Lecture #23: Nonconvex Optimization

Suppose we want to minimize a nonconvex function

$$f: \mathbb{R}^n \to \mathbb{R}$$

but we only want to reach a local minimum

def. x is a local minimum if for some $\varepsilon > 0$,

$$\|x - y\| \leq \varepsilon \Rightarrow f(y) \geq f(x)$$

Why would we be content with just a local minimum?

1. Minimizing a nonconvex function is NP-hard, so what else can we do?

2. the points gradient descent reaches in practice seem to be good enough

In fact the global minimum might not generalize well [Zhang et al]

Theorem [Murty, Kabadi] Deciding if a given point x is a local minimum of a nonconvex function is NP-hard

Why can't you just check the gradient and if the Hessian has negative eigenvalues?

The connection only goes in one direction:

Claim: If x is a local minimum and f is twice differentiable, then $\nabla f(x) = 0$ and $\lambda_{min}(\nabla^2 f(x)) \geq 0$

Proof: If either condition were violated, there would be a direction where the function decreases. ☐

The difficulty in checking if x is a local minimum is:

If $\nabla f(x) = 0$, $\nabla^2 f(x) = 0$ then checking next terms in Taylor expansion $\iff$ tensor optimization

But there is some hope:

Claim: If $\nabla f(x) = 0$ and $\lambda_{min}(\nabla^2 f(x)) > 0$ then x is a local minimum

Q: Can we impose conditions on f to make finding a local minimum easy?

Aside from local minima/maxima we can also have saddle points

def: x is a saddle point if $\nabla f(x) = 0$ s.t. for any $\varepsilon > 0$ there are y and z with

$$\|x-y\|, \|x-z\| \leq \varepsilon \text{ and } f(y) < f(x) < f(z)$$

E.g. $f(x_1, x_2) = x_1^2 - x_2^2$, then $x_1 = x_2 = 0$ is a saddle point

Claim: If $\nabla f(x) = 0$ and $\nabla^2 f(x)$ has positive and negative eigenvalues, x is a saddle point

In some cases, saddle points are inevitable

Imagine we have a function f where

(1) all local minima are global minima

(2) the set of all global minima are discrete and isolated (e.g. due to symmetries in parameterization)

Then we must have saddle points (if not, the function would be convex, which would contradict (2))

Q: Empirically, when gradient descent on a nonconvex function does not work well, is it because we got stuck in a bad local minimum or near a saddle point?

To tame these saddle points, let's assume

def: [Ge et al] A function f has the (θ, γ, δ) strict saddle property if for any x at least one of the following holds

(1) $\|\nabla f(x)\| \geq \theta$

(2) $\lambda_{min}(\nabla^2 f(x)) \leq -\gamma$

(3) x is δ-close to a local minimum

Under this condition, there are efficient algorithms for finding a point δ-close to a local minimum

Second order methods: (Trust Region) (Cubic Regularization)

Given a point x either

(1) $\nabla f(x)$ is large, can take a step $-\eta \nabla f(x)$

(2) $\lambda_{min}(\nabla^2 f(x))$ is negative, can take a step $+\eta y$ where y is the corresponding eigenvector

However computing and storing the Hessian might be too expensive

First Order Methods: Noisy gradient descent

$$x_{t+1} = x_t - \eta \nabla f(x_t) + \xi$$

(ξ: Gaussian noise)

Theorem [Ge et al] Noisy gradient descent can find a point δ-close to a local minimum in polynomial time

The number of iterations depends on $\theta, \gamma, \delta, B, \rho, \varepsilon, \log n$

(ρ: smoothness of Hessian; B: smoothness; ε: target accuracy)

[Lee et al] use the stable manifold theorem from dynamical systems, but no bound on convergence

Intuition: Low probability over random starting point that trajectory passes too close to a saddle point

There are some cases we can prove no bad local minima:

[Auffinger, Ben-Arous, Cerny]: spherical p-spin glass

e.g. random third order tensor T (Gaussian entries) yields distribution on S^{n-1} where

$$\Pr[X] = \frac{e^{\sum_{ijk} x_i y_j y_k T_{ijk}}}{Z}$$

(Z: partition function)

Through Kac-Rice theorem can calculate distribution of critical points and show

"all local minima are close in objective value to the global minimum"

[Choromanka et al]: deep network $\rightsquigarrow$ spherical p-spin glass

This "reduction" involves expanding each path into a separate random variable and treating them as independent

Note: In a spherical p-spin glass model the objective value of the global minimum is close to the objective value of any other point

If you spike it:

$$T + \lambda u \otimes u \otimes u$$

(u: unit vector)

(T: random, Gaussian entries)

then the optimization landscape is completely different

It now resembles [Frieze, Kannan]'s tensors for planted clique, which one believes to be hard to optimize

In another direction, for many matrix models we know characterizations like:

Matrix completion: with high probability the objective function

$n \times r$ $m \times r$

$$f(U,V) = \sum_{(i,j)\in\Omega} \| M_{ij} - (UV^T)_{ij} \|^2$$

has all local minima satisfy $UV^T = M$ and satisfies the strict saddle property

Also matrix sensing, robust PCA