A Mean Field Analysis Of Deep ResNet:
Towards Provable Optimization Via Overparameterization From Depth

Joint work with Chao Ma, Yulong Lu, Jianfeng Lu and Lexing Ying

Presenter:
Yiping Lu
Contact:
yplu@stanford.edu,
https://web.stanford.edu/~yplu/

Stanford
University

Global Convergence Proof Of NN

- Neural Tangent Kernel([Jacot et al.2019]):Linearize the model $f_{NN}(\theta) = f_{NN}(\theta_{init}) + < \nabla_\theta f_{NN}(\theta_{init}), \theta - \theta_{init} >$

Stanford University

Global Convergence Proof Of NN

- Neural Tangent Kernel([Jacot et al.2019]):Linearize the model $f_{NN}(\theta) = f_{NN}(\theta_{init}) + < \nabla_\theta f_{NN}(\theta_{init}), \theta - \theta_{init} >$
  - **Pro:** can provide proof of convergence for any structure of NN. ([Li et al. 2019])
  - **Con:** Feature is lazy learned, *i.e.* not data dependent. ([Chizat and Bach 2019.][Ghorbani et al.2019])

Stanford University

Global Convergence Proof Of NN

- Neural Tangent Kernel([Jacot et al.2019]):Linearize the model $f_{\mathrm{NN}}(\theta) = f_{\mathrm{NN}}(\theta_{\mathrm{init}}) + < \nabla_\theta f_{\mathrm{NN}}(\theta_{\mathrm{init}}), \theta - \theta_{\mathrm{init}} >$
  - **Pro:** can provide proof of convergence for any structure of NN. ([Li et al. 2019])
  - **Con:** Feature is lazy learned, *i.e.* not data dependent. ([Chizat and Bach 2019.][Ghorbani et al.2019])

- Mean Field Regime([Bengio et al.2006][Bach et al.2014][Suzuki et al.2015]): We consider properties of the loss landscape with respect to the distribution of weights $L(\rho) = \|\mathbb{E}_{\theta\sim\rho} g(\theta, x) - f(x)\|_2^2$, the objective is a convex function

Stanford University

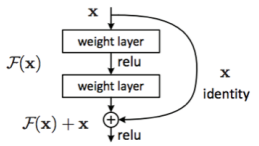
Global Convergence Proof Of NN

- Neural Tangent Kernel([Jacot et al.2019]):Linearize the model $f_{NN}(\theta) = f_{NN}(\theta_{init}) + < \nabla_\theta f_{NN}(\theta_{init}), \theta - \theta_{init} >$
  - **Pro:** can provide proof of convergence for any structure of NN. ([Li et al. 2019])
  - **Con:** Feature is lazy learned, *i.e.* not data dependent. ([Chizat and Bach 2019.][Ghorbani et al.2019])

- Mean Field Regime([Bengio et al.2006][Bach et al.2014][Suzuki et al.2015]): We consider properties of the loss landscape with respect to the distribution of weights $L(\rho) = \|\mathbb{E}_{\theta\sim\rho} g(\theta, x) - f(x)\|_2^2$, the objective is a convex function
  - **Pro:** SGD = Wasserstein Gradient Flow ([Mei et al.2018][Chizat et al.2018][Rotskoff et al.2018])
  - **Con:** Hard to generalize beyond two layer

Stanford University

Neural Ordinary Differential Equation



First-order ODE ⟹ Numerical solver: Euler's method ⟹ ResNet

[He et al. 2015]

$$\frac{dx}{dt} = F(x,t),$$

$$x(0) = x_0,$$

$$x_{l+1} = x_l + \gamma F(x_l, t_l),$$

$$x_0 \doteq x(0), x_l = x(\gamma l), \ldots \gamma - step\ size$$

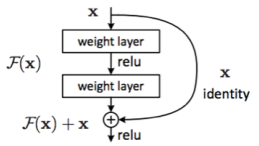
$$x_{l+1} = x_l + F(x_l, t_l),$$

[E. 2017] [Haber et al. 2017] [Lu et al. 2017] [Sho et al 2017] [Chen et al. 2018]

Figure: ResNet can be seen as the Euler discretization of a time evolving ODE

Stanford University

Neural Ordinary Differential Equation



First-order ODE ⟹ Numerical solver: Euler's method ⟹ ResNet

[He et al. 2015]

$$\frac{dx}{dt} = F(x,t),$$

$$x(0) = x_0,$$

$$x_{l+1} = x_l + \gamma F(x_l, t_l),$$

$$x_0 \doteq x(0), x_l = x(\gamma l), \ldots \gamma - step\ size$$

$$x_{l+1} = x_l + F(x_l, t_l),$$

[E. 2017] [Haber et al. 2017] [Lu et al. 2017] [Sho et al 2017] [Chen et al. 2018]

Figure: ResNet can be seen as the Euler discretization of a time evolving ODE

- Limit of depth$\rightarrow \infty$
- This analogy does not directly provide guarantees of global convergence even in the continuum limit.

Stanford University

Mean Field ResNet

Our Aim: Provide a **new** continuous limit for ResNet with good limiting landscape.

Idea:We consider properties of the loss landscape with respect to the distribution of weights.

Stanford University

Mean Field ResNet

Our Aim: Provide a **new** continuous limit for ResNet with good limiting landscape.

Idea:We consider properties of the loss landscape with respect to the distribution of weights.

$$\dot{X}_\rho(x,t) = \int_\theta f(X_\rho(x,t),\theta)\rho(\theta,t)d\theta$$

Residual block sample from ρ

Here:

- Input data is the initial condition $X_\rho(x,0) = \langle w_2, x\rangle$
- X is the feature, t represents the depth.
- Loss function:$E(\rho) = \mathbb{E}_{x\sim\mu}\left[\frac{1}{2}\left(\langle w_1, X_\rho(x,1)\rangle - y(x)\right)^2\right]$.

Stanford University

Adjoint Equation

To optimize the Mean Field model, we calculate the gradient $\frac{\delta E}{\delta \rho}$ via the *adjoint sensitivity method*.

Model

The loss function can be written as

$$\mathbb{E}_{x\sim\mu} E(x;\rho) := \mathbb{E}_{x\sim\mu} \frac{1}{2} \left| \langle w_1, X_\rho(x,1) \rangle - y(x) \right|^2 \tag{1}$$

where X_ρ satisfies the equation $\dot{X}_\rho(x,t) = \int_\theta f(X_\rho(x,t),\theta)\rho(\theta,t)d\theta$,

Stanford University

Adjoint Equation

To optimize the Mean Field model, we calculate the gradient $\frac{\delta E}{\delta \rho}$ via the *adjoint sensitivity method*.

Model

The loss function can be written as

$$\mathbb{E}_{x\sim\mu} E(x;\rho) := \mathbb{E}_{x\sim\mu} \frac{1}{2} \left| \langle w_1, X_\rho(x,1) \rangle - y(x) \right|^2 \tag{1}$$

where X_ρ satisfies the equation $\dot{X}_\rho(x,t) = \int_\theta f(X_\rho(x,t),\theta)\rho(\theta,t)d\theta$,

Adjoint Equation. The gradient can be represented as a second backwards-in-time augmented ODE.

$$\begin{aligned} \dot{p}_\rho(x,t) &= -\delta_X H_\rho(p_\rho, x, t) \\ &= -p_\rho(x,t) \int \nabla_X f(X_\rho(x,t),\theta)\rho(\theta,t)d\theta, \end{aligned}$$

Here the Hamiltonian is defined as $H_\rho(p,x,t) = p(x,t) \cdot \int f(x,\theta)\rho(\theta,t)d\theta$.

Adjoint Equation

Theorem

For $\rho \in \mathcal{P}^2$ let $\frac{\delta E}{\delta \rho}(\theta, t) = \mathbb{E}_{x\sim\mu} f(X_\rho(x,t),\theta))p_\rho(x,t)$, where p_ρ is the solution to the backward equation $\dot{p}_\rho(x,t) = -p_\rho(x,t)\int \nabla_X f(X_\rho(x,t),\theta)\rho(\theta,t)d\theta$. Then for every $\nu \in \mathcal{P}^2$, we have

$$E(\rho + \lambda(\nu - \rho)) = E(\rho) + \lambda \left\langle \frac{\delta E}{\delta \rho}, (\nu - \rho) \right\rangle + o(\lambda)$$

for the convex combination $(1-\lambda)\rho + \lambda\nu \in \mathcal{P}^2$ with $\lambda \in [0,1]$.

Adjoint equation is equivalent to the back propagation

Li Q, Chen L, Tai C, et al. Maximum principle based algorithms for deep learning. JMLR 2019
Zhang D, Zhang T, Lu Y, et al. You only propagate once: Painless adversarial training using maximal principle Neurips2019

Stanford University

Deep Residual Network Behaves Like an Ensemble Of Shallow Models

$$X^1 = X^0 + \frac{1}{L}\int_{\theta^0} \sigma(\theta^0 X^0)\rho^0(\theta^0)d\theta^0.$$

Stanford University

Deep Residual Network Behaves Like an Ensemble Of Shallow Models

$$X^1 = X^0 + \frac{1}{L}\int_{\theta^0} \sigma(\theta^0 X^0)\rho^0(\theta^0)d\theta^0.$$

$$X^2 = X^0 + \frac{1}{L}\int_{\theta^0} \sigma(\theta^0 X^0)\rho^0(\theta^0)d\theta^0 + \int_{\theta^1} \sigma(\theta^1(X^0 + \frac{1}{L}\int_{\theta^0} \sigma(\theta^0 X^0)\rho^0(\theta^0)d\theta^0))\rho^1(\theta^1)d\theta^1$$

$$= X^0 + \frac{1}{L}\int_{\theta^0} \sigma(\theta^0 X^0)\rho^0(\theta^0)d\theta^0 + \frac{1}{L}\int_{\theta^1} \sigma(\theta^1 X^0)\rho^1(\theta^1)d\theta^1 + \frac{1}{L^2}\int_{\theta_1} \nabla\sigma(\theta^1 X^0)\theta^1(\int_{\theta^0} \sigma(\theta^0 X^0)\rho^0(\theta^0)d\theta^0)\rho^1(\theta^1)d\theta^1$$

$$+ h.o.t.$$

Deep Residual Network Behaves Like an Ensemble Of Shallow Models

$$X^1 = X^0 + \frac{1}{L}\int_{\theta^0} \sigma(\theta^0 X^0)\rho^0(\theta^0)d\theta^0.$$

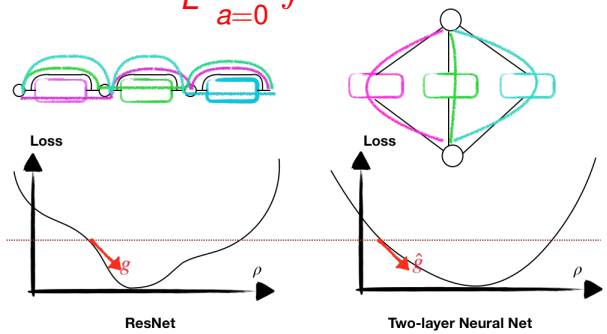
$$X^2 = X^0 + \frac{1}{L}\int_{\theta^0} \sigma(\theta^0 X^0)\rho^0(\theta^0)d\theta^0 + \int_{\theta^1} \sigma(\theta^1(X^0 + \frac{1}{L}\int_{\theta^0} \sigma(\theta^0 X^0)\rho^0(\theta^0)d\theta^0))\rho^1(\theta^1)d\theta^1$$

$$= X^0 + \frac{1}{L}\int_{\theta^0} \sigma(\theta^0 X^0)\rho^0(\theta^0)d\theta^0 + \frac{1}{L}\int_{\theta^1} \sigma(\theta^1 X^0)\rho^1(\theta^1)d\theta^1 + \frac{1}{L^2}\int_{\theta_1} \nabla\sigma(\theta^1 X^0)\theta^1(\int_{\theta^0} \sigma(\theta^0 X^0)\rho^0(\theta^0)d\theta^0)\rho^1(\theta^1)d\theta^1$$

$$+ h.o.t.$$

Iterating this expansion gives rise to

$$X^L \approx X^0 + \frac{1}{L}\sum_{a=0}^{L-1}\int \sigma(\theta X^0)\rho^a(\theta)d\theta + \frac{1}{L^2}\sum_{b>a}\int\int \nabla\sigma(\theta^b X^0)\theta^b\sigma(\theta^a X^0)\rho^b(\theta^b)\rho^a(\theta^a)d\theta^b\theta^a + h.o.t.$$



Veit A, Wilber M J, Belongie S. **Residual networks behave like ensembles of relatively shallow networks.** Advances in neural information processing systems. 2016: 550-558.

Stanford University

Deep Residual Network Behaves Like an Ensemble Of Shallow Models

Difference of back propagation process of two-layer net and ResNet.

Two-layer Network

$$\frac{\delta E}{\delta \rho}(\theta, t) = \mathbb{E}_{x\sim\mu} f(x, \theta))(X_\rho - y(x))$$

ResNet

$$\frac{\delta E}{\delta \rho}(\theta, t) = \mathbb{E}_{x\sim\mu} f(X_\rho(x, t), \theta))p_\rho(x, t)$$

We aim to show that the two gradient are similar.

Stanford University

Deep Residual Network Behaves Like an Ensemble Of Shallow Models

Difference of back propagation process of two-layer net and ResNet.

Two-layer Network

$$\frac{\delta E}{\delta \rho}(\theta, t) = \mathbb{E}_{x\sim\mu} f(x, \theta))(X_\rho - y(x))$$

ResNet

$$\frac{\delta E}{\delta \rho}(\theta, t) = \mathbb{E}_{x\sim\mu} f(X_\rho(x, t), \theta))p_\rho(x, t)$$

We aim to show that the two gradient are similar.

Lemma

The norm of the solution to the adjoint equation can be bounded by the loss

$$\|p_\rho(\cdot, t)\|_\mu \geq e^{-(C_1+C_2 r)} E(\rho), \forall t \in [0, 1]$$

,

Local = Global

Theorem

If $E(\rho) > 0$ *for distribution* $\rho \in \mathcal{P}^2$ *that is supported on one of the nested sets* Q_r*, we can always construct a descend direction* $\nu \in \mathcal{P}^2$*, i.e.*

$$\inf_{\nu \in \mathcal{P}^2} \left\langle \frac{\delta E}{\delta \rho}, (\nu - \rho) \right\rangle < 0$$

Stanford University

Local = Global

Theorem

If $E(\rho) > 0$ *for distribution* $\rho \in \mathcal{P}^2$ *that is supported on one of the nested sets* Q_r*, we can always construct a descend direction* $\nu \in \mathcal{P}^2$*, i.e.*

$$\inf_{\nu \in \mathcal{P}^2} \left\langle \frac{\delta E}{\delta \rho}, (\nu - \rho) \right\rangle < 0$$

Corollary

Consider a stationary solution to the Wasserstein gradient flow which is full support(informal), then it's a global minimizer.

Stanford University

Numerical Scheme

We may consider using a parametrization of ρ with n particles as

$$\rho_n(\theta, t) = \sum_{i=1}^{n} \delta_{\theta_i}(\theta) \mathbb{1}_{[\tau_i, \tau_i')}(t).$$

The characteristic function $\mathbb{1}_{[\tau_i, \tau_i')}$ can be viewed as a relaxation of the Dirac delta mass $\delta_{\tau_i}(t)$.

Given: A collection of residual blocks $(\theta_i, \tau_i)_{i=1}^n$
while training **do**
  Sort (θ_i, τ_i) based on τ_i to be (θ^i, τ^i) where $\tau^0 \leq \cdots \leq \tau^n$.
  Define the ResNet as $X^{\ell+1} = X^\ell + (\tau^\ell - \tau^{\ell-1})\sigma({}^\ell X^\ell)$ for $0 \leq \ell < n$.
  Use gradient descent to update both θ^i and τ^i.
end while

Stanford University

Numerical Results

	Vanilla	mean-field	Dataset
ResNet20	8.75	8.19	CIFAR10
ResNet32	7.51	7.15	CIFAR10
ResNet44	7.17	6.91	CIFAR10
ResNet56	6.97	6.72	CIFAR10
ResNet110	6.37	6.10	CIFAR10
ResNet164	5.46	5.19	CIFAR10
ResNeXt29(864d)	17.92	17.53	CIFAR100
ResNeXt29(1664d)	17.65	16.81	CIFAR100

Table: Comparison of the stochastic gradient descent and mean-field training (Algorithm 1.) of ResNet On CIFAR Dataset. Results indicate that our method our performs the Vanilla SGD consistently.

Stanford University

Take Home Message

- We propose a new continuous limit for deep resnet

$$\dot{X}_\rho(x,t) = \int_\theta f(X_\rho(x,t),\theta)\rho(\theta,t)d\theta,$$

  with initial $X_\rho(x,0) = \langle w_2, x\rangle$

- Local minimizer is global in ℓ_2 space.
- A potential scheme to approximate.

Stanford University

Take Home Message

- We propose a new continuous limit for deep resnet

$$\dot{X}_\rho(x,t) = \int_\theta f(X_\rho(x,t),\theta)\rho(\theta,t)d\theta,$$

with initial $X_\rho(x,0) = \langle w_2, x\rangle$

- Local minimizer is global in ℓ_2 space.
- A potential scheme to approximate.

TO DO List.

- Analysis of Wasserstein gradient flow. (Global Existence)
- Refined analysis of numerical scheme
- h.o.t in the expansion from ResNet to ensemble of small networks.

Stanford University

Thanks

Reference: Yiping Lu*, Chao Ma, Yulong Lu, Jianfeng Lu, Lexing Ying. "A Mean-field Analysis of Deep ResNet and Beyond: Towards Provable Optimization Via Overparameterization From Depth" arXiv:2003.05508 ICLR 2020 Workshop on Integration of Deep Neural Models and Differential Equations. **(Contributed Talk)**

Contact:yplu@stanford.edu