CS2910 - Optimization Summer 2023

Lecture 8 — Escaping Saddle Points

Lecturer: Yunwei Ren *Scribed by Zhidan Li*

Contents

1 Overview

In this lecture, we move to an advanced topic in optimization: how to escape the saddle points. The existence of saddle points might make the algorithms of gradient descent fail to (approximately) find a minimizer efficiently. We will show how to escape such kind of 'bad' points under some conditions.

1.1 Preliminary

The following anti-concentration of ball volume will be of great significance in our analysis.

Lemma 1. *Let $x \sim \mathrm{Unif}(r\mathbb{B}^d)$. For all $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$|x_1| \geq r\delta/(2\sqrt{d}).$$

2 Introduction

To precisely describe the topic, firstly we introduce the *local minimizer*.

Definition 2. *Given a function $f : \mathbb{R}^d \to \mathbb{R}$, we say $x \in \mathbb{R}^d$ is a* ***local minimizer*** *of f if there exists an open set $U \subseteq \mathbb{R}^d$ such that $x \in U$ and $f(x) \leq f(x')$ for all $x' \in U$.*

To show the local minimizer, it suffices to show the *second-order condition* holds.

Fact 3. *Given a function $f : \mathbb{R}^d \to \mathbb{R}$, for a point $x \in \mathbb{R}^d$, if*

$$\nabla f(x) = 0, \ \nabla^2 f(x) \succ 0,$$

then x is a strict local minimizer.

Remark 1. Note that, the condition $\nabla f(x) = 0, \ \nabla^2 f(x) \succeq 0$ does not necessarily imply the local minimizer. We say such a point x is a second-order stationary point.

Example: Consider the function $f(x) = x^3$ at the point $x = 0$. See Figure 1 as an illustration.

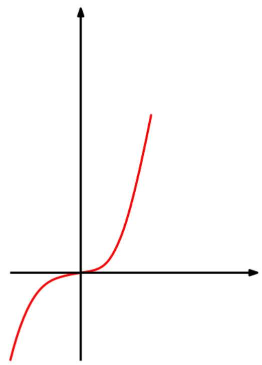

Figure 1: $y = x^3$. At point $x = 0$, $y' = y'' = 0$ but it is not a local minimizer.

Now we introduce the general stationary point.

Definition 4. *Given a function $f : \mathbb{R}^d \to \mathbb{R}$, assume that f is ρ-Hessian Lipschitz ($\nabla^2 f$ is ρ-Lipschitz). We say x is a ε-second order stationary point of f if*

$$\|\nabla f(x)\| \le \varepsilon, \ \lambda_{\min}(\nabla^2 f(x)) \ge -\sqrt{\rho\varepsilon}.$$

Remark 2. The conditions means x is approximately stationary and its Hessian is approximately positive semidefinite.

3 Noisy Gradient Descent

Unfortunately, it's not guaranteed that we can always escape any arbitrary saddle point. In fact, we will show, given $f : \mathbb{R}^d \to \mathbb{R}$, assume that it is ℓ-smooth and ρ-Hessian Lipschitz. Then the noisy gradient flow/descent can find an ε-second order stationary point efficiently (in time $O(\text{poly}\,(d))$).

3.1 A quadratic case

Consider $f(x) = x^\top A x$ where $A = \text{diag}\,(-1, 1, \ldots, 1) \in \mathbb{R}^{d \times d}$. It is trivial to see $\hat{x} = 0$ is a saddle point. We do gradient flow of f:

$$\frac{d}{dt} x_t = -\nabla f(x_t) = -2Ax_t.$$

Since A is a diagonal matrix, for each $k \in [d]$,

$$\frac{d}{dt}(x_t)_d = -2A_{k,k}(x_t)_d.$$

Solving these systems, we obtain

$$(x_t)_k = (x_0)_k \exp\left(-2A_{k,k}t\right).$$

This means as $t \to \infty$, $(x_t)_k \to 0$ for $k \neq 1$ and $|(x_t)_1| \to \infty$. This means as long as $|(x_0)_1| \geq 1/\mathrm{poly}(d)$, $|(x_t)_1|$ will become $\Omega(1)$ in $O(\log d)$ time.

3.2 General loss function

Now we illustrate the high-level idea to show how to escape the saddle point. Due to Lemma 1, to escape the saddle point, it might be possible to add a small perturbation ball if necessary. The algorithm can be roughly stated as

Algorithm 1: noisy gradient descent

for $t = 0, 1, \dots$ **do**
  if *some perturbation condition holds* **then**
    $\xi_t \sim \mathrm{Unif}(r\mathbb{B}^d)$;
    $x_t \leftarrow x_t + \xi_t$;
  $x_{t+1} = x_t - \eta\nabla f(x_t)$;

Since the analysis is quite technical, we will present the high-level idea here (for precise and detailed analysis, refer [JGN+17]). Suppose that x_0 is near a saddle point with at least one descent direction (to ensure that it is possible to escape). Assume that, $\nabla f(x_t)$ is small for all $t \in [0, T]$. Then x_t is near x_0. Since f is Hessian Lipschitz, we know

$$\nabla^2 f(x_t) \approx \nabla^2 f(x_0).$$

Thus we compute

$$\begin{aligned}\frac{d}{dt}\|\nabla f(x_t)\|^2 &= 2\left\langle \nabla f(x_t), \frac{d}{dt}\nabla f(x_t)\right\rangle \\ &= 2\left\langle \nabla f(x_t), \nabla^2 f(x_t)\dot{x_t}\right\rangle \\ &= -2\left\langle \nabla f(x_t), \nabla^2 f(x_t)\nabla f(x_t)\right\rangle \\ &\approx -2\left\langle \nabla f(x_t), \nabla^2 f(x_0)\nabla f(x_t)\right\rangle.\end{aligned}$$

This means $\nabla f(x_t)$ will blow up along the descent direction. Thus we know it will escape the saddle point (under some regularity conditions).

Theorem 5 (Theorem 2 in [JGN+17])**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be an ℓ-smooth, ρ-Hessian Lipschitz function. For all $\eta < \ell^2/\rho$, $\delta \in (0, 1)$, with probability at least $1 - \delta$, noisy gradient descent can output an ε-second order stationary point with*

$$O\left(\frac{\ell(f(x_0) - f_*)}{\varepsilon^2}\log^4\left(\frac{d\ell(f(x_0) - f_*)}{\eta^2\delta}\right)\right)$$

iterations.

References

[JGN+17] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. How to Escape Saddle Points Efficiently, 2017.