CS2910 - Optimization	Summer 2023
Lecture 2 — Gradient Descent	
Lecturer: Yunwei Ren	*Scribed by Zhidan Li*

Contents

1 Overview

In this lecture we focus on how to solve the optimization problem. Firstly we will introduce a time-continuous method called *gradient flow*, and analyze its efficiency. Then we turn our sight on a time-discrete method — *gradient descent*, and analyze the efficiency of gradient descent and compare it with gradient flow.

The following lemma known as *Gronwall lemma* will be useful in our analysis

Lemma 1 (Gronwall lemma)**.** *For a time-continuous non-negative process (or a path, for short) u_t, suppose that $\dot{u}_t \le \alpha_t u_t$. Then we have*

$$u_t \le u_0 \exp\left(\int_0^T \alpha_t \, dt\right).$$

Remark 1. The path u_t satisfying $\dot{u}_t = Au_i$ is called a *linear system*. Its solution is $u_T = u_0 \exp(AT)$. Its discrete version is the sequence $\{u_k\}_{k\in\mathbb{N}}$ satisfying

$$u_{k+1} - u_k = Au_k, \forall k \in \mathbb{N}.$$

Then $u_k = u_0(1 + A)^k$. This means the exponential growth/decreasing rate.

2 Gradient Flow

Now we introduce the method called *gradient flow* to solve the optimization problem.

Definition 2 (gradient flow (GF))**.** *For a function $f \in C^1(\mathbb{R}^d)$, we define the **gradient flow** of f with initial point $\hat{x} \in \mathbb{R}^d$ as the solution to the initial value problem:*

$$\dot{x}_t = -\nabla f(x_t), x_0 = \hat{x}.$$

Remark 2. By the chain rule,

$$\frac{d}{dt} f(x_t) = \langle \nabla f(x_t), \dot{x_t} \rangle = -\|\nabla f(x_t)\|^2 \leq 0.$$

This means $f(x_t)$ is not increasing.

Now we show that, the gradient flow will converge to the minimizer if f is strongly convex. The following proposition shows strong convexity implies linear convergence rate.

Proposition 3. *Suppose that the function $f : \mathbb{R}^d \to \mathbb{R}$ is a $C^1(\mathbb{R}^d)$, μ-strongly convex function. Let $x_* = \operatorname{argmin}_{x \in \mathbb{R}^d} f(x)$, x_t be the gradient flow of f. For all $\varepsilon > 0$, we have*

$$\|x_T - x_*\|_2 \leq \varepsilon, \forall T \geq \mu^{-1} \log\left(\frac{\|x_0 - x_*\|_2}{\varepsilon}\right).$$

Proof. Consider the function $t \mapsto \|x_t - x_*\|_2^2$. Then, by the chain rule,

$$\begin{aligned} \frac{d}{dt}\|x_t - x_*\|_2^2 &= 2\left\langle x_t - x_*, \frac{d}{dt} x_t \right\rangle \\ &= -2\left\langle x_t - x_*, \nabla f(x_t) \right\rangle \\ &= -2\left\langle x_t - x_*, \nabla f(x_t) - \nabla f(x_*) \right\rangle \\ &\leq -2\mu\|x_t - x_*\|_2^2 \end{aligned}$$

where the last inequality holds from the strong convexity. Then by Lemma 1, it holds that

$$\|x_T - x_*\|_2^2 \leq \|x_0 - x_*\|_2^2 \exp\left(-2\mu T\right).$$

□

If f is not necessarily strongly convex, for $f(x_t)$ we also have the following approximation.

Proposition 4. *Let $f \in C^1(\mathbb{R}^d)$ be a convex function, $(x_t)_t$ be the gradient flow of f and $x_* = \operatorname{argmin}_{x \in \mathbb{R}^d} f(x)$ be the minimizer. Then,*

$$f(x_T) \leq f(x_*) + \frac{\|x_0 - x_*\|_2^2}{2T}.$$

Proof. Since f is convex, by lower linear bound, we have

$$\langle \nabla f(x), x - x_* \rangle \geq f(x) - f(x_*)$$

which implies

$$\begin{aligned} \frac{d}{dt}\|x_t - x_*\|_2^2 &= -2\left\langle x_t - x_*, \nabla f(x_t) \right\rangle \\ &\leq -2(f(x_t) - f(x_*)). \end{aligned}$$

Integrating both sides, we obtain

$$\|x_T - x_*\|_2^2 - \|x_0 - x_*\|_2^2 \le -2\int_0^T f(x_t)\,dt + 2Tf(x_*) \le 2T(f(x_*) - f(x_T)).$$

where the last inequality holds since f is not increasing. Rearranging terms, we conclude

$$f(x_T) \le f(x_*) + \frac{\|x_0 - x_*\|_2^2}{2T}.$$

□

Remark 3. The above proposition illustrates a phenomenon that, when f is almost flat, although the movement of x_t is slow, since f is convex, $f(x_T)$ is near $f(x_*)$ in those regions. Thus we can track $f(x_T)$ as an approximation of $f(x_*)$.

2.1 Polyak-Lojasiewicz condition

Surprisingly, when f is not necessarily convex, gradient flow might be efficient when f meets some regular conditions.

Definition 5 (Polyak-Lojasiewicz condition)**.** *For a function $f \in C^1(\mathbb{R}^d)$ (not necessarily convex), let $f_* = \inf_{x\in\mathbb{R}^d} f(x)$. We say f satisfies the Polyak-Lojasivewicz (PL) condition with PL constant $\mu > 0$ if*

$$\frac{1}{2}\|\nabla f(x)\|^2 \ge \mu(f(x) - f_*).$$

Remark 4. Note that, since $\frac{d}{dt}f(x_t) = -\|\nabla f(x)\|^2$, PL condition implies the linear convergence. Also, it gives the message that, to get the rate of convergence, it suffices to lower bound $\|\nabla f(x)\|$. Then another strategy is picking a descent direction u

$$\|\nabla f(x_t)\| \ge \langle \nabla f(x_t), u/\|u\| \rangle .$$

Lemma 6. *For a μ-strongly convex function f, it also satisfies PL condition with μ.*

Proof. Suppose f is μ-strongly convex. Then we have

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

This means

$$\min_{y_1\in\mathbb{R}^d} f(y) \ge \min_{y_2\in\mathbb{R}^d} \left\{ f(x) + \langle \nabla f(x), y_2 - x\rangle + \frac{\mu}{2}\|y_2 - x\|^2 \right\}$$

which is exactly the following inequality:

$$f_* \ge f(x) - \frac{1}{2\mu}\|\nabla f(x)\|^2.$$

Rearranging the inequality we prove what we desire. □

3 Gradient Descent

Although the gradient flow is efficient (under some regular conditions), it is hard to implement it since it is a time-continuous path. Now we introduce its discrete version — gradient descent.

Definition 7 (gradient descent (GD)). *Given a function $f \in C^1(\mathbb{R}^d)$, the gradient descent of f with starting point $\hat{x}$ and a step size $\eta > 0$ is the sequence $\{x_k\}_{k\in\mathbb{N}}$ satisfying:*

$$x_{k+1} = x_k - \eta \nabla f(x_k), x_0 = \hat{x}.$$

We compare the gradient descent with the gradient flow. For the gradient descent,

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

For the gradient flow,

$$x_{(k+1)\eta} = x_{k\eta} - \int_{k\eta}^{(k+1)\eta} \nabla f(x_t)\, dt.$$

Directly from the comparison, intuitively we observe that, if ∇f doesn't change too fast, then GD $\approx$ GF.

Definition 8. *A function $f \in C^1(\mathbb{R}^d)$ is said to be L-smooth for $L \geq 0$ if its gradient is L-Lipschitz, i.e.,*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \forall x, y \in \mathbb{R}^d.$$

Remark 5. For L-smoothness, we have the reverse PL conditions, i.e.,

$$\frac{1}{2}\|\nabla f(x)\|^2 \leq L(f(x) - f_*).$$

Lemma 9 (equivalent definitions). *For a function $f \in C^1(\mathbb{R}^d)$, the followings are equivalent:*

(a) f is L-smooth.

(b) $\left\|\nabla^2 f(x)\right\|_2 \leq L$.

(c) (Two-sided) f has upper quadratic bound, i.e., for all $x, y \in \mathbb{R}^d$,

$$f(y) \in \left[f(x) + \langle \nabla f(x), y - x\rangle - \frac{L}{2}\|x - y\|^2, f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|x - y\|^2\right].$$

Now we focus on the efficiency of the gradient descent.

Lemma 10 (descent lemma). *For an L-smooth function $f \in C^1(\mathbb{R}^d)$ (not necessarily convex), and $\eta \leq 1/L$, we have*

$$f(x_{k+1}) \leq f(x_k) - \frac{\eta}{2}\|\nabla f(x_k)\|^2.$$

Proof. Since f is L-smooth, by the upper quadratic bound, we have

$$\begin{aligned} f(x_{k+1}) &\le f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ &= f(x_k) - \eta\|\nabla f(x_k)\|^2 + \frac{L\eta^2}{2}\|\nabla f(x_k)\|^2 \\ &= f(x_k) - \eta(1 - \frac{\eta L}{2})\|\nabla f(x_k)\|^2 \\ &\le f(x_k) - \frac{\eta}{2}\|\nabla f(x_k)\|^2. \end{aligned}$$

□

Remark 6. Note that if we only want to make $f(x_k)$ not increasing, then $\eta < 2/L$ is enough. Additionally, this lemma might be meaningless in non-convex optimization because of the existence of the EoS phenomenon.

The following corollary comes immediately from Lemma 10 by summing over both sides and rearranging terms.

Corollary 11. *Within $\frac{2}{\eta\varepsilon}(f(x_0) - f(x_*))$ iterations, GD with $\eta \le 1/L$ can find a point x with $\|\nabla f(x)\|^2 \le \varepsilon$.*

Remark 7. For a μ-strongly convex function f, since it also satisfies μ-PL condition, then the condition $\|\nabla f(x_k)\|^2 \le \varepsilon$ implies $f(x_k) - f_* \le \varepsilon/\mu$.

Then for strongly convex functions, we have the following convergence rate.

Proposition 12. *For a μ-strongly convex, L-smooth function $f \in C^1(\mathbb{R}^d)$, Let the minimizer $x_* = \operatorname{argmin}_{x\in\mathbb{R}^d} f(x)$ and $\eta \le 1/L$. Then we have*

$$\|x_k - x_*\|^2 \le (1 - \eta\mu)^k\|x_0 - x_*\|^2.$$

Proof. By definition,

$$\begin{aligned} \|x_{k+1} - x_*\|^2 &= \|x_k - \eta\nabla f(x_k) - x_*\|^2 \\ &= \|x_k - x_*\|^2 - 2\eta\langle \nabla f(x_k), x_k - x_* \rangle + \eta^2\|\nabla f(x_k)\|^2 \\ &\le \|x_k - x_*\|^2 - 2\eta\left(f(x_k) - f(x_*) + \frac{\mu}{2}\|x_k - x_*\|^2\right) + \eta^2\|\nabla f(x_k)\|^2 \\ &\le \|x_k - x_*\|^2 - 2\eta\left(f(x_k) - f(x_*) + \frac{\mu}{2}\|x_k - x_*\|^2\right) + 2\eta^2 L(f(x_k) - f_*) \end{aligned}$$

where the first inequality comes from the lower linear bound, and the second inequality comes from the reverse PL condition of L-smoothness. Rearranging terms, we obtain

$$\|x_{k+1} - x_*\|^2 \le (1 - \eta\mu)\|x_k - x_*\|^2 - 2\eta(1 - \eta L)(f(x_k) - f_*) \le (1 - \eta\mu)\|x_k - x_*\|^2.$$

□

Remark 8. Note that this bound is dimension-free. This bound is also tighter than the one deduced from descent lemma with PL condition. Consider the function $x \mapsto \frac{1}{2}\|x\|^2$. This proposition means we only need to choose a proper step size.

3.1 Convergence of gradient descent without strong convexity

Now we establish the convergence of gradient descent without strong convexity. Firstly we introduce some basic lemmas.

Lemma 13 (law of cosines)**.** *For all $x, y, z \in \mathbb{R}^d$, it holds that*

$$\langle z - x, y - x \rangle = \frac{1}{2}\left(\|y - x\|^2 + \|z - x\|^2 - \|y - z\|^2\right).$$

Lemma 14 (basic mirror descent lemma)**.** *For a C^1 convex function $f : \mathbb{R}^d \to \mathbb{R}$, for all $y \in \mathbb{R}^d$,*

$$f(x_k) \leq f(y) + \frac{1}{2\eta}\left(\|y - x_k\|^2 - \|y - x_{k+1}\|^2 + \|x_{k+1} - x_k\|^2\right).$$

Proof. By the lower linear bound, it holds that

$$\begin{aligned} f(y) &\geq f(x_k) + \langle \nabla f(x_k), y - x_k \rangle \\ &= f(x_k) + \frac{1}{\eta}\langle x_k - x_{k+1}, y - x_k \rangle \\ &= f(x_k) - \frac{1}{2\eta}\left(\|y - x_k\|^2 - \|y - x_{k+1}\|^2 + \|x_{k+1} - x_k\|^2\right). \end{aligned}$$

□

Remark 9. When $\|x_k - x_{k+1}\|$ is small and $y = x_*$, the lemma shows

$$f(x_k) - f_* \lesssim \frac{1}{2\eta}\left(\|x_* - x_k\|^2 - \|x_* - x_{k+1}\|^2\right).$$

Namely, we lower bound the distance x_k moves within one step by the suboptimality.

Now we establish the convergence rate.

Proposition 15. *For an L-smooth C^1 convex function $f : \mathbb{R}^d \to \mathbb{R}$, choose $\eta \leq 1/L$. Then it holds that*

$$f(x_T) \leq f(x_*) + \frac{\|x_0 - x_*\|^2}{\eta T}.$$

Proof. By Lemma 14, choose $y = x_*$,

$$f(x_k) \leq f(x_*) + \frac{1}{2\eta}\left(\|x_* - x_k\|^2 - \|x_* - x_{k+1}\|^2 + \|x_{k+1} - x_k\|^2\right).$$

Summing over from 0 to $T - 1$, we obtain

$$\begin{aligned} \sum_{k=0}^{T-1} f(x_k) &\leq T f(x_*) + \frac{1}{2\eta}\left(\|x_0 - x_*\|^2 - \|x_T - x_*\|^2 + \sum_{k=0}^{T-1}\|x_{k+1} - x_k\|^2\right) \\ &\leq T f(x_*) + \frac{1}{2\eta}\|x_0 - x_*\|^2 + \frac{1}{2\eta}\sum_{k=0}^{T-1}\|x_{k+1} - x_k\|^2 \\ &= T f(x_*) + \frac{1}{2\eta}\|x_0 - x_*\|^2 + \frac{1}{2}\eta\sum_{k=0}^{T-1}\|\nabla f(x_k)\|^2 \\ &\leq T f(x_*) + \frac{1}{2\eta}\|x_0 - x_*\|^2 + (f(x_0) - f(x_T)) \end{aligned}$$

where the last inequality holds by Lemma 10. Then

$$\begin{aligned}
f(x_T) &\le \frac{1}{T}\sum_{k=0}^{T-1} f(x_k) \\
&\le f(x_*) + \frac{1}{2T\eta}\|x_0 - x_*\|^2 + \frac{1}{T}(f(x_0) - f(x_T)) \\
&\le f(x_*) + \frac{1+\eta L}{2T\eta}\|x_0 - x_*\|^2 \\
&\le f(x_*) + \frac{\|x_0 - x_*\|^2}{\eta T}
\end{aligned}$$

where the third inequality comes from f is L-smooth. □