$ is positive semi-definite, i.e.*

$$\nabla^2 f(\mathbf{x}) \succeq 0.$$

Example 1: Linear Functions

$$f\left(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}\right) = \alpha f(\mathbf{x}) + (1-\alpha)f(\mathbf{y})$$

Example 2: Quadratic Functions

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{b}^T\mathbf{x},\ \lambda_{min}(\mathbf{A}) \geq 0$$

Example 3: Negative Entropy

$$f(\mathbf{x}) = \sum_i^d x_i \log x_i,\ \mathbf{x} \in \mathbb{R}^d$$

Example 4: Non-negative weighted sum of convex functions

$$F(\mathbf{x}) = \sum_{i=1}^n \alpha_i f_i(\mathbf{x}),\ \alpha_i \geq 0, \forall i$$

Example 5: Sum of squared difference loss

$$\sum_{i=1}^n \frac{1}{2}\left(y_i - \mathbf{x}^\top \mathbf{z}_i\right)^2$$

4.1 Strongly Convex Functions

Definition 11. *(**Zero Order Characterization of μ-Strongly Convex Functions**): 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, for any $\mathbf{x}, \mathbf{y} \in C$ and any $\alpha \in [0, 1]$ we have*

$$f\left((1-\alpha)\mathbf{x} + \alpha\mathbf{y}\right) \leq (1-\alpha)f(\mathbf{x}) + \alpha f(\mathbf{y}) - \frac{\mu}{2}\alpha(1-\alpha)||\mathbf{y} - \mathbf{x}||^2,$$

for some $\mu > 0$.

Theorem 6. *(**First Order Characterization of μ-Strongly Convex Functions**): 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 $\mu > 0$.

Theorem 7. *(**Second Order Characterization of μ-Strongly Convex Functions**): 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 $\mathbf{x} \in C$ we have*

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

for some $\mu > 0$.

Remark: Theorem 7 is useful when showing a function is strongly convex when we have the function's Hessian.

Bibliographic notes

More prelimiaries of calculus and linear algebra can be found in Chapter 2 of [1]. For Gradient Flow, there is nice blog article [2] and the introduction to Behrman's dissertation[3].

References

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

[2] Francis Bach. Effortless optimization through gradient flows `https://urldefense.com/v3/__https://francisbach.com/gradient-flows/__;!!Mih3wA!HgIABJB-qreJ1KQQ59Z8wY8Z76bb1cNv_aknz_3HziNXP0njTYuxKUk1fSak8MS_Me7L1u5i79ioON0$`

[3] William Behrman. An Efficient Gradient Flow Method for Unconstrained Optimization `https://web.stanford.edu/group/SOL/dissertations/behrmanthesis.pdf`