Convergence of gradient descent in over-parameterized networks

September 2022 · 6 min · 1257 words · Mina Ghashami | Suggest Changes

Neural networks typically have very large number of parameters. Depending on whether they have more parameters than training instances, they are over-parameterized or under-parameterized. In either case, their loss function is a multivariable, multidimensional and often non-convex function. In this post, we study over-parameterized neural networks and their loss landscape; we answer the question of why gradient descent (GD) and its variants converge to global minima in over-parameterized neural networks, even though their loss function is non-convex.

Notation and problem definition

Let us consider a training dataset $D = \{x_i, y_i\}_{i=1}^n$ with $x_i \in \mathbb{R}^d$, $y \in \mathbb{R}$, and a parametric family of models $f(w; x)$ e.g. neural networks where $w \in \mathbb{R}^m$. We wish to find a model with parameter w^* such that it fits the training data, i.e.

$$f(w^*; x_i) \approx y_i, \forall i \in [1, n]$$

Since x is not a model parameter, we can compact this notation and write it as:

$$F(w) = y$$
$$F : \mathbb{R}^m \to \mathbb{R}^n$$
$$w \in \mathbb{R}^m, y \in \mathbb{R}^n$$

Where $(F(w))_i = F(w; x_i)$. To minimize this system and find optimal parameter w^* a certain loss function $L(w)$ is defined, e.g., the least square loss $L(w) = \frac{1}{2}||F(w) - y||^2$. Irrespective the choice of loss function, $L(w)$ will be non-convex as F in neural network contains non-linear activations and is non-convex.

For vectors, we use $\|.\|$ to denote Euclidean norm. For a function $F : \mathbb{R}^m \to \mathbb{R}^n$, we use $\nabla F(w)$ to denote gradient of F with $(\nabla F(w))_{i,j} = \frac{\partial F_i}{\partial w_j}$. We denote Hessian of F as $H_F \in \mathbb{R}^{n \times m \times m}$ with $(H_F)_{ijk} = \frac{\partial^2 F_i}{\partial w_j \partial w_k}$. We denote Hessian of the loss function as H_L which is an $m \times m$ matrix. We denote neural tangent kernel of F as $K(w) = K(w; x, x') = \nabla F(w; x) \nabla F(w; x')^T$.

We denote eigenvalues of a function/matrix f with $\lambda(f)$, where $\lambda_{max}(f), \lambda_{min}(f)$ are largest and smallest eigenvalues of f, respectively.

Solving this system in over-parameterized setting i.e. where $m > n$ is the focus of this post; we show exact solutions exist.

Loss landscape

We refer to the representation of loss function with respect to model parameters as loss landscape. We know loss landscape of neural networks is highly non-convex; but there is a key difference between loss landscape of under-parameterized networks and that of over-parameterized networks.

Under-parameterized NNs have a globally non-convex loss landscape, however they are locally convex in a sufficiently small neighborhood of local minima. On the other hand, loss landscape of over-parameterized systems is *essentially non-convex*, meaning that they are globally as well as locally non-convex; i.e. in any arbitrarily small neighborhood around global minimum they are non-convex. As the result, we can not use convexity based optimization theorems for analyzing over-parameterized NNs.

Non-isolated local/global minima

For under-parameterized systems, local minima are generally isolated, however in over-parameterized networks, there is no isolated local/global minima in the loss landscape. Chaoyue Liu et. al prove the phenomena in appendix A of their paper, but to see it intuitively consider the following example:

Example: Consider an over-parameterized system with two training instances $\{(x_1, y_1), (x_2, y_2)\}$ where $x_i \in \mathbb{R}^2$, $y_i \in \mathbb{R}$, and a parameter space of $w \in \mathbb{R}^3$. To minimize the training loss, we need to find $w^* = (w^0, w^1, w^2)$ such that the following two equations hold:

$$w^0 + w^1 x_1^1 + w^2 x_1^2 = y_1$$
$$w^0 + w^1 x_2^1 + w^2 x_2^2 = y_2$$

Above equations are planes in $\mathbb{R}^3$, and their intersection which is a line is the set of global minima w^*; since this set is a line it is a non-isolated manifold.

Non-convexity of manifold of global minima

We saw that w^*s are non-isolated and form a manifold, here we show the manifold of global minima is non-convex. Let M denote manifold of w^*, where $\forall w \in M, w$ is a global minima of $L(w)$. If the manifold was convex, then for two $w_1, w_2 \in M$, the in-between points $v = \alpha w_1 + (1-\alpha) w_2$ for $\alpha \in (0,1)$ will have a lower loss i.e. $L(v) \leq L(w_1)$. But we already know M is the minizer of loss function, so the manifold can not be convex.

Polyak-Lojasiewicz (PL) condition

In absence of convex based optimization methods for over-parameterized networks, a new framework based on Polyak-Lojasiewicz (PL) condition helps us to analyze them.

μ-PL condition. A non-negative loss function L satisfies μ-PL condition for $\mu > 0$ if,

$$\|\nabla L(w)\|^2 \geq \mu L(w), \forall w \in B$$

Where $B = B(w_0, r) = \{w | \ \|w - w_0\| \leq r\}$ is a neighborhood around initial value w_0 with radius r.

We will show that if μ-PL condition is satisfied, then global minima w^* exists and GD converges to w^* exponentially fast.

Convergence under PL condition

Specifically we show under three conditioning assumptions:

1. $L(w)$ is μ-PL for $\forall w \in B$
2. $L(w)$ is β-smooth, i.e. $\lambda_{max}(H) \leq \beta$
3. GD's learning rate is small enough i.e. $\eta = 1/\beta$

GD will converge to global minima with an exponentially fast convergence rate.

proof. Gradient descent's update rule is $w_{t+1} = w_t - \eta \nabla L(w_t)$. Using Taylor's expansion we will have:

$$
\begin{aligned}
L(w_{t+1}) &= L(w_t) + \underbrace{(w_{t+1} - w_t)^T}_{-\eta \nabla L(w_t)^T} \nabla L(w_t) + \frac{1}{2} \underbrace{(w_{t+1} - w_t)^T}_{-\eta \nabla L(w_t)^T} H(w_t) \underbrace{(w_{t+1} - w_t)}_{-\eta \nabla L(w_t)} \\
&\leq L(w_t) - \eta \|\nabla L(w_t)\|^2 + \frac{\eta^2}{2} \underbrace{\nabla L(w_t)^T H(w) \nabla L(w_t)}_{\leq \beta \|\nabla L(w_t)\|^2} \\
&\leq L(w_t) - \eta \|\nabla L(w_t)\|^2 (1 - \frac{\eta \beta}{2}) \\
&\leq L(w_t) - \frac{\eta}{2} \underbrace{\|\nabla L(w_t)\|^2}_{\geq \mu L(w_t)} \\
&\leq (1 - \eta \mu) L(w_t)
\end{aligned}
$$

Therefore after t steps of gradient descent, $L(w_{t+1}) \leq (1 - \eta\mu)^t L(w_0)$. This shows loss decays exponentially fast with the decay rate of $(1 - \eta\mu)$.

Over-parameterized NNs satisfy PL condition

While the second and third condition in convergence proof are commonly used in convex cases as well, it is not obvious why first condition holds for over-parameterized NNs. In this section, we intuitively show why $\mu - PL$ condition holds for these networks.

First we note that in over-parameterized NNs, the least eigenvalue of NTK is separated from zero, i.e. $\lambda_{min}(K(w)) > 0$. The reason for this is that

$$
\begin{aligned}
\text{rank}(K(w)) &= \text{rank}(\nabla F(w) \nabla F(w)^T) \\
&= \text{rank}(F(w)) \\
&= \min(m, n) \\
&= n
\end{aligned}
$$

And since $K(w) \in \mathbb{R}^{n \times n}$ it can not be degenerate i.e. does not have zero eigenvalues. So we can always assume $\lambda_{min}(K(w)) \geq \mu$ for a positive μ.

Lemma. If $\lambda_{min}(K(w)) \geq \mu$ for all $w \in B$ then, the square loss function $L(w) = \frac{1}{2}\|F(w) - y\|^2$ is μ-PL on $w \in B$.

proof.

$$\begin{aligned}\frac{1}{2}\|\nabla L(w)\|^2 &= \frac{1}{2}\|(F(w)-y)^T\nabla F(w)\|^2 \\ &= \frac{1}{2}(F(w)-y)^T\nabla F(w)\nabla F(w)^T(F(w)-y) \\ &= \frac{1}{2}(F(w)-y)^T K(w)(F(w)-y) \\ &\geq \frac{1}{2}\lambda_{min}(K(w))\|F(w)-y\|^2 \\ &= \lambda_{min}(K(w))L(w)\end{aligned}$$

and therefore it satisfies PL condition.

I would like to finish this post by noting that $\eta\mu$ in the convergence rate, is actually the inverse of condition number of the NN function:

$$\text{Condition number}(F) = K_F = \frac{\sup_B \lambda_{max}(H)}{\inf_B \lambda_{min}(K)}$$

Using notation of above section, we see that $K_F = \frac{\beta}{\mu} = \frac{1}{\eta\mu}$. The smaller condition number is the better it is as it leads to faster convergence of GD. This paper proves that overparamterization helps with condition number meaning as number of parameters $m \to \infty$, the condition number $K_F \to 1$.

References

1. Simon S. Du et. al GRADIENT DESCENT PROVABLY OPTIMIZES OVER-PARAMETERIZED NEURAL NETWORKS
2. Chaoyue Liu et. al Loss landscapes and optimization in over-parameterized non-linear systems and neural networks

Thank you

If you have any questions please reach out to me:

mina.ghashami@gmail.com

https://www.linkedin.com/in/minaghashami/

Follow me on medium for more content: https://medium.com/@mina.ghashami

« PREV

Transformer: Concept and code from scratch