A MEAN-FIELD ANALYSIS OF DEEP RESNET AND BEYOND: TOWARDS PROVABLE OPTIMIZATION VIA OVERPARAMETERIZATION FROM DEPTH

Anonymous authors
Paper under double-blind review

ABSTRACT

To understand the success of SGD achieving zero training loss for training deep neural networks, this work presents a mean-field analysis of deep residual networks, based on a line of works that interpret the continuum limit of the deep residual network as an ordinary differential equation when the network capacity tends to infinity. Specifically, we propose a **new continuum limit** of deep residual networks, which enjoys a good landscape in the sense that **every local minimizer is global**, without assuming the convexity of the loss landscape. Our proof relies on a zero-loss assumption at the global minimizer that can be achieved when the model shares a universal approximation property. Key to our result is the observation that a deep residual network resembles a shallow network ensemble (47), *i.e.* a two-layer network. We bound the difference between the shallow network and our ResNet model via the adjoint sensitivity method, which enables us to apply existing mean-field analyses of two-layer networks to deep networks. Furthermore, we propose several novel training schemes based on the new continuous model, including one training procedure that switches the order of the residual blocks and results in strong empirical performance on the benchmark datasets.

1 INTRODUCTION

Neural networks have become state-of-the-art models in numerous machine learning tasks and strong empirical performance is often achieved by deeper networks. One landmark example is the residual network (ResNet) (22; 23), which can be efficiently optimized even at extremely large depth such as 1000 layers. However, there exists a gap between this empirical success and the theoretical understanding: ResNets can be trained to almost zero loss with standard stochastic gradient descent, yet it is known that larger depth leads to increasingly non-convex landscape even the the presence of residual connections (49). While global convergence can be obtained in the so-called "lazy" regime e.g. (24; 15), such kernel models cannot capture fully-trained neural networks (45; 13; 19).

In this work, we aim to demonstrate the provable optimization of ResNet beyond the restrictive "lazy" regime. To do so, we build upon recent works that connect ordinary differential equation (ODE) models to infinite-depth neural networks (17; 33; 43; 20; 10; 16; 51; 46; 44; 32). Specifically, each residual block of a ResNet can be written as $x_{n+1} = x_n + \Delta t f(x_n, \theta_n)$, which can be seen as the Euler discretization of the ODE $\dot{x}_t = f(x, t)$. This turns training the neural network into solving an optimal control problem (26; 18; 30), under which backpropagation can be understood as simulating the adjoint equation (10; 26; 27; 50; 28). (46) proved the Gamma-convergence of ResNets in the asymptotic limit. However, there are two points of that approach that require further investigation. First, (46) introduced a regularization term $n \sum_{i=1}^{n} \|\theta_i - \theta_{i-1}\|^2$, where n is the depth of the network. This regularization becomes stronger as the network gets deeper, which implies a more constrained space of functions that the network can represent. However, this analogy does not directly provide guarantees of global convergence even in the continuum limit.

To address the problem of global convergence, we propose **a new limiting ODE model** of ResNets. Formally, we model deep ResNets via the following mean-field ODE

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

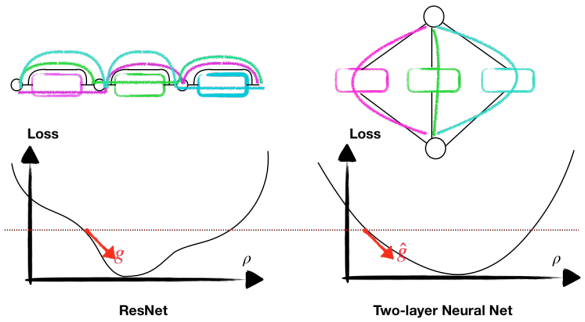

Figure 1: Illustration that ResNet behaves like shallow network ensemble, *i.e.* a two-layer overparameterized neural network. The high-level intuition is to show that the gradient of the two models are at the same scale when the loss are comparable.

This model considers every residual block $f(\cdot, \theta_i)$ as a particle and optimizes over the empirical distribution $\rho(\theta, t)$ of particles, where θ denotes the weight of the residual block and t denotes the layer index of the residual block. We consider properties of the loss landscape with respect to the distribution of weights, an approach similar to (7; 4; 34; 41) which states that two-layer neural networks enjoy a convex landscape. Inspired by (47) that a deep ResNet behaves like an ensemble of shallow models, we compare a deep ResNet with its counterpart two-layer network and show that the gradients of the two models are close to each other. This leads us to conclude that, although the loss landscape may not be convex, every local minimizer is a global one.

1.1 Contribution

- We derive a new continuous depth limit of deep ResNets. In this new model, each residual block is regarded as a particle and the training dynamics is captured by the gradient flow on the distribution ρ of the particle.
- We analyze the loss landscape with respect to ρ and show that all local minima have zero loss, which implies that every local optima is global. This property leads to the conclusion that a full support stationary point of the Wasserstein gradient flow is a global optimum. To the best of our knowledge, this is the **first global convergence result for multi-layer neural networks in the mean-field regime** without the convexity assumption on the loss landscape.
- We propose novel numerical schemes to approximate the mean-field limit of the deep ResNets and demonstrate that they achieve superior empirical results on real-world datasets.

2 Limiting Model

2.1 A New Continuous Model

The goal is to minimize the l_2 loss function

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

with respect to the parameter distribution $\rho(\theta, t)$ where the parameter θ is defined on a compact set Ω and $t \in [0, 1]$. Here $X_\rho(x, t)$ is the solution of the ODE

$$\dot{X}_\rho(x,t) = \int_\theta f(X_\rho(x,t),\theta)\rho(\theta,t)d\theta, \quad X_\rho(x,0) = \langle w_2, x\rangle\,. \tag{2}$$

The ODE above is understood as an integral equation, *i.e.*, for fixed distribution $\rho(\cdot,\cdot)$ and input $x \in \mathbb{R}^{d_1}$, the solution path $X_\rho(x,t), t \in [0,1]$ satisfies

$$X_\rho(x,t) = X_\rho(x,0) + \int_0^t \int_\Omega f(X_\rho(x,s),\theta)\rho(\theta,s)d\theta ds.$$

Here $y(x) = \mathbb{E}[y|x] \in \mathbb{R}$ is the function to be estimated. The parameter $w_2 \in \mathbb{R}^{d_1 \times d_2}$ represents the first convolution layer in the ResNet (22; 23), which extracts feature before sending them to the residual blocks. To simplify the analysis, we let w_2 to a predefined linear transformation (*i.e.* not training the first layer parameters w_2) with the technical assumption that $\min\{\sigma(w_2)\} \geq \sigma_1$ and $\max\{\sigma(w_2)\} \leq \sigma_2$, where $\sigma(w_2)$ denotes the set of singular values. We remark that this assumption is not unrealistic, for example (38) let w_2 be a predefined wavelet transform and still achieved the state-of-the-art result on several benchmark datasets. Here $f(\cdot, \theta)$ is the residual block with parameter θ that aims to learn a feature transformation from $\mathbb{R}^{d_2}$ to $\mathbb{R}^{d_2}$. For simplicity, we assume that the residual block is a two layer neural network, thus $f(x, \theta) = \sigma(\theta x), \theta \in \Omega \subset \mathbb{R}^{d_2 \times d_2}$ and $\sigma : \mathbb{R} \to \mathbb{R}$ is an activation function, such as sigmoid and relu. Note that in our notation $\sigma(\theta x)$ the activation function σ is applied separately to each component of the vector. Finally, $w_1 \in \mathbb{R}^{d_2 \times 1}$ is a pooling operator that transfers the final feature $X_\rho(x, 1)$ to the classification result and an l_2 loss function is used for example. We also assume that w_1 is a predefined linear transform with satisfies $\|w_1\|_2 = 1$, which can be easily achieved via an operator used in realistic architectures such as the global average pooling (29). Due to the page limit, we list the necessary regularity assumptions and results in the appendix.

3 Landscape Analysis of the Mean-Field Model

In this section we show that the landscape of a deep residual network satisfies a remarkable property that any local optima is global, by comparing the gradient of deep residual network with the mean-field model of two-layer neural network (34; 13; 36). To estimate the accuracy of the first order approximation (*i.e.* linearization), we apply the adjoint sensitivity analysis (8) and show that the difference between the gradient of two models can be bounded via the stability constant of the backward adjoint equation. More precisely, the goal is to show the backward adjoint equation will only affect the gradient in a bounded constant.

3.1 Gradient via the Adjoint Sensitivity Method

Adjoint Equation. To optimize the objective equation 1, we calculate the gradient $\frac{\delta E}{\delta \rho}$ via the *adjoint sensitivity method* (8). We first derive the adjoint equation which represents the gradient as a second backwards-in-time augmented ODE. The derivative of $E(\rho)$ with respect to $X_\rho(x, t)$ satisfies at any previous time $s \leq 1$ the adjoint equation of the ODE

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

with terminal condition $p_\rho(x, 1) := \frac{\partial E(x;\rho)}{\partial X_\rho(x,1)} = \big(\langle w_1, X_\rho(x, 1)\rangle - y(x)\big) w_1$. Here the Hamiltonian is defined as $H_\rho(p, x, t) = p(x, t) \cdot \int f(x, \theta) \rho(\theta, t) d\theta$.

Utilizing the adjoint equation, we can characterize the gradient of our model with respect to the distribution ρ. More precisely, we may characterize the variation of the loss function with respect to the distribution as the following theorem. A detailed proof is presented in the Appendix.

Theorem 1. *(Gradient of the parameter) For $\rho \in \mathcal{P}^2$ let $\frac{\delta E}{\delta \rho}(\theta, t) = \mathbb{E}_{x \sim \mu} f(X_\rho(x, t), \theta)) p_\rho(x, t)$. Then for every $\nu \in \mathcal{P}^2$, $E(\rho + \lambda(\nu - \rho)) = E(\rho) + \lambda \left\langle \frac{\delta E}{\delta \rho}, (\nu - \rho) \right\rangle + o(\lambda)$ for the convex combination $(1 - \lambda)\rho + \lambda\nu \in \mathcal{P}^2$ with $\lambda \in [0, 1]$.*

3.2 Landscape Analysis

In this section we aim to show that the proposed model enjoys a good landscape in the L_2 geometry. Specifically, we can always find a descent direction around a point whose loss is strictly larger than 0, which means that all local minimum is a global one. We list here a proof sketch and the details are given in the Appendix.

Theorem 2. *If $E(\rho) > 0$ for a distribution $\rho \in \mathcal{P}^2$ supported on one of the nested sets Q_r, we can always construct a descend direction $\nu \in \mathcal{P}^2$,* i.e. $\inf_{\nu \in \mathcal{P}^2} \left\langle \frac{\delta E}{\delta \rho}, (\nu - \rho) \right\rangle < 0$.

3.3 Discussion of the Wasserstein gradient flow

As described in the introduction, we consider each residual block as a particle and trace the evolution of the empirical distribution ρ_s of the particles during the training (here the variable s denotes the training time). While using gradient descent or stochastic gradient descent with small time steps, we move each particle through a velocity field $\{v_s\}_{s\geq 0}$ and the evolution can be expressed by a PDE $\partial_s \rho_s = \text{div}(\rho_s v_s)$, where div is the divergence operator (34; 12; 41; 42). Thus in this section, we consider the gradient flow of the the objective function in the Wasserstein space, given by a McKean–Vlasov type equation (9; 2; 25; 37; 36) $\frac{\partial_{(\theta,t)}\rho}{\partial s} = \text{div}_{(\theta,t)}\left(\rho \nabla_{(\theta,,t)} \frac{\delta E}{\delta \rho}\right)$. We aim to show some global convergence property of the stationary point of the Wasserstein gradient flow.

Theorem 3. *(Informal) When the residual block (X, θ) is positively p-homogeneous respective to θ. Let $(\rho_s)_{s\geq 0}$ be the solution of the the Wasserstein gradient $\frac{\partial_{(\theta,t)}\rho}{\partial s} = div_{(\rho,t)}(\rho \nabla_{(\rho,t)} \frac{\delta E}{\delta \rho})$ of our mean-field model (2). If $(\rho_s)_{s\geq 0}$ converge to ρ_∞ in W_2 and ρ^* concentrates in a ball $B(0, r_b)$ and separates the spheres $r_a \mathbb{S}^{d-1} \times [0,1]$ and $r_b \mathbb{S}^{d-1} \times [0,1]$, then ρ^* is the global minimum satisfies $E(\rho_\infty) = 0$.*

The precise statement and the proof are presented in Appendix, where we also analyze the regularity of our objective function in the Wasserstein space. The result guarantees that, when the gradient flow converges, it has to reach the global minimum of the loss function.

4 Deep ResNet as a Numerical Scheme

To turn the continuous model into a ResNet training algorithm, we use a set of particles to approximate the the distribution (36; 3; 31) and the Euler scheme to numerically solve the ODE model which leads to a simple Residual Network (33). Since ρ characterizes the distribution of the pairs (θ, t), each particle in our representation would carry the parameter θ, together with information on the activation time period of the particle. Therefore, also different from the usual ResNet, we also need to allow the particle to move in the gradient direction corresponding to t. We may consider using a parametrization of ρ with n particles as $\rho_n(\theta, t) = \sum_{i=1}^n \delta_{\theta_i}(\theta) 1_{[\tau_i, \tau_i']}(t)$. The characteristic function $1_{[\tau_i, \tau_i']}$ can be viewed as a relaxation of the Dirac delta mass $\delta_{\tau_i}(t)$. However, this parametrization comes with a difficulty in practice, namely, the intervals $[t_i, t_i']$ may overlap significantly with each other, and in the worst case, though unlikely, all the time intervals of the n particles coincide, which leads to heavy computational cost in the training process.

Therefore, for practical implementation, we constrain that every time instance t is just contained in the time interval of a single particle. We realize this by adding a constraint $\tau_i' = \tau_{i+1}$ between consecutive intervals. More precisely, given a set of parameters (θ^i, τ^i), we first sort them according to τ^i values. Assuming τ^i are ordered, we define the architecture as

$$X^{\ell+1} = X^\ell + (\tau^\ell - \tau^{\ell-1})\sigma(\theta^\ell X^\ell), \quad 0 \leq \ell < n; \tag{3}$$

$$X^0 = x. \tag{4}$$

Both θ and τ parameters can be trained with SGD and n is the depth of the network. The order of τ may change during the training (thus to make each particle indistinguishable to guarantee the mean-field behavior), thus after every update, we *sort* the τ_i to get the new order of the residual blocks. The algorithm is listed in Algorithm 1. The empirical results is listed in Table 5.1.

5 Discussion and Conclusion

5.1 Conclusion

To better understand the reason that the stochastic gradient descent can optimize the complicated landscape. Our work directly consider an infinitely deep residual network. We proposed a new continuous model of deep ResNets and established an asymptotic global optimality property by bounding the difference between the gradient of the deep residual network and an associated two-layer network. To the best of the author's knowledge, this is the first global convergence result of deep neural networks beyond the lazy training (kernel) regime (24; 15; 1; 52; 39).

	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 1: 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.

Algorithm 1 Training Of Mean-Field Deep Residual Network

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 $X^{\ell+1} = X^\ell + (\tau^\ell - \tau^{\ell-1})\sigma(\theta^\ell X^\ell)$ for $0 \leq \ell < n$.
  Use gradient descent to update both θ^i and τ^i.
end while

REFERENCES

[1] Z. ALLEN-ZHU, Y. LI, AND Z. SONG, *A convergence theory for deep learning via over-parameterization*, arXiv preprint arXiv:1811.03962, (2018).

[2] L. AMBROSIO, N. GIGLI, AND G. SAVARÉ, *Gradient flows: in metric spaces and in the space of probability measures*, Springer Science & Business Media, 2008.

[3] J. BA, M. A. ERDOGDU, M. GHASSEMI, T. SUZUKI, S. SUN, D. WU, AND T. ZHANG, *Towards characterizing the high-dimensional bias of kernel-based particle inference algorithms*, (2019).

[4] F. BACH, *Breaking the curse of dimensionality with convex neural networks*, The Journal of Machine Learning Research, 18 (2017), pp. 629–681.

[5] P. L. BARTLETT, S. N. EVANS, AND P. M. LONG, *Representing smooth functions as compositions of near-identity functions with implications for deep network optimization*, arXiv preprint arXiv:1804.05012, (2018).

[6] P. L. BARTLETT, D. P. HELMBOLD, AND P. M. LONG, *Gradient descent with identity initialization efficiently learns positive-definite linear transformations by deep residual networks*, Neural computation, 31 (2019), pp. 477–502.

[7] Y. BENGIO, N. L. ROUX, P. VINCENT, O. DELALLEAU, AND P. MARCOTTE, *Convex neural networks*, in Advances in neural information processing systems, 2006, pp. 123–130.

[8] V. BOLTYANSKIY, R. GAMKRELIDZE, Y. MISHCHENKO, AND L. PONTRYAGIN, *Mathematical theory of optimal processes*, (1962).

[9] J. A. CARRILLO, R. J. MCCANN, C. VILLANI, ET AL., *Kinetic equilibration rates for granular media and related equations: entropy dissipation and mass transportation estimates*, Revista Matematica Iberoamericana, 19 (2003), pp. 971–1018.

[10] T. Q. CHEN, Y. RUBANOVA, J. BETTENCOURT, AND D. K. DUVENAUD, *Neural ordinary differential equations*, in Advances in Neural Information Processing Systems, 2018, pp. 6572–6583.

[11] L. CHIZAT, *Sparse optimization on measures with over-parameterized gradient descent*, arXiv preprint arXiv:1907.10300, (2019).

[12] L. CHIZAT AND F. BACH, *On the global convergence of gradient descent for over-parameterized models using optimal transport*, in Advances in neural information processing systems, 2018, pp. 3036–3046.

[13] L. CHIZAT, E. OYALLON, AND F. BACH, *On lazy training in differentiable programming*, (2019).

[14] L. CHIZAT, G. PEYRÉ, B. SCHMITZER, AND F.-X. VIALARD, *An interpolating distance between optimal transport and fisher–rao metrics*, Foundations of Computational Mathematics, 18 (2018), pp. 1–44.

[15] S. S. DU, J. D. LEE, H. LI, L. WANG, AND X. ZHAI, *Gradient descent finds global minima of deep neural networks*, arXiv preprint arXiv:1811.03804, (2018).

[16] E. DUPONT, A. DOUCET, AND Y. W. TEH, *Augmented neural odes*, in Advances in Neural Information Processing Systems, 2019, pp. 3134–3144.

[17] W. E, *A proposal on machine learning via dynamical systems*, Communications in Mathematics and Statistics, 5 (2017), pp. 1–11.

[18] W. E, J. HAN, AND Q. LI, *A mean-field optimal control formulation of deep learning*, Research in the Mathematical Sciences, 6 (2019), p. 10.

[19] B. GHORBANI, S. MEI, T. MISIAKIEWICZ, AND A. MONTANARI, *Limitations of lazy training of two-layers neural networks*, arXiv preprint arXiv:1906.08899, (2019).

[20] E. HABER AND L. RUTHOTTO, *Stable architectures for deep neural networks*, Inverse Problems, 34 (2017), p. 014004.

[21] K. HE, X. ZHANG, S. REN, AND J. SUN, *Delving deep into rectifiers: Surpassing human-level performance on imagenet classification*, in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1026–1034.

[22] ——, *Deep residual learning for image recognition*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[23] ——, *Identity mappings in deep residual networks*, in European conference on computer vision, Springer, 2016, pp. 630–645.

[24] A. JACOT, F. GABRIEL, AND C. HONGLER, *Neural tangent kernel: Convergence and generalization in neural networks*, in Advances in neural information processing systems, 2018, pp. 8571–8580.

[25] R. JORDAN, D. KINDERLEHRER, AND F. OTTO, *The variational formulation of the fokker–planck equation*, SIAM journal on mathematical analysis, 29 (1998), pp. 1–17.

[26] Q. LI, L. CHEN, C. TAI, AND W. E, *Maximum principle based algorithms for deep learning*, The Journal of Machine Learning Research, 18 (2017), pp. 5998–6026.

[27] Q. LI AND S. HAO, *An optimal control approach to deep learning and applications to discrete-weight neural networks*, in Proceedings of the 35th International Conference on Machine Learning, J. Dy and A. Krause, eds., vol. 80 of Proceedings of Machine Learning Research, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018, PMLR, pp. 2985–2994.

[28] X. LI, T.-K. L. WONG, R. T. CHEN, AND D. DUVENAUD, *Scalable gradients for stochastic differential equations*, arXiv preprint arXiv:2001.01328, (2020).

[29] M. LIN, Q. CHEN, AND S. YAN, *Network in network*, arXiv preprint arXiv:1312.4400, (2013).

[30] G.-H. LIU AND E. A. THEODOROU, *Deep learning theory review: An optimal control and dynamical systems perspective*, arXiv preprint arXiv:1908.10920, (2019).

[31] Q. LIU AND D. WANG, *Stein variational gradient descent: A general purpose bayesian inference algorithm*, in Advances in neural information processing systems, 2016, pp. 2378–2386.

[32] Y. LU, Z. LI, D. HE, Z. SUN, B. DONG, T. QIN, L. WANG, AND T.-Y. LIU, *Understanding and improving transformer from a multi-particle dynamic system point of view*, arXiv preprint arXiv:1906.02762, (2019).

[33] Y. LU, A. ZHONG, Q. LI, AND B. DONG, *Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations*, arXiv preprint arXiv:1710.10121, (2017).

[34] S. MEI, A. MONTANARI, AND P.-M. NGUYEN, *A mean field view of the landscape of two-layer neural networks*, Proceedings of the National Academy of Sciences, 115 (2018), pp. E7665–E7671.

[35] C. A. MICCHELLI, Y. XU, AND H. ZHANG, *Universal kernels*, Journal of Machine Learning Research, 7 (2006), pp. 2651–2667.

[36] A. NITANDA AND T. SUZUKI, *Stochastic particle gradient descent for infinite ensembles*, arXiv preprint arXiv:1712.05438, (2017).

[37] F. OTTO, *The geometry of dissipative evolution equations: the porous medium equation*, (2001).

[38] E. OYALLON, E. BELILOVSKY, AND S. ZAGORUYKO, *Scaling the scattering transform: Deep hybrid networks*, in Proceedings of the IEEE international conference on computer vision, 2017, pp. 5618–5627.

[39] S. OYMAK AND M. SOLTANOLKOTABI, *Towards moderate overparameterization: global convergence guarantees for training shallow neural networks*, arXiv preprint arXiv:1902.04674, (2019).

[40] G. ROTSKOFF, S. JELASSI, J. BRUNA, AND E. VANDEN-EIJNDEN, *Neuron birth-death dynamics accelerates gradient descent and converges asymptotically*, in International Conference on Machine Learning, 2019, pp. 5508–5517.

[41] G. M. ROTSKOFF AND E. VANDEN-EIJNDEN, *Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error*, arXiv preprint arXiv:1805.00915, (2018).

[42] J. SIRIGNANO AND K. SPILIOPOULOS, *Mean field analysis of neural networks: A central limit theorem*, Stochastic Processes and their Applications, (2019).

[43] S. SONODA AND N. MURATA, *Double continuum limit of deep neural networks*, in ICML Workshop Principled Approaches to Deep Learning, 2017.

[44] ——, *Transport analysis of infinitely deep neural network*, The Journal of Machine Learning Research, 20 (2019), pp. 31–82.

[45] T. SUZUKI, *Adaptivity of deep relu network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality*, arXiv preprint arXiv:1810.08033, (2018).

[46] M. THORPE AND Y. VAN GENNIP, *Deep limits of residual neural networks*, arXiv preprint arXiv:1810.11741, (2018).

[47] A. VEIT, M. J. WILBER, AND S. BELONGIE, *Residual networks behave like ensembles of relatively shallow networks*, in Advances in neural information processing systems, 2016, pp. 550–558.

[48] S. XIE, R. GIRSHICK, P. DOLLÁR, Z. TU, AND K. HE, *Aggregated residual transformations for deep neural networks*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1492–1500.

[49] C. YUN, S. SRA, AND A. JADBABAIE, *Are deep resnets provably better than linear predictors?*, in Advances in Neural Information Processing Systems, 2019, pp. 15660–15669.

[50] D. ZHANG, T. ZHANG, Y. LU, Z. ZHU, AND B. DONG, *You only propagate once: Painless adversarial training using maximal principle*, arXiv preprint arXiv:1905.00877, (2019).

[51] X. ZHANG, Y. LU, J. LIU, AND B. DONG, *Dynamically unfolding recurrent restorer: A moving endpoint control method for image restoration*, in International Conference on Learning Representations, 2019.

[52] D. ZOU, Y. CAO, D. ZHOU, AND Q. GU, *Stochastic gradient descent optimizes over-parameterized deep relu networks*, arXiv preprint arXiv:1811.08888, (2018).

A APPENDIX

Here is the supplementary material for the paper: "A Mean Field Analysis Of Deep ResNet And Beyond: Towards Provable Optimization Via Overparameterization From Depth". The supplementray meterial is organized as following

- Section B: Notations and Preliminaries
- Section C: Properties of our continuous model.
- Section D: Detailed Proofs For Landscape Analysis.
- Section E: Properties of the loss function in the Wasserstein space.
- Section F: An analog of ResNet behaves like ensembe of shallow networks.
- Section G: Experiment.

B NOTATIONS AND PRELIMINARIES

Notations. Let $\delta(\cdot)$ denote the Dirac mass and 1_Ω be the indicator function on Ω. We denote by $\mathcal{P}^2$ the set of probability measures endowed with the Wasserstein-2 distance (see below for definition). Let μ be the population distribution of the input data and the induced norm by $\|f\|_\mu = \sqrt{\mathbb{E}_{x\sim\mu}[f(x)^\top f(x)]}$.

Fréchet Derivative. We extend the notion of the gradient to infinite dimensional space. For a functional $f : X \to \mathbb{R}$ defined on a Banach space X, the Fréchet derivative is an element in the dual space $df \in X^*$ that satisfies

$$\lim_{\delta\in X, \delta\to 0} \frac{f(x+\delta) - f(x) - df(\delta)}{\|\delta\|} = 0, \quad \text{for all } x \in X.$$

In this paper, $\frac{\delta f}{\delta X}$ is used to denote the Fréchet derivative.

Wasserstein Space. The Wasserstein-2 distance between two probability measures $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ is defined as

$$W_2(\mu,\nu) := \left(\inf_{\gamma\in\mathcal{T}(\mu,\nu)} \int |y-x|^2 d\gamma(x,y)\right)^{1/2}.$$

Here $\mathcal{T}(\mu,\nu)$ denotes the set of all couplings between μ and ν, i.e., all probability measures $\gamma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ with marginals μ on the first factor and ν on the second.

Bounded Lipschitz norm. We say that a sequence of measures $\mu_n \in \mathcal{M}(\mathbb{R}^d)$ *weakly* (or *narrowly*) converges to μ if, for all continuous and bounded function $\varphi : \mathbb{R}^d \to \mathbb{R}$ it holds $\int \varphi \underset{\cdot}{\mu}_n \to \int \varphi \underset{\cdot}{\mu}$. For sequences which are bounded in total variation norm, this is equivalent to the convergence in Bounded Lipschitz norm. The latter is defined, for $\mu \in (\mathbb{R}^d)$, as

$$\|\mu\|_{\text{BL}} := \sup\left\{\int \varphi \underset{\cdot}{\mu}\ ;\ \varphi : \mathbb{R}^d \to \mathbb{R},\ \text{Lip}(\varphi) \le 1,\ \|\varphi\|_\infty \le 1\right\} \tag{5}$$

where (φ) is the smallest Lipschitz constant of φ and $\|\cdot\|_\infty$ the supremum norm.

Assumptions. All proofs in this appendix are based on the following assumptions

Assumption 1. *1. (Boundedness of data and target distribution) The input data x lies μ-almost surely in a compact ball,* i.e. $\|x\| \le R_1$ *for some constant $R_1 > 0$. At the same time the target function is also bounded $\|y(\cdot)\|_\infty \le R_2$ for some constant $R_2 > 0$.*

2. (Lipschitz continuity of distribution with respect to depth) There exist a constant C_ρ such that

$$\|\rho(\cdot,t_1) - \rho(\cdot,t_2)\|_{BL} \le C_\rho |t_1 - t_2|$$

for all $t_1, t_2 \in [0,1]$.

3. *The kernel $k(x_1, x_2) := g(x_1, x_2) = \sigma(x_1^\intercal x_2)$ is a universal kernel (35), i.e. the span of $\{k(x, \cdot) : x \in \mathbb{R}^{d_2}\}$ is dense in L^2.*

4. *(Locally Lipschitz derivative with sub-linear growth (12)) There exists a family $\{Q_r\}_{r>0}$ of nested nonempty closed convex subsets of Ω that satisfies:*

   - $\{u \in \Omega \mid (u, Q_r) \leq r'\} \subset Q_{r+r'}$ *for all* $r, r' > 0$.
   - *There exist constants $C_1, C_2 > 0$ such that*

     $$\sup_{\theta \in Q_r, x} \|\nabla_x f(x, \theta)\| \leq C_1 + C_2 r$$

     holds for all $r > 0$. Also the gradient of $f(x, \theta)$ with respect of x is a Lipschitz function with Lipschitz constant $L_r > 0$.
   - *For each r, the gradient respect to the parameter θ is also bounded*

     $$\sup_{\|x\| \leq R_1, \theta \in Q_r} \|\nabla_\theta f(x, \theta)\| \leq C_{3,r}$$

     for some constant $C_{3,r}$.

Let us elaborate on these assumptions in the neural network setting. For Assumption 1.4, $k(x_1, x_2) := g(x_1, x_2) = \sigma(x_1^\intercal x_2)$ is a universal kernel holds for the sigmoid and ReLU activation function. The local regularity Assumption 1.5 concerning function $f(x, \theta)$ can easily be satisfied, for $\nabla_\theta \sigma(\theta^\intercal x) = \sigma'(\theta^\intercal x)x$ and $\nabla_x \sigma(\theta^\intercal x) = \sigma'(\theta^\intercal x)\theta$. Hence, in order to satisfy the local regularity condition, one possible solution is that we utlize a Lipschitz gradient activation function and set the local set Q_r to be a ball with radius r centered at origin.

C PROPERTIES OF OUR CONTINUOUS MODEL.

Theorem 4. *Under Assumption 1 and we further assume that there exists a constant $r > 0$ such that μ is concentrated on one of the nested sets Q_r. Then, the ODE in equation 2 has a unique solution in $t \in [0, 1]$ for any initial condition $x \in \mathbb{R}^{d_1}$. Moreover, for any pair of distributions ρ_1 and ρ_2, there exists a constant C such that*

$$\|X_{\rho_1}(x, 1) - X_{\rho_2}(x, 1)\| < C W_2(\rho_1, \rho_2), \tag{6}$$

where $W_2(\rho_1, \rho_2)$ is the 2-Wasserstein distance between ρ_1 and ρ_2.

Proof. We first show the existence and uniqueness of $X_\rho(x, t)$. From now on, let

$$F_\rho(X, t) = \int_\theta f(X, t)\rho(\theta, t)d\theta. \tag{7}$$

Then, the ODE equation 2 becomes

$$\dot{X}_\rho(x, t) = F_\rho(X_\rho(x, t), t), \tag{8}$$

and by the condition of the theorem and assumption 1 we have

$$\|F_\rho(X, t)\| \leq C_f^r \left| \int_\theta \rho(\theta, t)d\theta \right| < C_f^r C_\rho. \tag{9}$$

This is because, for the continuous function $f(x, \theta)$ is now defined on the domain for which θ lies in a compact set Q_r and $\|x\| < R_1$, which leads to an upper bound C_f^r such that $\sup_{\|x\|<R} f(x, \theta) < C_f^r$ holds for all $\theta \in Q_r$. The notation C_f^r will continuously used in the following section.

Hence, $F_\rho(X_\rho, t)$ is bounded. On the other hand, $F_\rho(X, t)$ is integrable with respect to t and Lipschitz continuous with respect to X in any bounded region (by 2 of assumption 1). Therefore, consider the region $[X_0 - C_f^r C_\rho, X_0 + C_f^r C_\rho] \times [0, 1]$, where $X_0 = X_\rho(x, 0)$. By the existence and uniqueness theorem of ODE (the Picard–Lindelöf theorem), the solution of equation 8 initialized from X_0 exists and is unique on $[0, 1]$.

Next, we show the continuity of $X_\rho(x,t)$ with respect to ρ. Letting $\Delta(x,t) = \|X_{\rho_1}(x,t) - X_{\rho_2}(x,t)\|$, we have

$$\begin{aligned}\Delta(x,t) &= \left\| \int_0^t \dot{X}_{\rho_1}(x,s) - \dot{X}_{\rho_2}(x,s) ds \right\| \\ &= \left\| \int_0^t F_{\rho_1}(X_{\rho_1},s) - F_{\rho_1}(X_{\rho_2},s) ds + \int_0^t F_{\rho_1}(X_{\rho_2},s) - F_{\rho_2}(X_{\rho_2},s) ds \right\| \\ &\leq \int_0^t \|F_{\rho_1}(X_{\rho_1},s) - F_{\rho_1}(X_{\rho_2},s)\| ds + \left\| \int_0^t F_{\rho_1}(X_{\rho_2},s) - F_{\rho_2}(X_{\rho_2},s) ds \right\|. \end{aligned} \tag{10}$$

Let $C_m = \max\{C_{\rho_1}, C_{\rho_2}\}$. For the first term in equation 10, since both X_{ρ_1} and X_{ρ_2} are controlled by $X_0 + C_f^r C_m$, by 2 of Assumption 1 we have the following Lipschitz condition for

$$\|F_{\rho_1}(X_{\rho_1},s) - F_{\rho_1}(X_{\rho_2},s)\| \leq (C_1 + C_2 X_0 + C_2 C_f^r C_m) C_m \Delta(x,s). \tag{11}$$

For the second term of equation 10, we have

$$\left\| \int_0^t F_{\rho_1}(X_{\rho_2},s) - F_{\rho_2}(X_{\rho_2},s) ds \right\| = \left\| \int_0^t \int_\theta f(X_{\rho_2},\theta)(\rho_1(\theta,s) - \rho_2(\theta,s)) d\theta ds \right\|. \tag{12}$$

Since X_{ρ_2} is $C_f^r C_m$-Lipschitz continuous with respect to t and also bounded by $X_0 + C_f^r C_m$, we have $f(X_{\rho_2},\theta)$ is $(C_1 + C_2 X_0 + C_2 C_f^r C_m) C_f^r C_m$-Lipschitz continuous w.r.t t. On the other hand, still by Assumption 1, broken link $f(X,\theta)$ is $C_{3,r}$-Lipschitz with respect to θ. As a result, the function $f(X_{\rho_2},\theta)$ is C-Lipschitz continuous on (t,θ) with $C = (C_1 + C_2 X_0 + C_2 C_f^r C_m) C_f^r C_m + C_{3,r}$, which implies

$$\left\| \int_0^t \int_\theta f(X_{\rho_2},\theta)(\rho_1(\theta,s) - \rho_2(\theta,s)) d\theta ds \right\| \leq C W_2(\rho_1,\rho_2). \tag{13}$$

Finally, by defining

$$\hat{C} = \max\{(C_1 + C_2 X_0 + C_2 C_f^r C_m) C_m, C\}, \tag{14}$$

we have by equation 10

$$\Delta(x,t) \leq \int_0^t \hat{C}\Delta(x,t) + \hat{C} W_2(\rho_1,\rho_2). \tag{15}$$

Applying Gronwall's inequality gives

$$\Delta(x,t) \leq \hat{C} e^{\hat{C}t} W_2(\rho_1,\rho_2), \tag{16}$$

and specifically for $t = 1$ we have

$$\|X_{\rho_1}(x,1) - X_{\rho_2}(x,1)\| \leq \hat{C} e^{\hat{C}} W_2(\rho_1,\rho_2). \tag{17}$$

□

Theorem. *(Restate of Theorem 1.) For $\rho \in \mathcal{P}^2$ let*

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

then for every $\nu \in \mathcal{P}^2$ we have

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

Proof. To simplify the notation, we use $\hat{\rho}_\lambda = \rho + \lambda(\rho - \nu)$, From Theorem 4 (the well-poseness of the model), we know that the function $f(\lambda) = E(\hat{\rho}_\lambda) - E(\rho)$ is a continuous function with $f(0) = 0$ and thus

$$\begin{aligned} E(\hat{\rho}_\lambda) - E(\rho) &= \mathbb{E}_{x\sim\mu} |\langle w_1, X_{\hat{\rho}_\lambda}(x,1)\rangle - y(x)|^2 - \mathbb{E}_{x\sim\mu} |\langle w_1, X_\rho(x,1)\rangle - y(x)|^2 \\ &= \mathbb{E}_{x\sim\mu} (\langle w_1, X_\rho\rangle - y(x))(X_{\hat{\rho}_\lambda}(x,1) - X_\rho(x,1)) + O(X_{\hat{\rho}_\lambda}(x,1) - X_\rho(x,1)) \end{aligned}$$

Now we bound $X_{\hat{\rho}_\lambda}(x,1) - X_\rho(x,1)$. First, notice that the adjoint equation is a linear equation:

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

with solution

$$p(x,t) = p(x,1)\exp(\int_t^1 \int \nabla_X f(X_\rho(x,t),\theta)\rho(\theta,t)d\theta dt).$$

Next, we bound $\Delta(x,t) = \|X_{\hat{\rho}_\lambda}(x,t) - X_\rho(x,t) - \lambda \int_t \int_\theta (\rho(x,\theta) - \nu(x,\theta))p_\rho(x,t)\|$ in order to show that $\Delta(x,t) = o(\lambda)$. The way to estimate the difference is to utilize the Duhamel's principle.

$$\begin{aligned}
&\frac{d}{dt}\left[e^{-\int_0^t \int \nabla_X f(X_\rho(x,t),\theta)\rho(\theta,s)d\theta ds}(X_{\hat{\rho}_\lambda}(x,s) - X_\rho(x,s))\right] \\
&= e^{-\int_0^t \int \nabla_X f(X_\rho(x,t),\theta)\rho(\theta,s)d\theta ds}\left[\dot{X}_{\hat{\rho}_\lambda}(x,s) - \dot{X}_\rho(x,s) - \int_\theta \nabla_X f(X_\rho(x,t),\theta)\rho(\theta,t)d\theta(X_{\hat{\rho}_\lambda}(x,s) - X_\rho(x,s))\right]
\end{aligned}$$

At the same time we have

$$\begin{aligned}
\dot{X}_{\hat{\rho}_\lambda}(x,s) - \dot{X}_\rho(x,s) &= F_\rho(X_{\hat{\rho}_\lambda},s) - F_\rho(X_\rho,s) + F_{\hat{\rho}_\lambda}(X_{\hat{\rho}_\lambda},s) - F_\rho(X_{\hat{\rho}_\lambda},s) \\
&= \left(\int_\theta \nabla_X f(X_\rho(x,t),\theta)\rho(\theta,t)d\theta\right)(X_{\hat{\rho}_\lambda}(x,s) - X_\rho(x,s)) + o(\lambda) \\
&+ \lambda \int_\theta f(X_{\hat{\rho}_\lambda}(x,s),\theta)(\rho-\nu)(\theta,s)d\theta \\
&= \left(\int_\theta \nabla_X f(X_\rho(x,t),\theta)\rho(\theta,t)d\theta\right)(X_{\hat{\rho}_\lambda}(x,s) - X_\rho(x,s)) + o(\lambda) \\
&+ \lambda\left(\int_\theta \nabla_X f(X_\rho(x,s),\theta)(\rho-\nu)(\theta,s)d\theta\right)(X_{\hat{\rho}_\lambda}(x,s) - X_\rho(x,s)) + o(\lambda) \\
&+ \lambda \int_\theta f(X_\rho(x,s),\theta)(\rho-\nu)(\theta,s)d\theta \\
&= \left(\int_\theta \nabla_X f(X_\rho(x,t),\theta)\rho(\theta,t)d\theta\right)(X_{\hat{\rho}_\lambda}(x,s) - X_\rho(x,s)) \\
&+ \lambda \int_\theta f(X_\rho(x,s),\theta)(\rho-\nu)(\theta,s)d\theta + o(\lambda).
\end{aligned}$$

Here $F_\rho(X,t) = \int_\theta f(X,t)\rho(\theta,t)d\theta$, and the last equality holds because $\|X_{\hat{\rho}_\lambda}(x,s) - X_\rho(x,s)\| \leq \hat{C}e^{\hat{C}}d(\rho_1,\rho_2) = O(\lambda)$. This leads us to

$$\begin{aligned}
&\frac{d}{dt}\left[e^{-\int_0^t \int \nabla_X f(X_\rho(x,t),\theta)\rho(\theta,s)d\theta ds}(X_{\hat{\rho}_\lambda}(x,s) - X_\rho(x,s))\right] \\
&= e^{-\int_0^t \int \nabla_X f(X_\rho(x,t),\theta)\rho(\theta,s)d\theta ds}\left[\dot{X}_{\hat{\rho}_\lambda}(x,s) - \dot{X}_\rho(x,s) - \int_\theta \nabla_X f(X_\rho(x,t),\theta)\rho(\theta,t)d\theta(X_{\hat{\rho}_\lambda}(x,s) - X_\rho(x,s))\right] \\
&= e^{-\int_0^t \int \nabla_X f(X_\rho(x,t),\theta)\rho(\theta,s)d\theta ds}\left[\lambda \int_\theta f(X_\rho(x,s),\theta) + o(\lambda)\right]
\end{aligned}$$

Thus

$$X_{\hat{\rho}_\lambda}(x,1) - X_\rho(x,1) = \int_0^1 \int_\theta e^{\int_t^1 \int \nabla_X f(X_\rho(x,s),\theta)\rho(\theta,s)d\theta ds} f(X_\rho(x,s),\theta)(\rho-\nu)(\theta,t)d\theta dt + o(\lambda)$$

Combining with the definition of the adjoint equation $p(x,t) = p(x,1)e^{\int_t^1 \int \nabla_X f(X_\rho(x,t),\theta)\rho(\theta,t)d\theta dt}$ and $p_\rho(x,1) := \frac{\partial E(x;\rho)}{\partial X_\rho(x,1)} = \big(\langle w_1, X_\rho(x,1)\rangle - y(x)\big)w_1$, we have

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

□

Corollary 1. *(2) For distribution ρ satisfies $\rho(Q_r) = 1$, for any admissible transport plan γ and a vector field $v = \nabla \frac{\delta E}{\delta \rho}$, we have*

$$E(\pi_{\#}\rho) \geq E(\rho) + \int v(y) \cdot (x - y) d\gamma(x, y) + o\left(\left(\int |y - x|^2 d\gamma(x, y)\right)^{1/2}\right).$$

D Detailed Proofs For Landscape Analysis.

Theorem. *(Restate of Theorem 2.) If $E(\rho) > 0$ for some probability distribution $\rho \in \mathcal{P}^2$ which concentrates on one of the nested sets Q_r, then there exists a descend direction $v \in \mathcal{P}^2$ s.t.*

$$\left\langle \frac{\delta E}{\delta \rho}, (\rho - v) \right\rangle > 0$$

Proof. First we lower bound the gradient respect to the feature map $X_\rho(\cdot, t)$ by the loss function to show that changing feature map can always leads to a lower loss. This is observed by (5; 6) where they mean by

Lemma 1. *The norm of the solution to the adjoint equation can be bounded by the loss*

$$\|p_\rho(\cdot, t)\|_\mu^2 \geq e^{-(C_1 + C_2 r)} E(\rho), \qquad \forall\, t \in [0, 1].$$

Proof. By definition,

$$\|p_\rho(\cdot, 1)\| = \|\big(\langle w_1, X_\rho(\cdot, 1)\rangle - y(\cdot)\big) w_1\| = |\langle w_1, X_\rho(\cdot, 1)\rangle - y(\cdot)|,$$

which implies that $\|p_\rho(\cdot, 1)\|_\mu^2 = 2E(\rho)$.

By assumption there exist a constant $C_\rho > 0$ such that

$$\Big| \int \rho(\theta, t) d\theta - \int \rho(\theta, s) d\theta \Big| \leq \|\rho(\cdot, t - s) - \rho(\cdot, s)\|_{BL} \leq C_\rho |t - s|, \ \forall t, s \in [0, 1].$$

Integrating the inequality above with respect to s over $[0, 1]$, and using the fact that $\int_\theta \int_t \rho(\theta, t) = 1$, one obtains that $\int \rho(\theta, t) d\theta \leq 1 + C_\rho \int_0^1 |t - s| ds \leq 1 + \frac{C_\rho}{2}$.

Recall that p_ρ solves the adjoint equation

$$\dot{p}_\rho(x, t) = -p_\rho(x, t) \int \nabla_X f(X_\rho(x, t), \theta) \rho(\theta, t) d\theta \tag{18}$$

where by the assumption on f and the above bound on $\int \rho(\theta, t) d\theta$, we have for any x

$$\| \int \nabla_X f(X_\rho(x, t), \theta) \rho(\theta, t) d\theta \| \leq \sup_{x, \theta} |\nabla_X f(X_\rho(x, t), \theta)| \int_\theta \rho(\theta, t) d\theta \leq (C_1 + C_2 r).$$

It then follows from the Gronwall's inequality that

$$\|p_\rho(\cdot, t)\|_\mu \geq e^{-\int_0^1 \sup_x \| \int \nabla_X f(X_\rho(x,t), \theta) \rho(\theta, t) d\theta \| dt} \|p_\rho(\cdot, 1)\|_\mu \geq e^{-(C_1 + C_2 r)} E(\rho)^{1/2}.$$

The claim of the Lemma then follows by squaring the inequality (and redefining constants C_1 and C_2). □

Thanks to the existence and uniqueness of the solution of the ODE model as stated in Theorem 4, the solution map of the ODE is invertible so that there exists an inverse map $X_{\rho,t}^{-1}$ such that we can construct an inversion function $X_{\rho,t}^{-1}(X_\rho(x, t)) = x$. With $X_{\rho,t}^{-1}$, we define $\hat{p}_\rho(x, t) = p_\rho(X_{\rho,t}^{-1}(x), t)$.

Since $\rho(\theta, t)$ is a probability density, i.e., $\int \int \rho(\theta, t) d\theta dt = 1$, there exists $t_* \in (0, 1)$ such that $\int_\theta \rho(\theta, t_*) d\theta > \frac{1}{2}$. Since $k(x_1, x_2) = f(x_1, x_2)$ is a universal kernel (35), for any $g(x)$ satisfying

that $\|g\|_{\hat{\mu}} < \infty$ for some probability measure $\hat{\mu}$ and for any fixed $\epsilon > 0$, there exists a probability distribution $\delta\hat{\nu} \in \mathcal{P}^2(\mathbb{R}^{d_2})$ such that

$$\|g(x) - \int_\theta f(x,\theta)\delta\hat{\nu}(\theta)d\theta\|_{\hat{\mu}} \leq \epsilon, \tag{19}$$

In particular, in what follows we consider the function $g(x)$ and the measure $\hat{\mu}$ given by

$$g(x) := -\hat{p}(x,t_*) + \frac{1}{\int_\theta \rho(\theta,t_*)d\theta}\int_\theta f(x,\theta)\rho(\theta,t_*)d\theta \text{ and } \hat{\mu} = \hat{\mu}_{\rho,t_*} := X_\rho(\cdot,t_*)_\#\mu.$$

The value of ϵ will be chosen later in the proof. Moreover, we also define the perturbed measure

$$\delta\nu = \left(\delta\hat{\mu}(\theta) - \frac{\rho(\theta,t_*)}{\int_\theta \rho(\theta,t_*)d\theta}\right)\phi(t), \tag{20}$$

where $\phi(t)$ is a smooth non-negative function integrates to 1 and compactly supported in the interval $(0,1)$, so that it is clear that $\delta\nu$ satisfies the regularity assumptions. We will consider the perturbed probability density ν defined as

$$\nu = \rho + \delta r\delta\nu \text{ for some } \delta r > 0.$$

Lemma 2. *The constructed ν with ϵ sufficiently small gives a descent direction of our model with the estimate*

$$\left\langle \frac{\delta E}{\delta\rho}, (\nu-\rho)\right\rangle \leq -\frac{\delta r}{2}e^{-2(C_1+C_2r)}E(\rho) < 0. \tag{21}$$

Proof. An application of the Gronwall inequality to equation 18 implies that

$$p_\rho(x,t_1)p_\rho(x,t_2) \geq e^{-|t_1-t_2|(C_1+C_2r)}\big(p_\rho(x,t_1)^2 \vee p_\rho(x,t_2)^2\big) \tag{22}$$

for all $x \in \mathbb{R}^d, 1 \geq t_2 \geq t_1 \geq 0$.

As a result of equation 20,

$$\begin{aligned}
&\left\langle \frac{\delta E}{\delta\rho}, (\nu-\rho)\right\rangle = \mathbb{E}_{x\sim\mu}\left\langle f(X_\rho(x,t),\cdot))p_\rho(x,\cdot), \delta r\delta\nu\right\rangle \\
&= \delta r\int \mathbb{E}_{x\sim\hat{\mu}_{\rho,t}}\hat{p}_\rho(x,t)\int_\theta f(x,\theta)\delta\nu(\theta,t)d\theta\phi(t)dt \\
&= \delta r\int \mathbb{E}_{x\sim\hat{\mu}_{\rho,t}}\Big[\hat{p}_\rho(x,t)\int_\theta f(x,\theta)\delta\hat{\nu}(\theta)d\theta\Big]\phi(t)dt \\
&\qquad - \delta r\int \mathbb{E}_{x\sim\hat{\mu}_{\rho,t}}\Big[\hat{p}_\rho(x,t)\underbrace{\frac{\int f(x,\theta)\rho(\theta,t_*)d\theta}{\int_\theta \rho(\theta,t_*)d\theta}}_{=g+\hat{p}(x,t_*)}\Big]dt \\
&= \delta r\int \mathbb{E}_{x\sim\hat{\mu}_{\rho,t}}\Big[\hat{p}_\rho(x,t)\Big(\int_\theta f(x,\theta)\delta\hat{\nu}(\theta)d\theta - g(x)\Big)\Big]\phi(t)dt \\
&\qquad - \delta r\int \mathbb{E}_{x\sim\hat{\mu}_{\rho,t}}\Big[\hat{p}_\rho(x,t)\hat{p}(x,t_*)\Big]\phi(t)dt \\
&=: I_1 + I_2.
\end{aligned}$$

The last equation defines I_1 and I_2 which will be estimated separately below.

Thanks to equation 19, for I_1, we have

$$\begin{aligned}
I_1 &\leq \delta r\int \|\hat{p}_\rho(\cdot,t)\|_{\hat{\mu}_{\rho,t}}\Big\|\int_\theta f(x,\theta)\delta\hat{\nu}(\theta)d\theta - g(x)\Big\|_{\hat{\mu}_{\rho,t}}\phi(t)dt \\
&= \delta r\int \|p_\rho(\cdot,t)\|_\mu\Big\|\int_\theta f(x,\theta)\delta\hat{\nu}(\theta)d\theta - g(x)\Big\|_{\hat{\mu}_{\rho,t}}\phi(t)dt \\
&\leq \delta r\int \|p_\rho(\cdot,t)\|_\mu\,\epsilon\sup_x\Big|\frac{d\hat{\mu}_{\rho,t}}{d\hat{\mu}_{\rho,t_*}}\Big|\phi(t)dt \\
&= \delta r\int \|p_\rho(\cdot,t)\|_\mu\,\epsilon\sup_x\big|J_\rho(x;t,t_*)\big|\phi(t)dt,
\end{aligned}$$

where $J_\rho(x;t,s)$ is the Jacobian of the flow at time t with respect to time s assuming starting at x at time 0; which is bounded by the Lipschitz assumption of the f. Thus, we have

$$I_1 \leq C\epsilon\delta r \int \|p_\rho(\cdot,t)\|_\mu \phi(t)dt. \tag{23}$$

Thanks to equation 22, one has

$$\begin{aligned} I_2 &\leq -\delta r \int e^{-|t-t_*|(C_1+C_2 r)} \|\hat{p}_\rho(\cdot,t)\|^2_{\hat{\mu}_{\rho,t}} \phi(t)dt \\ &= -\delta r \int e^{-|t-t_*|(C_1+C_2 r)} \|p_\rho(\cdot,t)\|^2_{\mu} \phi(t)dt \\ &\leq -e^{-(C_1+C_2 r)} \delta r \int \|p_\rho(\cdot,t)\|^2_{\mu} \phi(t)dt. \end{aligned} \tag{24}$$

Combined the above together, and choosing ϵ sufficiently small that the rhs of equation 23 is bounded by a half of the rhs of equation 24 (note that the constants and the integral in the rhs of equation 23 and equation 24 do not depend on ϵ), we arrive at

$$\begin{aligned} I_1 + I_2 &\leq -\frac{1}{2} e^{-(C_1+C_2 r)} \delta r \int \|p_\rho(\cdot,t)\|^2_{\mu} \phi(t)dt \\ &\leq -\frac{1}{2} e^{-(C_1+C_2 r)} \delta r \int e^{-(C_1+C_2 r)} E(\rho) \phi(t)dt \\ &= -\delta r \frac{1}{2} e^{-2(C_1+C_2 r)} E(\rho), \end{aligned}$$

where the last inequality follows from Lemma 1. □

As Lemma 2 illustrated, if the loss $E(\rho)$ is not equal to zero, then we can always find a direction to decrease the loss, which proves Theorem 2. □

E PROPERTIES OF THE LOSS FUNCTION IN THE WASSERSTEIN SPACE.

In this section we analyze the objective function following the theory of the gradient flow developed in (2). First we will prove that our objective function shares the same regularity with the two-layer neural network as shown in (13). Then we will analyze the situation that the gradient flow will converge to the global minima.

REGULARITY IN WASSERSTEIN SPACE

To address the regularity of the Wasserstein gradient flow, following (12), we first analyze the regularity in the ball $F_r = \{E_\rho | \rho \in \mathcal{P}^2, \rho(Q_r) = 1\}$ then extending the result to show that the finite time gradient flow will lie in a ball F_r using the sublinear growth assumption.

Theorem 5. *(Geodesically Semiconvex Properties of F^r in Wasserstein geometry) Further assume that $f(x,\theta)$ have second order smoothness,* i.e. *$f(x,\theta)$ has a smooth Hessian. Then for all $r > 0$, F_r is proper and continuous in W_2 space on its closed domain, Moreover, for $\forall \rho_1, \rho_2 \in \mathcal{P}^2$ and an admissible transport plan γ, denote the interpolation plan in Wasserstein space as $\mu_t^\gamma := ((1-t)\rho_1 + t\rho_2)_\# \gamma$. There exists a $\lambda > 0$ such that the function on the Wasserstein geodesic $t \to F^r(\mu_t^\gamma)$ is differentiable with a $\lambda C(\gamma)$-Lipschitz derivative. Here $C(\gamma)$ is the transport cost $C(\gamma) = \left(\int |y-x|^2 d\gamma(x,y)\right)^{1/2}$*

Proof. To prove the regularity of our objective in Wasserstein Space, we first provide some analysis of the objective function.

Lemma 3. *The gradient of the objective function has the following bound,* i.e.

$$\sup_{\theta \in Q_r} \left\| \frac{\delta E}{\delta \rho}(\theta,t) \right\| = \sup_{\theta \in Q_r} \|\mathbb{E}_{x\sim\mu} f(X_\rho(x,t),\theta)) p_\rho(x,t)\| \leq e^{(C_1+C_2 r)} \sigma_3(\sigma_2 R_1 + R_2 + C_f^r).$$

Proof. First the output of the neural network satisfies

$$\|X_\rho(x,1)\| \le \|X_\rho(x,0)\| + \|\int_0^1 \int_\theta f(X_\rho(x,t),\theta)\rho(\theta,t)d\theta dt\| \le \sigma_2 R_1 + C_f^r,$$

thus $\|p_\rho(x,1)\| := \|\frac{\partial E(x;\rho)}{\partial X_\rho(x,1)}\| = \|\big(\langle w_1, X_\rho(x,1)\rangle - y(x)\big)\| \le \sigma_3(\sigma_2 R_1 + R_2 + C_f^r)$.

At the same time, for the adjoint process $p_\rho(x,t)$ satisfying the adjoint equation, using Gronwall inequality we have, similarly to the proof of Lemma 1

$$\|p_\rho(\cdot,t)\| \le e^{\int_0^1 \|\int_\theta \nabla_X f(X_\rho(x,t),\theta)\rho(\theta,t)d\theta\| dt}\|p_\rho(\cdot,1)\| \le e^{(C_1+C_2 r)}\sigma_3(\sigma_2 R_1 + R_2 + C_f^r). \quad (25)$$

The conclusion then follows as f is bounded on the compact space.

Lemma 4. *The gradient of objective function respect to the feature $X_\rho(x,t)$ is also Lipschitz both in $\mathcal{P}^2$,* i.e. *there exists a constant L_{g_1} satisfies*

$$\sup_{\rho_1\neq\rho_2}\sup_{s\in(0,1)} \frac{\|p_{\rho_1}(x,s) - p_{\rho_2}(x,s)\|}{\|\rho_1-\rho_2\|} \le L_{g_1}.$$

Furthermore, the Frechet derivative $\frac{\delta p_\rho}{\delta \rho}$ exists.

Proof. As proved in Theorem 1, $\|X_{\rho_1}(x,1) - X_{\rho_2}(x,1)\| \le \hat{C}e^{\hat{C}} d_W(\rho_1,\rho_2) \le \frac{\hat{C}e^{\hat{C}}}{R_r^2}\|\rho_1-\rho_2\|$, which leads to $\|p_{\rho_1}(x,1) - p_{\rho_2}(x,1)\| = |(\langle w_1, X_{\rho_1}(x_1,1)\rangle - y(x))| - |(\langle w_1, X_{\rho_2}(x_1,1)\rangle - y(x))| \le \hat{C}e^{\hat{C}} d_W(\rho_1,\rho_2) \le \frac{\hat{C}e^{\hat{C}}}{R_r^2}\|\rho_1-\rho_2\|$. To propagate the estimates to $t\le 1$, we control

$$\begin{aligned}
\|\dot{p}_{\rho_1}(x,s) - \dot{p}_{\rho_2}(x,s)\| &= \Bigg\| \left(\int_\theta \nabla_X f(X_{\rho_1}(x,s),\theta)\rho_1(\theta,s)d\theta\right) p_{\rho_1}(x,s) \\
&\qquad - \left(\int_\theta \nabla_X f(X_{\rho_2}(x,s),\theta)\rho_2(\theta,s)d\theta\right) p_{\rho_2}(x,s)\Bigg\| \\
&\le \left\| \left(\int_\theta \nabla_X f(X_{\rho_1}(x,s),\theta)\rho_1(x,s)d\theta\right)(p_{\rho_1}(x,s) - p_{\rho_2}(x,s))\right\| \\
&\qquad + \left\| \left(\int_\theta \nabla_X f(X_{\rho_2}(x,s),\theta)(\rho_2(x,s)-\rho_1(x,s))d\theta\right) p_{\rho_2}(x,s)\right\| \\
&\le (C_1 + C_2 r)\left(\int \rho_1(\theta,s)d\theta\right)\|p_{\rho_1}(x,s) - p_{\rho_2}(x,s)\| \\
&\qquad + (C_1+C_2 r)\|p_{\rho_2}(x,s)\|\left(\int_\theta (\rho_1(\theta,s)-\rho_2(\theta,s))^2 d\theta\right)^{1/2} \\
&\overset{equation\ 25}{\le} (C_1 + C_2 r)\left(\int \rho_1(\theta,s)d\theta\right)\|p_{\rho_1}(x,s) - p_{\rho_2}(x,s)\| \\
&\qquad + (C_1+C_2 r)e^{(C_1+C_2 r)}\sigma_3(\sigma_2 R_1 + R_2 + C_f^r) \\
&\qquad\qquad \times \left(\int_\theta (\rho_1(\theta,s)-\rho_2(\theta,s))^2 d\theta\right)^{1/2}.
\end{aligned}$$

Introduce the short hand $M := (C_1+C_2r)e^{(C_1+C_2r)}\sigma_3(\sigma_2R_1+R_2+C_f^r)$ and applying the Gronwall inequality, we obtain

$$\begin{aligned}
\|p_{\rho_1}(x,s) - p_{\rho_2}(x,s)\| &\leq \frac{\hat{C}e^{\hat{C}+(C_1+C_2r)\int_0^1\int\rho_1(\theta,s)d\theta ds}}{R_r^2}\|\rho_1-\rho_2\| \\
&\quad + \int_0^1 Me^{(C_1+C_2r)\int_t^1(\int\rho_1(\theta,s)d\theta)ds}\left(\int_\theta(\rho_1(\theta,s)-\rho_2(\theta,s))^2\,d\theta\right)^{1/2}dt \\
&\leq \frac{\hat{C}e^{\hat{C}+(C_1+C_2r)}}{R_r^2}\|\rho_1-\rho_2\| \\
&\quad + Me^{(C_1+C_2r)\int_0^1\int\rho_1(\theta,s)d\theta ds}\int_0^1\left(\int_\theta(\rho_1(\theta,s)-\rho_2(\theta,s))^2\,d\theta\right)^{1/2}dt \\
&\leq \left(\frac{\hat{C}e^{\hat{C}+(C_1+C_2r)}}{R_r^2} + Me^{(C_1+C_2r)}\right)\|\rho_1-\rho_2\|,
\end{aligned}$$

where last inequality follows from Jensen's inequality

$$\int_0^1\left(\int_\theta(\rho_1(\theta,s)-\rho_2(\theta,s))^2\,d\theta\right)^{1/2}dt \leq \left(\int_0^1\int_\theta(\rho_1(\theta,s)-\rho_2(\theta,s))^2\,d\theta dt\right)^{1/2} = \|\rho_1-\rho_2\|.$$

The existence of the Frechet derivative follows from the smoothness of the activation function, in particuar the assumption that the Hessian is bounded. □

Now we show the continuity of the objective function in the Wasserstein Space. By denoting $h(t) = F^{(r)}(\mu_t^\gamma)$

$$\begin{aligned}
h'(\tau) &= \frac{d}{dt}E^{(r)}(\mu_t^\gamma) \\
&= \left\langle \frac{\delta E}{\delta\rho}[\mu_\tau^\gamma], \frac{d}{d\tau}\mu_\tau^\gamma\right\rangle \\
&= \int d\frac{\delta E}{\delta\rho}[\mu_\tau^\gamma]((1-\tau)(\theta_1,t_1)+\tau(\theta_2,t_2))((\theta_1,t_1)-(\theta_2,t_2))d\gamma((\theta_1,t_1),(\theta_2,t_2)). \quad (26)
\end{aligned}$$

For any $\tau_1,\tau_2\in[0,1]$, we have $h'(\tau_1)-h'(\tau_2) = I+J$ with

$$\begin{aligned}
I &= \int d\frac{\delta E}{\delta\rho}[\mu_{\tau_1}^\gamma]((1-\tau_1)(\theta_1,t_1)+\tau_1(\theta_2,t_2))((\theta_1,t_1)-(\theta_2,t_2))d\gamma((\theta_1,t_1),(\theta_2,t_2)) \\
&\quad - \int d\frac{\delta E}{\delta\rho}[\mu_{\tau_2}^\gamma]((1-\tau_1)(\theta_1,t_1)+\tau_1(\theta_2,t_2))((\theta_1,t_1)-(\theta_2,t_2))d\gamma((\theta_1,t_1),(\theta_2,t_2)), \quad (27) \\
J &= \int d\frac{\delta E}{\delta\rho}[\mu_{\tau_2}^\gamma]((1-\tau_1)(\theta_1,t_1)+\tau_1(\theta_2,t_2))((\theta_1,t_1)-(\theta_2,t_2))d\gamma((\theta_1,t_1),(\theta_2,t_2)) \\
&\quad - \int d\frac{\delta E}{\delta\rho}[\mu_{\tau_2}^\gamma]((1-\tau_2)(\theta_1,t_1)+\tau_2(\theta_2,t_2))((\theta_1,t_1)-(\theta_2,t_2))d\gamma((\theta_1,t_1),(\theta_2,t_2)). \quad (28)
\end{aligned}$$

For I, we have

$$|I| \leq L_{g_1}\cdot 2r\|\mu_{\tau_1}^\gamma - \mu_{\tau_2}^\gamma\| \leq 2rL_{g_1}C_2(\gamma)|\tau_1-\tau_2|. \quad (29)$$

Similarly, for J we have

$$|J| \leq L_{g_1}|\tau_1-\tau_2|\int((\theta_1,t_1)-(\theta_2,t_2))^2d\gamma((\theta_1,t_1),(\theta_2,t_2)) = L_{g_1}C_2^2(\gamma)|\tau_1-\tau_2|. \quad (30)$$

Finally, combining the estimates for I and J shows that $h'(\tau)$ is Lipschitz continuous.

□

With the proved regularity, we are in position to prove the well-posedness of Wasserstein gradient flows

Corollary 2. *((2) Thm. 11.2.1) There exists a $\Delta t > 0$ such that there exists a unique solution $\{\rho_s\}_{s\in[0,\Delta t]}$ of the Wasserstein gradient flow $\frac{\partial_{(\theta,t)}\rho}{\partial s} = div_{(\rho,t)}(\rho\nabla_{(\rho,t)}\frac{\delta E}{\delta\rho})$ starting from any $\mu_0 \in \mathcal{P}_2$ concentrated on Q_r.*

CONVERGENCE RESULTS FOR THE WASSERSTEIN GRADIENT FLOW

We move on to prove that the wasserstein gradient flow will achieve the global optimum with a support related assumption of the stationary point. Following (12), we introduce an assumption of the homogeneity of the activation function which is a central property for our global convergence results

Homogeneity. A function f between vector spaces is *positively p-homogeneous* when for all $\lambda > 0$ and argument x, $f(\lambda x) = \lambda^p f(x)$. We assume that the functions $f(X,\theta)$ that constitute the residual block obtained through the lifting share the property of being positively p-homogeneous($p > 0$) in the variable θ. As (12) remarked the Relu function is a 1-homogeneity function which leads to the 2-homogeneity respect to θ of $f(X,\theta)$ when the residual block is implemented via a two-layer neural network.

Theorem. *(Restate of Theorem 3.) When the residual block (X,θ) is positively p-homogeneous respective to θ. Let $(\rho_s)_{s\geq 0}$ be the solution of the the Wasserstein gradient $\frac{\partial_{(\theta,t)}\rho}{\partial s} = div_{(\rho,t)}(\rho\nabla_{(\rho,t)}\frac{\delta E}{\delta\rho})$ of our mean-field model (2). If $(\rho_s)_{s\geq 0}$ converge to ρ_∞ in W_2 and ρ^* concentrates in a ball $B(0,r_b)$ and separates the spheres $r_a\mathbb{S}^{d-1}\times[0,1]$ and $r_b\mathbb{S}^{d-1}\times[0,1]$. Then ρ^* is the global minimum satisfies $E(\rho_\infty) = 0$.*

Proof. First we use the conclusion of (36) which characterize the condition of the stationary points in the Wasserstein space.

Lemma 5. *(36) The steady state ρ of the Wasserstein gradient flow $\frac{\partial_{(\theta,t)}\rho}{\partial s} = div_{(\rho,t)}(\rho\nabla_{(\rho,t)}\frac{\delta E}{\delta\rho})$ must satisfy $\nabla_{(\theta,t)}\frac{\delta E}{\delta\rho}(\theta,t) = 0, \rho a.e.$.*

Then we will use the homogeneity of the activation function and the separation property of the support of ρ_∞ to prove $\nabla_{(\theta,t)}\frac{\delta E}{\delta\rho}|_{\rho=\rho_\infty}(\theta,t) = 0$, a.e.. Due to the separation assumption of the support of the distribution, for any $(\theta,t)\in\mathbb{R}^{d_1\times d_1}\times[0,1]$ there exist $r > 0$ such that $(r\theta,t)\in\text{supp}(\rho_\infty)$. For $\frac{\delta E}{\delta\rho}(r\theta,t) = \mathbb{E}_{x\sim\mu}f(X_\rho(x,t),r\theta))p_\rho(x,t) = r^p\mathbb{E}_{x\sim\mu}f(X_\rho(x,t),\theta))p_\rho(x,t) = r^p\frac{\delta E}{\delta\rho}(r\theta,t)$, which also leads to $\nabla_{(\theta,t)}\frac{\delta E}{\delta\rho}(r\theta,t) = r^p\nabla_{(\theta,t)}\frac{\delta E}{\delta\rho}(\theta,t)$. For $\nabla_{(\theta,t)}\frac{\delta E}{\delta\rho}|_{\rho=\rho_\infty}(\theta,t) = 0, \rho_\infty$a.e., we know that $\nabla_{(\theta,t)}\frac{\delta E}{\delta\rho}|_{\rho=\rho_\infty}(\theta,t) = 0$, a.e..

For we have $\nabla_{(\theta,t)}\frac{\delta E}{\delta\rho}|_{\rho=\rho_\infty}(\theta,t) = 0$, a.e., which means that the gradient is constant function that is $\frac{\delta E}{\delta\rho}|_{\rho=\rho_\infty}(\theta,t) = c$

If $E(\rho_\infty)\neq 0$, according to Theorem 3, there exists another distribution $\mu\in\mathcal{P}^2$ s.t.

$$\left\langle \frac{\delta E}{\delta\rho}|_{\rho=\rho_\infty}, (\rho-\mu)\right\rangle > 0.$$

However $\left\langle \frac{\delta E}{\delta\rho}|_{\rho=\rho_\infty}, (\rho-\mu)\right\rangle = c\left(\int\rho(\theta,t)d\theta dt - \int\rho(\theta,t)d\theta dt\right) = 0$ for the mass conservation of the probability measure, which leads to a contradiction. Thus the converged probability measure must satisfies $E(\rho_\infty) = 0$, which states that it is a global optimal.

□

□

F AN ANALOG OF RESNET BEHAVES LIKE ENSEMBE OF SHALLOW NETWORKS

In this section, we briefly explain the intuition behind our analysis, *i.e.* deep residual network can be approximated by a two-layer neural network. (47) introduced an unraveled view of the ResNets and showed that deep ResNets behave like ensembles of shallow models. First, we offer a formal derivation to reveal how to make connection between a deep ResNet and a two-layer neural network. The first residual block is formulated as

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

By Taylor expansion, the second layer output is given by

$$\begin{aligned}
X^2 &= X^1 + \frac{1}{L}\int_{\theta^1} \sigma(\theta^1 X^1)\rho^1(\theta^1)d\theta^1 \\
&= X^0 + \frac{1}{L}\int_{\theta^0} \sigma(\theta^0 X^0)\rho^0(\theta^0)d\theta^0 \\
&+ \int_{\theta^1} \sigma(\theta^1(X^0 + \frac{1}{L}\int_{\theta^0} \sigma(\theta^0 X^0)\rho^0(\theta^0)d\theta^0))\rho^1(\theta^1)d\theta^1 \\
&= X^0 + \frac{1}{L}\int_{\theta^0} \sigma(\theta^0 X^0)\rho^0(\theta^0)d\theta^0 \\
&+ X^0 + \frac{1}{L}\int_{\theta^1} \sigma(\theta^1 X^0)\rho^1(\theta^1)d\theta^1 \\
&+ \frac{1}{L^2}\int_{\theta_1} \nabla\sigma(\theta^1 X^0)\theta^1(\int_{\theta^0} \sigma(\theta^0 X^0)\rho^0(\theta^0)d\theta^0)\rho^1(\theta^1)d\theta^1 \\
&+ h.o.t.
\end{aligned}$$

Iterating this expansion gives rise to

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

Here we only keep the terms that are at most quadratic in ρ. A similar derivation shows that at order k in ρ there are $\binom{L}{k}$ terms with coefficient $\frac{1}{L^k}$ each. This implies that the k-th order term in ρ decays as $O(\frac{1}{k!})$, suggesting that one can approximate a deep network by the keeping a few leading orders.

F.1 DISCUSSION AND FUTURE DIRECTION

Our work gives qualitative analysis of the loss landscape of a deep residual network and shows that its gradient differs from the gradient of a two-layer neural network by at most a bounded factor when the loss is at the same level. This indicates that the deep residual network's landscape may not be much more complicate than a two-layer network, which inspires us to formulate a mean-field analysis framework for deep residual network and suggests a possible framework for the optimization of the deep networks beyond the kernel regime. (49) has shown that deep residual network may not be better than a linear model in terms of optimization, but our work suggests that this is caused by the lack of overparameterization. In the highly overparameterization regime, the landscape of deep ResNet can still be nice. Based on the initiation and framework proposed in our paper, there are several interesting directions related to understanding and improving the residual networks.

Firstly, to ensure the full support assumption, we can consider extending the neural birth-death (40; 11) to deep ResNets. Neural birth-death dynamics considers the gradient flow in the Wasserstein-Fisher-Rao space(14) rather than the Wasserstein space and ensures convergence. Another approach for removing the full support assumption is to follow (12) and utilize the homogeneity properties of the active function. One drawback of this approach is the introduction of an additional Sard-type regularity assumption, which is hard to check both theoretically and practically.

Secondly, as shown in the derivation in Section **??**, the two-layer network approximation is just the lowest order approximation to the deep residual network and it is interesting to explore the higher order terms.

G EXPERIMENT

In this section, we aim to show that our algorithm is not only designed from theoretical consideration but also realizable on practical datasets and network structures. We implement our algorithm for ResNet/ResNeXt on CIFAR 10/100 datasets and demonstrate that our "mean-field training" method consistently outperforms the vanilla stochastic gradient descent.

Our new algorithm only introduces n parameters, as n is the depth which is around 100 in practice, thus the number of extra parameters is negligible comparing to the 1M+ parameter number typically used in usual ResNet architectures. The sorting of $\{\tau_i\}_{i=1}^n$ also induces negligible cost per step.

We also remark that the flexibility of τ^ℓ can be also viewed as an adaptive time marching scheme of the ODE model for x, as $\tau^\ell - \tau^{\ell-1}$ can be understood as the time step in the Euler discretization. Since the parameters $\{\tau^\ell\}$ are learned from data, as a by-product, our scheme also naturally yields a data-adaptive discretization scheme.

As the number of particles n becomes large, the expected time evolution of ρ_n should be close to the gradient flow $\frac{\partial_{(\theta,t)}\rho}{\partial s} = \mathrm{div}_{(\rho,t)}(\rho\nabla_{(\rho,t)}\frac{\delta E}{\delta\rho})$. The rigorous proof of this is however non-trivial, which will be left for future works.

Implementation Details. On CIFAR, we follow the simple data augmentation method in (22; 23) for training: 4 pixels are padded on each side, and a 32×32 crop is randomly sampled from the padded image or its horizontal flip. For testing, we only evaluate the single view of the original 32×32 image. For the experiments of ResNet on CIFAR, we adopt the original design of the residual block in (22), i.e. using a small two-layer neural network as the residual block, whose layered structure is bn-relu-conv-bn-relu-conv. We start our networks with a single 3×3 conv layer, followed by 3 residual blocks, a global average pooling, and a fully-connected classifier. Parameters are initialized following the method introduced by (21). Mini-batch SGD is used to optimize the parameters with a batch size of 128. During training, we apply a weight decay of 0.0001 for ResNet and 0.0005 for ResNeXt, and a momentum of 0.9. For ResNet on CIFAR10 (CIFAR100), we start with the learning rate of 0.1, divide it by 10 at 80 (150) and 120 (225) epochs and terminate the training at 160 (300) epochs. For ResNeXt on CIFAR100, we start with the learning rate of 0.1 and divide it by 10 at 150 and 225 epochs, and terminate the training at 300 epochs. We would like to mention that here the ResNeXt is a preact version which is different from the original (48). This difference leads to a small performance drop on the final result. For each model and dataset, we report the average test accuracy over 3 runs in Table 5.1.