S&DS 671 (Spring 2019) Selected Topics in Neural Nets **Lecture: 12**

First-order Methods Almost Always Avoid Saddle Points

Lecturer: Harrison Zhou *Scribe: Cynthia Yue, Apr 10, 2019*

1 Introduction

In their 2017 paper *First-order Methods Almost Always Avoid Saddle Points*, Lee et al. demonstrate that first-order methods, such as gradient descent, proximal point algorithm, and coordinate descent, almost always avoid saddle points. They show this by applying the stable manifold theorem to this problem.

The paper discusses several first-order methods, and we will focus on the following:

1. Gradient descent
2. Proximal point algorithm
3. Coordinate descent

Let us first define a strict saddle point:

Definition. Strict saddle points.

1. $\nabla f(x^*) = 0$.
2. $\lambda_{\min}(\nabla^2 f(x^*)) < 0$: There exists an eigenvector v such that $Hv = \lambda_{\min} v$, and f is decreasing in the direction of v at the point x^*. If $\lambda_{\min} < 0$, then in the direction of the eigenvector v of $\lambda_{\min}$, f is decaying/decreasing.

$\mathcal{X}^*$: set of strict saddle points

In this particular paper, we are also focused on iterative algorithms. These algorithms start at an initial point x_0, and the iterates are generated as follows:

$$x_{k+1} = g(x_k) = g^2(x_{k-1}) = \ldots = g^{k+1}(x_0).$$

The main result of the paper emerges from looking at the convergence of these algorithms under some conditions:

$$m(x_0 : \lim_{k\to\infty} g^k(x_0) \in \mathcal{X}^*) = 0.$$

In other words, the the length of the set of starting points such that the algorithm converges to strict saddle points is 0, or the collection of points that the algorithm takes to saddle points is negligible.

Let us first study the intuition behind the main result. Let us consider some examples of first-order methods that do not converge to saddle points.

Example 1. We consider the case of a non-convex quadratic:

$$f(x) = \frac{1}{2}x^\top Hx \qquad \text{where} \quad x \in \mathbb{R}^n.$$

$$H = \begin{pmatrix} \lambda_1 & & & & & & \\ & \ddots & & & & & \\ & & \lambda_m & & & & \\ & & & \lambda_{m+1} & & & \\ & & & & \ddots & & \\ & & & & & \lambda_n \end{pmatrix} \quad \text{where} \quad \lambda_1, \ldots, \lambda_m > 0 \quad \text{and} \quad \lambda_{m+1}, \ldots, \lambda_n < 0.$$

Are there critical points?

$$\nabla f = 0 \Rightarrow Hx = 0 \Rightarrow x = 0.$$

What about saddle points?

$$\nabla^2 f = H\lambda_{\min}(\nabla^2 f(0)) < 0.$$

Let us also define the following:

$$\begin{aligned} L &= \sup |\lambda_i(H)| \\ &= \|H\|. \end{aligned}$$

Here, we see that the point at $x = 0$ satisfies the properties of a strict saddle point. Let us now consider gradient descent and see how it converges.

Gradient Descent

With gradient descent, we have the following iterative process:

$$\begin{aligned} x_{k+1} &= g(x_k) \\ &= x_k - \alpha \nabla f(x_k) && \text{where} \quad 0 < \alpha < \frac{1}{L}, \quad \text{so} \quad |\alpha\lambda_i| < 1 \\ &= x_k - \alpha H x_k \\ &= (I - \alpha H)x_k. \\ x_k &= (I - \alpha H)^k x_0 \\ &= \begin{pmatrix} (I-\alpha\lambda_1)^k & & & & & & \\ & \ddots & & & & & \\ & & (I-\alpha\lambda_m)^k & & & & \\ & & & (I-\alpha\lambda_{m+1})^k & & & \\ & & & & \ddots & & \\ & & & & & (I-\alpha\lambda_n)^k \end{pmatrix}. \end{aligned}$$

Now, let us ask the question, for what x_0 do we have $\lim_{k\to\infty} g^k(x_0) = 0$? In other words, what initial point leads to the saddle point $x = 0$? We find the initial point to be:

$$x_0 = \begin{pmatrix} x_0 1 \\ \vdots \\ x_0 m \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Here, we see that the measure is 0 because a set of dimension m in a space of dimension n, where $n > m$, has measure 0. Hence, the collection of points that gradient descent takes this to saddle points is negligible; essentially, we almost always avoid saddle points here using gradient descent.

Example 2. We consider the case of a non-quadratic function:

$$f(x, y) = \frac{1}{2}x^2 + \frac{1}{4}y^4 - \frac{1}{2}y^2$$

Are there critical points?

$$\nabla f = \begin{pmatrix} x \\ y^3 - y \end{pmatrix} = 0.$$
$$\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ -1 \end{pmatrix}$$

What about saddle points?

$$\nabla^2 f = \begin{pmatrix} 1 & 0 \\ 0 & 3y^2 - 1 \end{pmatrix} \Rightarrow \text{If the Hessian is diagonal, then the values on the diagonal are eigenvalues.}$$

Saddle point is at $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$.

Gradient Descent

In this example, we have the following iterative process for gradient descent:

$$\begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix} = g\Big(\begin{pmatrix} x_k \\ y_k \end{pmatrix}\Big) = \begin{pmatrix} x_k \\ y_k \end{pmatrix} - \alpha \begin{pmatrix} x_k \\ y_k^3 - y_k \end{pmatrix}.$$

Now, we ask the same question, for what $\begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$ do we have $\lim_{k\to\infty} g^k \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$? What initial point leads to the saddle point? We find the initial point to be:

$$\begin{pmatrix} x_0 \\ 0 \end{pmatrix}.$$

Once again, we see that the measure is 0 because we have a set of dimension 1 in a space of dimension 2, so we have measure 0, demonstrating we almost always avoid saddle points in this example using gradient descent.

We will now consider stable manifold theorem and its relevance to optimization with regards to saddle points.

2 Stable Manifold Theorem

Theorem 2.1. (Stable Manifold Theorem)

$$x_{k+1} = g(x_k)$$
$$A_g^* = \{x : g(x) = x,\ \max|\lambda_i(\mathrm{D}g)| > 1\}$$

If $|\det(\mathrm{D}g)| \neq 0$ for all x, then

$$m(x_0 : \lim_{k\to\infty} g^k(x_0) \in A_g^*) = 0.$$

Corollary 2.2. If $\mathcal{X}^* \subset A_g^*$, then

$$m(x_0 : \lim_{k\to\infty} g^k(x_0) \in \mathcal{X}*) = 0.$$

By considering the assumptions needed for stable manifold theorem and its corollary, we can study methods used for optimization and their performance with regards to saddle points by specifically checking if the following properties hold:

1. $|\det(\mathrm{D}g)| \neq 0$
2. $g(x) = x$
3. $\max|\lambda_i(\mathrm{D}g)| > 1$

3 Applications to Optimization with Saddle Points

We will study the applications of stable manifold theorem to optimization with regards to saddle points. We particularly study the applications of stable manifold theorem to gradient descent, proximal point algorithm, and coordinate descent.

3.1 Gradient Descent

With gradient descent, we have $f(x) : x \in \mathbb{R}^n$ and the assumption $\sup_x \|\nabla^2 f\| \leq L$.

The gradient descent algorithm is:

$$x_{k+1} = x_k - \alpha \nabla f(x_k) \quad \text{where} \quad 0 < \alpha < \frac{1}{L}.$$

We will prove that gradient descent has the properties required for stable manifold theorem.

Proof. (i) In order to check if gradient descent has the property $|\det(\mathrm{D}g)| \neq 0$, we see:

$$\begin{aligned} \det(\mathrm{D}g(\lambda)) &= I - \alpha \nabla^2 f(x) \\ &= \prod_{i=1}^{n} (1 - \alpha \lambda_i) > 0. \end{aligned}$$

(ii) In order to show $\mathcal{X}^* \subset A_g^*$, we see:

$$\begin{aligned} g(x^*) &= x^* \\ \sup_i |\lambda_i(I - \alpha \nabla^2 f(x^*))| &> 1. \end{aligned}$$

□

Example 3. $Y_i, \ldots, Y_n$ i.i.d.

$$\frac{1}{2}\mathcal{N}(-\theta, I) + \frac{1}{2}\mathcal{N}(\theta, I)$$

In this example, we are studying the expectation-maximization algorithm, estimating θ.

$$\begin{aligned}\theta_{k+1} &= \frac{1}{n}\sum_{i=1}^{n} Y_i \tanh(\theta_k^\top Y_i)\\ \theta_{k+1} &= \mathbb{E}Y \tanh(\theta_k^\top Y)\\ &= \theta_k + \nabla l(\theta_k).\end{aligned}$$

3.2 Proximal Point Algorithm

The proximal point algorithm has the following iteration:

$$\begin{aligned}x_{k+1} &= g(x_k)\\ &= \arg\min_z f(z) + \frac{1}{2\alpha}\|x_k - z\|^2 \quad \text{where} \quad \alpha < \frac{1}{L}.\end{aligned}$$

One of the assumptions for proximal point algorithm is $\sup_x \|\nabla^2 f(x)\| \le L$.

We will prove that proximal point algorithm has the properties required for stable manifold theorem.

Proof.

$$g(x) = \arg\min_z f(z) + \frac{1}{2\alpha}\|x - z\|^2.$$

(i) In order to show $\det|\mathrm{D}g(x)| \neq 0$, we need to find the Jacobian of g.

$$\begin{aligned}\nabla f(z) - \frac{1}{\alpha}(x - z) &= 0\\ \nabla f(g(x)) - \frac{1}{\alpha}(x - g(x)) &= 0\\ \alpha\nabla f(g(x)) - (x - g(x)) &= 0\\ \alpha\nabla f(g(x)) + g(x) &= x = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}\\ \alpha\nabla^2 f(g(x))\mathrm{D}g(x) + \mathrm{D}g(x) &= \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix}\\ (\alpha\nabla^2 f(g(x)) + I)\mathrm{D}g(x) &= I\\ \det(\mathrm{D}g(x)) &\neq 0.\end{aligned}$$

(ii)

$$\begin{aligned}\nabla f(z) - \frac{1}{\alpha}(x^* - z) &= 0\\ g(x^*) &= x^*.\end{aligned}$$

(iii)

$$\begin{aligned}\mathrm{D}g(x^*) &= (I + \alpha\nabla^2 f(g(x^*)))^{-1}\\ \max_i |\lambda_i(\mathrm{D}g(x))| &> 1.\end{aligned}$$

□

3.3 Coordinate Descent

With coordinate descent, the idea is that each iteration of the algorithm updates a single coordinate. The iteration for coordinate descent corresponds to the following update:

$$\begin{aligned} x_0 : x_0^{(1)} &= x_0 - \alpha(e_1^\top \nabla f(x_0) e_1, 0, \ldots, 0) \\ x_0^{(2)} &= x_0^{(1)} - \alpha(0, e_2^\top \nabla f(x_0^{(1)}) e_2, 0, \ldots, 0) \\ &\vdots \\ x_1 &= x_0^{(n)} \\ \text{so } x_{k+1} &= g(x_k) \\ &= g_n \circ g_{n-1} \circ \ldots \circ g_1(x_k). \end{aligned}$$

One of the assumptions for coordinate descent is $\sup_{i,x} |e_i^\top \nabla^2 f(x) e_i| \leq L$ where $0 < \alpha < \frac{1}{L}$.

We will prove that coordinate descent has the properties required for stable manifold theorem.

Proof. (i) We first want to check if coordinate descent has the property $|\det(\mathrm{D}g(x))| \neq 0$. In order to do so, we first know that as long as the determinant of each function $\neq 0$, then $|\det(\mathrm{D}g(x))| \neq 0$. If we consider the first coordinate $x_0^{(1)}$, we take the derivative in terms of x_0 to get $\mathrm{D}g_1(x) = (I - \alpha e_1 e_1^\top \nabla^2 f(x))$, which results in $|\det(\mathrm{D}g(x))| \neq 0$.

(ii) For x^* is a strict saddle point, we can see that

$$g(x^*) = x^*.$$

(iii)

$$\begin{aligned} H &= \nabla^2 f(x^*) \\ J &= \mathrm{D}g(x^*) \\ &= \prod_{j=1}^{n} (I - \alpha e_{n-j+1} e_{n-j+1}^\top H). \end{aligned}$$

Let y_0 be the eigenvector corresponding to the smallest eigenvalue of H, which is $-\lambda_0$.

Define $y_t = J^t y_0$ for $t \in N$.

We will show that

$$\begin{aligned} &\|J\|^t \|y_0\|_2 \geq \|J^t y_0\|_2 \geq C(1+\epsilon)^t \quad \text{where} \quad \epsilon > 0 \\ &\Rightarrow \|J\|(\|y_0\|)^{\frac{1}{t}} \geq C^{\frac{1}{t}}(1+\epsilon). \end{aligned}$$

We first claim:

$$y_{t+1}^\top H y_{t+1} \leq (1+4\epsilon) y_t^\top H y_t.$$

Under this claim, we have:

$$\begin{aligned}
-\lambda_0\|y_t\|^2 \leq y_{t+1}^\top H y_{t+1} &\leq (1+4\epsilon) y_t \top H y_t && \text{for all} && t \\
-\lambda_0\|y_t\|^2 \leq y_t^\top H y_t &\leq (1+4\epsilon)^t y_0^\top H y_0 \\
&\leq (1+4\epsilon)^t(-\lambda_0\|y_0\|^2) && \text{since } Hy_0 = -\lambda_0 y_0 \\
\text{so } \|y_t\|^2 &\geq (1+4\epsilon)^t\|y_0\|^2) \\
\|y_t\| &\geq \{(1+4\epsilon)^{\frac{1}{2}}\}^t\|y_0\| \\
&\geq (1+\epsilon)^t\|y_0\| && \text{Let us call } \|y_0\| \ C. \\
\|J^t y_0\| &= (1+\epsilon)^t C.
\end{aligned}$$

As $\|y_0\| = C$, we have $\|J\| \geq (1+\epsilon)$.

$\|J\|$ is the absolute value of the product of eigenvalues. As the product > 1, there exists an eigenvalue > 1.

□

4 Time Complexity

Although first-order methods almost always avoid saddle points, we find that gradient descent can take exponential time to escape. In the later 2017 paper *Gradient Descent Can Take Exponential Time to Escape Saddle Points*, Du et al. demonstrates this for both "un-natural" and "natural" initialization schemes as well as non-pathological functions.

They introduce the following theorem:

Theorem 4.1. Suppose the initialization point is uniformly sampled from $[-1,1]^d$. There exists a function f defined on $\mathbb{R}^d$ that is B-bounded, l-gradient Lipschitz and p-Hessian Lipschitz with parameters B, l, p at most poly(d) such that:

1. with probability one, gradient descent with step size $\eta \leq \frac{1}{l}$ will be $\Omega(1)$ distance away from any local minima for any $T \leq e^{\Omega(d)}$.
2. for any $\epsilon > 0$, with probability $1 - e^{-d}$, perturbed gradient descent will find a point x such that $\|x - x^*\|_2 \leq \epsilon$ for some local minimum x^* in poly$(d, \frac{1}{\epsilon})$ iterations.

For the proof of this result, please refer to [1, Theorem 4.1].

References

[1] Simon S. Du, Chi Jin, Jason D. Lee, Michael I. Jordan, Barnabás Póczos, and Aarti Singh. Gradient descent can take exponential time to escape saddle points, 2017.

[2] Jason D. Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. First-order methods almost always avoid saddle points, 2017.