Fact[section] Example[section]

The Convergence of Over-parametrized Linear Networks Optimized Via Gradient Descent

Ziqing Xu, Salma Tarmoun, Hancheng Min, Enrique Mallada, Rene Vidal

Johns Hopkins University

02.14.2023

Introduction

The empirical success of neural networks on various applications, such as natural language processing, computer vision and decision-making, has motivated significant research on theoretically understanding why neural networks work so well in practice.
Question: Why over-parametrized neural networks trained with gradient descent (GD) enjoy fast convergence even if their loss landscape is non-convex?

Related work and their limitations

- neural tangent kernel: large width, large initialization
- mean-field analysis: infinitesimal stepsize, exponentially large width w.r.t. time
- convergence of linear networks: infinitesimal stepsize, special initialization(balanced, spectral)

This work: finite width, finite stepsize and general initialization for linear networks

Problem setting in the square loss

- training data $X \in \mathbb{R}^{n\times d}, Y \in \mathbb{R}^{n\times m}$.
- weight: $W_1 \in \mathbb{R}^{d\times h}, W_2 \in \mathbb{R}^{h\times m}$.
- loss function: $L(t) = \frac{1}{2}\|Y - XW_1(t)W_2(t)\|_F^2$.

A more general setting,

$$\min_{W_1, W_2} L(W), W = W_1 W_2$$

where $L(W)$ satisfies K-smoothness and $\mu-$PL condition w.r.t. W.

Notation

- product: $W(t) = W_1(t)W_2(t)$
- imbalance: $D(t) = W_1(t)^T W_1(t) - W_2(t)W_2(t)^T$
- condition number of data matrix: $\kappa = \frac{\lambda_{\max}(XX^T)}{\lambda_{\min}(XX^T)}$
- gradient w.r.t. W:$\nabla\ell(W)$

Gradient flow and gradient descent

Gradient descent

$$\begin{aligned} W_1(t+1) &= W_1(t) - \eta \nabla_{W_1} L(t), \\ W_2(t+1) &= W_2(t) - \eta \nabla_{W_2} L(t), \end{aligned} \tag{1}$$

Gradient flow

$$\begin{pmatrix} \dot{W}_1 \\ \dot{W}_2 \end{pmatrix} = - \begin{pmatrix} \nabla_{W_1} L(W_1, W_2) \\ \nabla_{W_2} L(W_1, W_2) \end{pmatrix} = - \begin{pmatrix} \nabla \ell(W) W_2^\top \\ W_1^\top \nabla \ell(W) \end{pmatrix}, \tag{2}$$

where $\nabla \ell(W) = \nabla_W L(W)$.

Convergence under gradient flow

We define the following linear operator which is the gradient of over-parametrized model

$$\gamma(\nabla \ell(W); W_1, W_2) := \begin{pmatrix} \nabla \ell(W) W_2^\top \\ W_1^\top \nabla \ell(W) \end{pmatrix}, \tag{3}$$

Using γ, one can show that the evolution of loss under GF is

$$\begin{aligned} \dot{L}(W_1, W_2) &= \left\langle \frac{\partial L}{\partial W_1}(W_1, W_2), \dot{W}_1 \right\rangle + \left\langle \frac{\partial L}{\partial W_2}(W_1, W_2), \dot{W}_2 \right\rangle \\ &= - \ \langle \gamma(\nabla \ell(W); W_1, W_2), \gamma(\nabla \ell(W); W_1, W_2) \rangle \\ &= - \langle \nabla \ell(W), \gamma^* \circ \gamma(\nabla \ell(W); W_1, W_2) \rangle , \end{aligned} \tag{4}$$

Convergence under gradient flow

Therefore, the dynamics of loss are defined by the following positive semi-definite Hermitian linear operator on $\nabla\ell(W)$:

$$\begin{aligned}\tau(\nabla\ell(W); W_1, W_2) &:= \gamma^* \circ \gamma(\nabla\ell(W); W_1, W_2) \\ &= \nabla\ell(W)\ W_2^\top W_2 + W_1 W_1^\top \nabla\ell(W).\end{aligned} \tag{5}$$

Then, from equation 4 and the min-max principle of Hermitian operators, we have

$$\dot{L}(t) = -\langle \nabla\ell(t), \tau_t(\nabla\ell(t))\rangle \leq -\lambda_{\min}(\tau_t)\|\nabla\ell(t)\|_F^2 \leq -2\mu\lambda_{\min}(\tau_t)L(t), \tag{6}$$

Convergence under gradient flow

How to prove $\lambda_{\min}(\tau_t)$ has a uniform positive lower bound?
There exists an non-negative function $\alpha(D, \sigma_{\min}(W))$ that depends on imbalance and product, such that for all $t \geq 0$,

$$\begin{aligned}\lambda_{\min}(\tau_t) &\geq \alpha(D(t), \sigma_{\min}(W(t))) \\ &= \alpha(D(0), \sigma_{\min}(W(t))) \\ &= \alpha(D(0), \sigma_{\min}(W(0))).\end{aligned} \tag{7}$$

Toy example

Objective $L(w_1, w_2) = \frac{1}{2}(y - w_1 w_2)^2$ where $y, w_1, w_2 \in \mathbb{R}$. Using same derivations, we can show

$$\begin{aligned} \dot{L}(t) &\leq -\big(w_1(t)^2 + w_2(t)^2\big) L(t) \\ &= -\sqrt{\big(w_1(t)^2 - w_2(t)^2\big)^2 + 4\big(w_1(t) w_2(t)\big)^2}\, L(t) \\ &= -\sqrt{\big(w_1(0)^2 - w_2(0)^2\big)^2 + 4\big(w_1(t) w_2(t)\big)^2}\, L(t) \end{aligned} \tag{8}$$

Regarding the product, one can show

$$\begin{aligned} |w_1(t) w_2(t)| &\geq |y| - |y - w_1(t) w_2(t)| \\ &\geq |y| - |y - w_1(0) w_2(0)| \\ &= |y| - |L(0)| \end{aligned} \tag{9}$$

Thus, we have

$$\dot{L}(t) \leq -\sqrt{\big(w_1(0)^2 - w_2(0)^2\big)^2 + 4\big(|y| - |L(0)|\big)^2}\, L(t) \tag{10}$$

Difference Between Gradient Flow and Gradient Descent

When using gradient flow(GF), imbalance is invariant,

$$\dot{D}(t) = 0 \tag{11}$$

When using gradient descent(GD), imbalance changes at each iteration,

$$D(t+1) = D(t) + O(\eta^2) \tag{12}$$

Convergence of non-overparametrized model under GD

Notice that $\ell(t)$ is K-smooth and satisfies μ-PL condition, where $K = \sigma^2_{\max}(X), \mu = \sigma^2_{\min}(X)$. Then, the following smoothness inequality holds for any W, W^+:

$$\ell(W^+) \leq \ell(W) + \langle \nabla \ell(W), W^+ - W \rangle + \frac{K}{2} \|W^+ - W\|_F^2 \quad (13)$$

After substituting the GD update with fixed step size η

$$W(t+1) = W(t) - \eta \nabla \ell(t). \quad (14)$$

into the smoothness inequality in equation 13 we obtain

$$\begin{aligned} \ell(t+1) &\leq \ell(t) - \eta \|\nabla \ell(t)\|_F^2 + \frac{K}{2}\eta^2 \|\nabla \ell(t)\|_F^2 \\ &= \ell(t) - \eta \left(1 - K\frac{\eta}{2}\right) \|\nabla \ell(t)\|_F^2 \\ &\leq (1 - 2\eta\mu + K\mu\eta^2)\ell(t) \end{aligned} \quad (15)$$

if the step size satisfies $\eta < \frac{2}{K}$.

Convergence of overparametrized model under GD

The update of the product is

$$\begin{aligned} W(t+1) &= W_1(t+1)W_2(t+1) \\ &= \left(W_1(t) - \eta \nabla \ell(t) W_2(t)^\top\right)\left(W_2(t) - \eta W_1(t)^\top \nabla \ell(t)\right) \\ &= W(t) - \eta \tau_t(\nabla \ell(t)) + \eta^2 \nabla \ell(t) W(t)^\top \nabla \ell(t). \end{aligned} \tag{16}$$

Then, we plug in the update of the product in the smoothness inequality

$$\begin{aligned} \ell(t+1) \leq \ell(t) + &\langle \nabla \ell(t), W(t+1) - W(t) \rangle \\ &+ \frac{K}{2} \| W(t+1) - W(t) \|_F^2 \end{aligned} \tag{17}$$

Convergence of overparametrized model under GD

Lemma

If at the t-th iteration of GD applied to the over-parametrized loss L, the step size η satisfies

$$\begin{aligned}&\lambda_{\min}(\tau_t) - \eta\|\nabla \ell(t)\|_F \|W(t)\|_F \\ &- \frac{K\eta}{2}\left[\lambda_{\max}(\tau_t) + \eta\|\nabla \ell(t)\|_F \|W(t)\|_F\right]^2 \geq 0\,,\end{aligned} \tag{18}$$

then the following inequality holds

$$L(t+1) \leq \rho(\eta, t) L(t)\,, \tag{19}$$

where

$$\begin{aligned}\rho(\eta, t) = 1 &- 2\eta\mu\lambda_{\min}(\tau_t) + K\mu\eta^2\lambda^2_{\max}(\tau_t) \\ &+ 2\eta^2\mu\sigma_{\max}(W(t))\|\nabla \ell(t)\|_F \\ &+ 2\eta^3\mu K\lambda_{\max}(\tau_t)\sigma_{\max}(W(t))\|\nabla \ell(t)\|_F \\ &+ \eta^4\mu K\sigma^2_{\max}(W(t))\|\nabla \ell(t)\|^2_F.\end{aligned} \tag{20}$$

Comparison

The convergence rate of non-overparametrized model is

$$\rho(\eta, t) = 1 - 2\eta\mu + K\mu\eta^2 \tag{21}$$

The convergence rate of overparametrized model is

$$\begin{aligned}\rho(\eta, t) = 1 &- 2\eta\mu\lambda_{\min}(\tau_t) + K\mu\eta^2\lambda^2_{\max}(\tau_t) \\ &+ 2\eta^2\mu\sigma_{\max}(W(t))\|\nabla\ell(t)\|_F \\ &+ 2\eta^3\mu K\lambda_{\max}(\tau_t)\sigma_{\max}(W(t))\|\nabla\ell(t)\|_F \\ &+ \eta^4\mu K\sigma^2_{\max}(W(t))\|\nabla\ell(t)\|^2_F.\end{aligned} \tag{22}$$

Towards linear convergence

- spectral bound for τ_t and $W(t)$.

$$p_1 \leq \sigma_{\min}(W(t)) \leq \sigma_{\max}(W(t)) \leq p_2$$
$$\alpha(D(t), \sigma_{\min}(W(t))) \leq \lambda_{\min}(\tau_t) \leq \lambda_{\max}(\tau_t) \leq \beta(D(t), \sigma_{\max}(W(t)))$$

- control of imbalance: we show that if loss decreases linearly, then $\|D(t) - D(0)\|_F \sim O(\eta)$
- uniform bounds on τ_t: when η is small but not infinitesimal

$$c_1\alpha_0 \leq \lambda_{\min}(\tau_t) \leq \lambda_{\max}(\tau_t) \leq c_2\beta_0 \tag{23}$$

where $0 < c_1 < 1, c_2 > 1$.

Theorem: Uniform bound on τ and W

Assume $\alpha_0 > 0$, and choose $0 < c_1 < 1$, and $c_2 > 1$. Let $\eta_1^{\max}$ and $\eta_2^{\max}$ be, respectively, the unique positive roots of the following two polynomials in η

$$\begin{aligned} a_4(0)\eta^3 + a_3(0)\eta^2 + \big(a_2(0) + \frac{4c_2 L(0)\sigma_{\max}^2(X)}{c_2 - 1}\big)\eta &= a_1, \\ a_4(0)\eta^3 + a_3(0)\eta^2 + \big(a_2(0) + \frac{8c_2\beta_0 L(0)\sigma_{\max}^2(X)}{(1 - c_1)\alpha_0}\big)\eta &= a_1. \end{aligned} \tag{24}$$

Then, for any $0 < \eta \leq \eta_{\max} := \min\{\eta_1^{\max}, \eta_2^{\max}\}$, the following holds for all $t = 0, 1, \ldots$

$$\begin{aligned} c_1\alpha_0 \leq \lambda_{\min}(\tau_t) \leq \lambda_{\max}(\tau_t) \leq c_2\beta_0 \\ p_1 \leq \sigma_{\min}(W(t)) \leq \sigma_{\max}(W(t)) \leq p_2. \end{aligned} \tag{25}$$

where

$$\begin{aligned} a_1 &= 2(c_1\alpha_0)\sigma_{\min}^2(X), \\ a_2(t) &= 2\sqrt{2\kappa L(t)\sigma_{\min}^6(X)}p_2 + \kappa\sigma_{\min}^4(X)(c_2\beta_0)^2, \\ a_3(t) &= 2\sqrt{2\kappa^3 L(t)\sigma_{\min}^{10}(X)}c_2\beta_0 p_2, \\ a_4(t) &= 2\kappa^2\sigma_{\min}^6(X)p_2^2 L(t). \end{aligned} \tag{26}$$

Theorem (Convergence rate of gradient descent on two-layer linear networks)

Under the same assumptions, for any $0 < \eta \leq \eta_{\mathsf{max}} := \min\{\eta_1^{\mathsf{max}}, \eta_2^{\mathsf{max}}\}$*, the loss function under GD satisfies*

$$L(t+1) \leq f(\eta, t)L(t),$$

for $f(\eta, t) = 1 - a_1\eta + a_2(t)\eta^2 + a_3(t)\eta^3 + a_4(t)\eta^4$ *is the upper bound of* $\rho(t)$*, and with*

$$0 < f(\eta, t) \leq f(\eta, 0) < 1, \quad \forall t \geq 0. \tag{27}$$

Thus, the loss converges linearly, i.e.,

$$L(t) \leq \Pi_{k=0}^{t} f(\eta, k)\, L(0) \leq f(\eta, 0)^t L(0). \tag{28}$$

with rate given by $f(\eta, 0)$*.*

Gradient Descent with Adaptive Learning Rate

The descent lemma we have is the following,

$$L(t+1) \leq \{1 - a_1\eta + a_2(t)\eta^2 + a_3(t)\eta^3 + a_4(t)\eta^4\} L(t) := f(\eta, t) L(t), \tag{29}$$

where

$$\begin{aligned} a_1 &= 2(c_1\alpha_0)\sigma_{\min}^2(X), \\ a_2(t) &= 2\sqrt{2\kappa L(t)\sigma_{\min}^6(X)}p_2 + \kappa\sigma_{\min}^4(X)(c_2\beta_0)^2, \\ a_3(t) &= 2\sqrt{2\kappa^3 L(t)\sigma_{\min}^{10}(X)}c_2\beta_0 p_2, \\ a_4(t) &= 2\kappa^2\sigma_{\min}^6(X)p_2^2 L(t). \end{aligned} \tag{30}$$

For each step, we can actually choose the learning rate which minimize the upper bound,

$$\eta_t^* = \arg\min_{\eta>0} f(\eta, t)$$

. Then, we get a sequence of learning rate $\{\eta_t^*\}_{t=1}^{\infty}$.

Asymptotic Convergence rate

The convergence rate we have is a fourth-order polynomial, it's hard to interpret. However, one observation is

$$\lim_{t\to\infty} a_3(t) = 0, \qquad \lim_{t\to\infty} a_4(t) = 0. \tag{31}$$

Thus, the rate becomes a quadratic term when $t \to \infty$,

$$f(\eta,\infty) = 1 - a_1\eta + a_2(\infty)\eta^2, \tag{32}$$

and

$$\min_{\eta} f(\eta,\infty) = 1 - \frac{\alpha^2 c_2^2}{\kappa\beta^2 c_1^2} \geq 1 - \frac{1}{\kappa}. \tag{33}$$

Current work: by studying the smoothness constant and PL constant w.r.t. (W_1, W_2), we can prove

$$\min_{\eta} f(\eta,\infty) = 1 - \frac{\alpha c_2^2}{\kappa\beta c_1^2}. \tag{34}$$

Simulation: comparison with related work

The data generation and weight initialization is the following

$$X = I_{20}, Y = XW(0) + 0.01\varepsilon\,,$$
$$W(0) \in \mathbb{R}^{20\times 1}, W(0)[i,j] \sim \mathcal{N}(0, 1/4)\,, \tag{35}$$
$$\varepsilon \in \mathbb{R}^{20\times 1}, \varepsilon[i,j] \sim \mathcal{N}(0,1)\,.$$

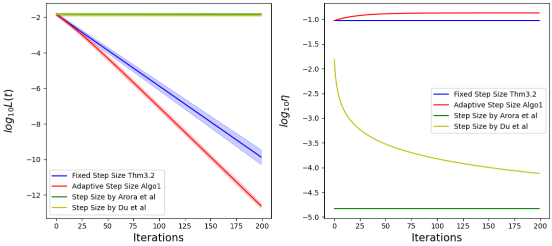

Figure 1

Simulation: over-parametrized vs non-overparametrized

The data generation and weight initialization is the following

$$\begin{aligned} & X, Y \in R^{1000\times 20}, Y = XW(0) + 0.001\varepsilon\,, \\ & W(0) \in \mathbb{R}^{20\times 20}, W(0)[i,j] \sim \mathcal{N}(0,1)\,, \\ & \varepsilon \in \mathbb{R}^{20\times 20}, \varepsilon[i,j] \sim \mathcal{N}(0,1)\,. \end{aligned} \tag{36}$$

We monitor the number of iterations needed to reach error 10^{-8}.

	over-parametrization	non-overparametrization
normal	18.92	14
NTK	12.7	9.08
xavier	14.74	12
He	13.8	10.96
uniform	17	12.96

Conclusion and Future Work

The contribution of our work is the following,

- We prove in the small learning rate regime, linear networks optimized via GD has linear convergence.
- We design a learning rate scheduler based on our theory.

For future work:

- Our work is in the small learning rate regime, what will happen in the large learning rate regime?
- How does imbalance interact with other phenomenon in deep learning, such as edge of instability, flat minima?