1 Outline

In this lecture, we study

- complexity of nonconvex optimization,
- finding stationary points,
- second-order stationary points.

2 Introduction to Nonconvex Optimization

Figure 9.1 shows a nonconvex function with two variables. As we can see, the function has one

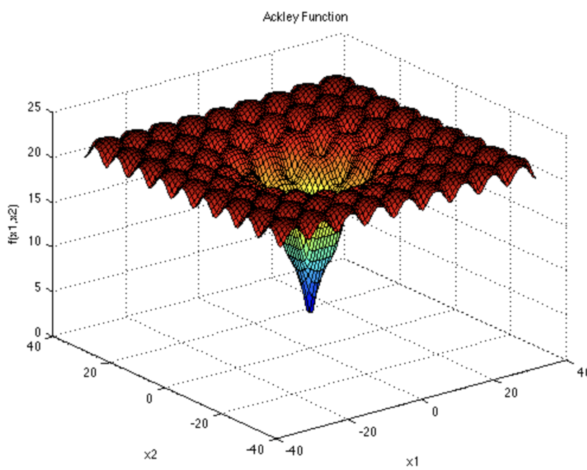

Figure 9.1: Ackley Function with Two Variables

global minimum but it has multiple local minima. Recall that for a convex function f, if $\nabla f(x) = 0$, then the corresponding x is a minimizer of the function f. Gradient-based methods for convex optimization basically seek for a solution x with $\nabla f(x) = 0$. However, for a nonconvex function f, a point x with $\nabla f(x) = 0$ does not guarantee global optimality as it can be a locally optimal solution.

In fact, finding the global minimum of a nonconvex function is a difficult task in general. To be more precise, the following theorem shows that there exists a smooth nonconvex function, to find the global minimum of which we need an exponential number of function evaluations.

Theorem 9.1. *For any $\beta > 0$, there exists a β-smooth function $f : [0,1]^d \to \mathbb{R}$ on $[0,1]^d$ such that any algorithm requires at least $(\beta/\epsilon)^{\Omega(d)}$ function queries to find an ϵ-optimal solution x with*

$$f(x) \leq \min_{x\in\mathbb{R}^d} f(x) + \epsilon.$$

Proof. We partition $[0,1]$ into k invervals of equal length, which gives rise to k^d boxes partitioning $[0,1]^d$. We can construct a function f such that a box contains a point x^* with $f(x^*) = -\epsilon$ but f has value 0 in the other boxes. This means that checking a box not containing x^* does not provide any information about the location of x^*. This means that to find the box containing x^*, we need at least $\Omega(k^d)$ function evalutions. To make function f smooth with parameter β, we can make f behaves like

$$f(x) \simeq f(x^*) + \frac{\beta}{2}\|x - x^*\|_2^2.$$

At the same time, to impose the condition that $f(x) = 0$ if x is not contained in the box with x^*, we can set $k = O(\sqrt{\beta/\epsilon})$. Therefore, we need at least $O\left((\sqrt{\beta/\epsilon})^d\right)$ function evaluations. □

3 Nonconvex Function Landscape

Recall that a point $x \in \mathbb{R}^d$ is a **local minimum** for a function f if there is some $\delta > 0$ such that

$$f(x) \leq f(y) \quad \text{for all } y \in \mathbb{R}^d \text{ with } \|x - y\| \leq \delta.$$

A global minimum of f over $\mathbb{R}^d$ is a point x^* with

$$f(x^*) \leq f(y) \quad \text{for all } y \in \mathbb{R}^d.$$

We learned that for a convex function, a local minimum is a global minimum. However, for a nonconvex function, we may have local minima that are not a global minimum as depicted in Figure 9.2.

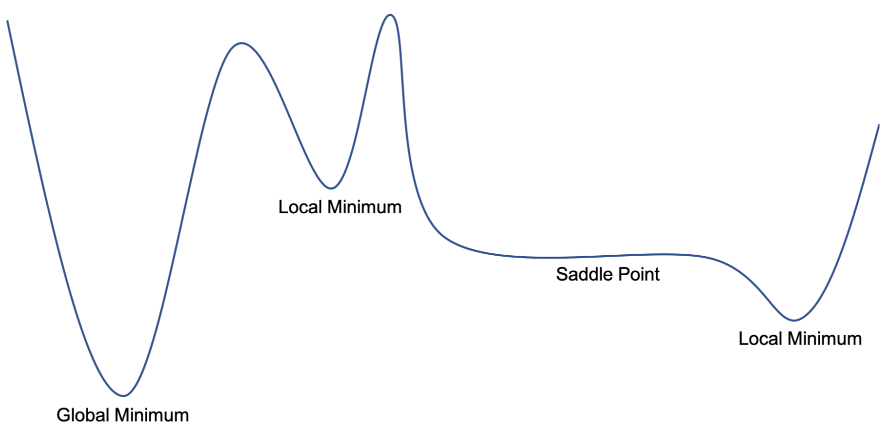

Figure 9.2: Landscape of a nonconvex function

For a differentiable function f, we say that a point $x \in \mathbb{R}^d$ is a **stationary point** if $\nabla f(x) = 0$. For a convex function, it follows from the optimality condition that a stationary point is a global minimum. For a nonconvex function, we can argue that a local minimum is a stationary point. However, there exists a stationary point that is not a local minimum. In Figure 9.2, the function landscape has a flat region where the gradient is zero that does not correspond to a local minimum. We refer to such a stationary point as a **saddle point**. In general, saddle points can form a very large flat region, and it is often hard for a basic implementation of gradient descent to escape from a saddle point.

The following provides necessary conditions for a local minimum.

Proposition 9.2. *Let x be a local minimum of a function $f : \mathbb{R}^d \to \mathbb{R}$ over $\mathbb{R}^d$. Then $\nabla f(x) = 0$, which means that x is a stationary point. Moreover, if f is twice continuously differentiable, then $\nabla^2 f(x) \succeq 0$.*

In fact, the condition that $\nabla f(x) = 0$ and $\nabla^2 f(x) \succeq 0$ is not sufficient to guarantee that x is a local minimum. For example, one may consider $f(x) = x^3$ with $\nabla f(x) = 3x^2$ and $\nabla^2 f(x) = 6x$. Then we have $\nabla f(0) = 0$ and $\nabla f^2(0) = 0$, but $x = 0$ is not a local minimum of $f(x) = x^3$ as depicted in Figure 9.3. The following proposition provides a sufficient condition for a local minimum.

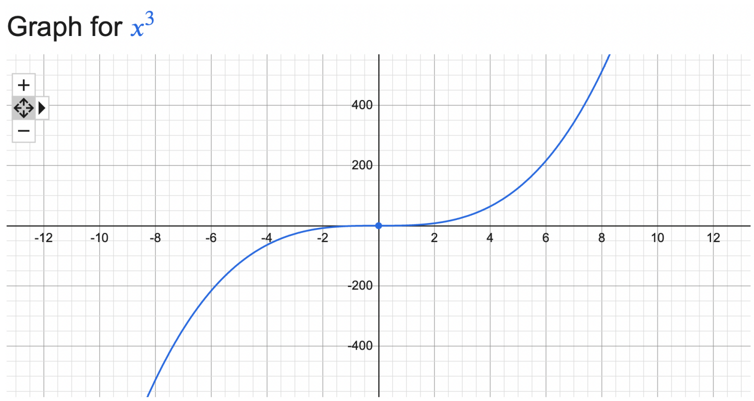

Figure 9.3: A cubic function $f(x) = x^3$

Proposition 9.3. *Suppose that a function $f : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable over $\mathbb{R}^d$. If $\nabla f(x) = 0$ and $\nabla^2 f(x) \succ 0$, then x is a local minimum.*

Although a stationary point with a positive definite Hessian is a local minimum, we should note that there exists a local minimum at which the Hessian is not positive definite but positive semidefinite.

4 Finding Stationary Points

We saw that finding a global minimum of a nonconvex function can take an exponential number of iterations even if the function is smooth (see Figure 9.4). In a high-dimensional space, finding

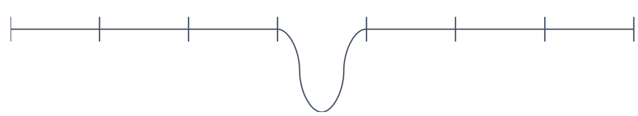

Figure 9.4: Hard nonconvex optimization instance

a local minimum can also be difficult. Then as a first step, one may attempt to find a stationary point. We define an **ϵ-stationary point** as a point $x \in \mathbb{R}^d$ with

$$\|\nabla f(x)\|_2 \leq \epsilon.$$

Next, we argue that gradient descent can find an ϵ-stationary point for a smooth function.

Theorem 9.4. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a β-smooth function in the ℓ_2-norm. Let x_{t+1} denote the solution generated by gradient descent with step size $\eta = 1/\beta$ after t iterations. Then for any*

$$t \geq \frac{2\beta(f(x_1) - f(x^*))}{\epsilon^2},$$

we have
$$\min\{\|\nabla f(x_s)\|_2 : s=1,\dots,t+1\} \le \epsilon.$$

Proof. Since f is β-smooth, it follows that

$$f(x_{t+1}) \le f(x_t) + \nabla f(x_t)^\top (x_{t+1}-x_t) + \frac{\beta}{2}\|x_{t+1}-x_t\|_2^2.$$

As $x_{t+1} = x_t - \eta \nabla f(x_t)$, the inequality is equivalent to

$$f(x_{t+1}) \le f(x_t) - \frac{1}{2\beta}\|\nabla f(x_t)\|_2^2.$$

If $\|\nabla f(x_t)\|_2 > \epsilon$, then
$$f(x_{t+1}) \le f(x_t) - \frac{1}{2\beta}\epsilon^2.$$

This means that the number of time steps t with $\|\nabla f(x_t)\|_2 > \epsilon$ is at most

$$\frac{2\beta(f(x_1)-f(x^*))}{\epsilon^2}.$$

Therefore, if
$$t \ge \frac{2\beta(f(x_1)-f(x^*))}{\epsilon^2},$$

we have
$$\min\{\|\nabla f(x_s)\|_2 : s=1,\dots,t+1\} \le \epsilon,$$

as required. □

In fact, we can also show that stochastic gradient descent finds a ϵ-stationary point for a smooth function.

Theorem 9.5. *Let $f:\mathbb{R}^d \to \mathbb{R}$ be a β-smooth function in the ℓ_2-norm. Let $x_1,\dots,x_{T+1}$ be the iterates generated by stochastic gradient descent with a constant step size $\eta = 1/\sqrt{T}$. Assume that the stochastic gradient g_t at time step t satisfies $\|g_t\|_2 \le L$. Then*

$$\min\left\{\mathbb{E}\left[\|\nabla f(x_t)\|_2^2\right] :\ t=1,\dots,T\right\} \le \frac{1}{\sqrt{T}}\left(f(x_1)-f(x^*)+\frac{\beta L^2}{2}\right)$$

where $x^* \in \mathrm{argmin}_{x\in\mathbb{R}^d} f(x)$.

Proof. By smoothness of f, we have

$$\begin{aligned} f(x_{t+1}) - f(x_t) &\le \nabla f(x_t)^\top (x_{t+1}-x_t) + \frac{\beta}{2}\|x_{t+1}-x_t\|_2^2 \\ &= -\eta \nabla f(x_t)^\top g_t + \frac{\beta\eta^2}{2}\|g_t\|_2^2 \\ &\le -\eta \nabla f(x_t)^\top g_t + \frac{\beta\eta^2 L^2}{2}. \end{aligned}$$

Taking the expectation conditioned on x_t, we obtain

$$\mathbb{E}\left[f(x_{t+1}) \mid x_t\right] - f(x_t) \le -\eta\|\nabla f(x_t)\|_2^2 + \frac{\beta\eta^2 L^2}{2}.$$

This implies that

$$\mathbb{E}\left[\|\nabla f(x_t)\|_2^2\right] \leq \frac{1}{\eta}\left(\mathbb{E}\left[f(x_t)\right] - \mathbb{E}\left[f(x_{t+1})\right]\right) + \frac{\beta\eta L^2}{2}.$$

Then it follows that

$$\begin{aligned}
\min\left\{\mathbb{E}\left[\|\nabla f(x_t)\|_2^2\right] :\ t = 1, \ldots, T\right\} &\leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla f(x_t)\|_2^2\right] \\
&\leq \frac{1}{T\eta}\left(f(x_1) - \min_{x\in\mathbb{R}^d} f(x)\right) + \frac{\beta\eta L^2}{2},
\end{aligned}$$

as required. □

5 Second-Order Stationary Points

In the previous section, we looked for an approximate stationary point that has a bounded norm. Remember that a local minimum not only is a stationary point but also has a positive semidefinite Hessian. Motivated by this, we consider methods for finding a point that approximately satisfies the necessary condition for a local minimum.

We say that a point x is an (ϵ, δ)**-SOSP** where an SOSP stands for a second-order stationary point if x satisfies

$$\|\nabla f(x)\|_2 \leq \epsilon \quad \text{and} \quad \nabla^2 f(x) \succeq -\delta I.$$

Here, $\nabla^2 f(x) \succeq -\delta I$ means that $\nabla^2 f(x) + \delta I \succeq 0$ which states that $\nabla^2 f(x) + \delta I$ is positive semidefinite. We will show an algorithm that computes an (ϵ, δ)**-SOSP** under some smoothness assumptions. We say that the Hessian of f is γ-Lipshitz continuous if

$$\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \leq \gamma\|x - y\|_2$$

where $\|\nabla^2 f(x) - \nabla^2 f(y)\|_2$ denotes the **spectral norm** of $\nabla^2 f(x) - \nabla^2 f(y)$ and the spectral norm of a matrix is its largest singular value.

Lemma 9.6. *Suppose that f is twice continuously differentiable and γ-Lipshitz continuous, then for any $x, y \in \mathbb{R}^d$,*

$$\left|f(y) - \left(f(x) + \nabla f(x)^\top(y - x) + \frac{1}{2}(y - x)^\top\nabla^2 f(x)(y - x)\right)\right| \leq \frac{\gamma}{6}\|y - x\|_2^3.$$

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a β-smooth function whose Hessian is γ-Lipschitz continuous. Then we apply the following algorithm.

0. Initialize a point $x = x_1 \in \mathbb{R}^d$.
1. Repeat the following gradient descent update until $\|\nabla f(x)\|_2 \leq \epsilon$:

$$x \leftarrow x - \frac{1}{\beta}\nabla f(x).$$

2. If $\nabla^2 f(x) \succeq -\delta I$, then x is an (ϵ, δ)-SOSP, so return x.
3. Find a unit vector v such that $v^\top\nabla^2 f(x)v < -\delta$.

4. For a step size $\eta > 0$, we update x as follows.

$$x \leftarrow \begin{cases} x + \eta v, & \text{if } f(x+\eta v) \le f(x-\eta v) \\ x - \eta v, & \text{otherwise.} \end{cases}.$$

5. Go back to step 1.

Theorem 9.7. *Let $x_1 \in \mathbb{R}^d$ denote the initial point, and let $x^* \in argmin_{x\in\mathbb{R}^d} f(x)$. Then the above algorithm computes an (ϵ, δ)-SOSP after at most*

$$\frac{2\beta(f(x_1) - f(x^*))}{\epsilon^2}$$

gradient computations and at most

$$\frac{3\gamma^2(f(x_1) - f(x^*))}{\delta^3}$$

Hessian computations.

Proof. Note that we apply gradient descent only if $\|\nabla f(x)\|_2 > \epsilon$. Recall that applying the gradient descent update on x satisfies

$$f\left(x - \frac{1}{\beta}\nabla f(x)\right) \le f(x) - \frac{1}{2\beta}\|\nabla f(x)\|_2^2.$$

Hence, each step of applying gradient descent reduces the function value by at least $\epsilon^2/2\beta$. This means that the total number of gradient descent updates is at most

$$\frac{2\beta(f(x_1) - f(x^*))}{\epsilon^2}.$$

Moreover, note that we apply the Hessian gradient only if $\nabla^2 f(x) \prec -\delta I$, in which case there exists a unit vector v with $v^\top \nabla^2 f(x) v < -\delta$. Here,

$$\begin{aligned}
\min\{f(x+\eta v), f(x-\eta v)\} &= \frac{1}{2}\left(f(x+\eta v) + f(x-\eta v)\right) \\
&\le f(x) + \frac{\eta^2}{2} v^\top \nabla^2 f(x) v + \frac{\gamma\eta^3}{6}\|v\|_2^3 \\
&< f(x) - \frac{\eta^2}{2}\delta + \frac{\gamma\eta^3}{6} \\
&= f(x) - \frac{\delta^3}{3\gamma^2}.
\end{aligned}$$

Therefore, one iteration of the Hessian gradient step reduces the function value by at least $\delta^3/3\gamma^2$. Hence, the total number of Hessian gradient descent updates is at most

$$\frac{3\gamma^2(f(x_1) - f(x^*))}{\delta^3},$$

as required. □

References