Appendix of "A Mean Field Analysis Of Deep ResNet And Beyond: Towards Provable Optimization Via Overparameterization From Depth"

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 supplementary material is organized as following

- Appendix A: Properties of our continuous model.
- Appendix B: Detailed Proofs For Landscape Analysis.
- Appendix C: Properties of the loss function in the Wasserstein space.

First we introduce the notations and assumptions we used in the appendix.

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 $\|f\|_\mu = \mathbb{E}_{x\sim\mu}[f(x)^\top f(x)]$.

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 $\int \varphi \mathrm{d}\mu_n \to \int \varphi \mathrm{d}\mu$. For sequences which are bounded in total variation norm, this is equivalent to the convergence in Bounded Lipschitz norm. The latter is defined, for $\mu \in \mathcal{M}(\mathbb{R}^d)$, as

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

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

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 exists 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 (Micchelli et al., 2006), 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 (Chizat & Bach, 2018)) There exists a family $\{Q_r\}_{r>0}$ of nested nonempty closed convex subsets of Ω that satisfies:*
   - *$\{u \in \Omega \mid \mathrm{dist}(u, Q_r) \le r'\} \subset Q_{r+r'}$ for all $r, r' > 0$.*
   - *There exist constants $C_1, C_2 > 0$ such that*

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

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

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

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

A Properties of our continuous model.

Our model is aimed to minimize the l_2 loss function

$$E(\rho) = \mathbb{E}_{x \sim \mu} \frac{1}{2} \big(\langle w_1, X_\rho(x, 1) \rangle - y(x) \big)^2. \tag{A.1}$$

over parameter distributions $\rho(\theta, t)$ for θ in a compact set Ω and $t \in [0, 1]$. Here $X_\rho(x, t)$ is the solution of the ODE

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

Theorem 1. *(Well-posedness of the Forward Model) 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* (A.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*

$$\|X_{\rho_1}(x, 1) - X_{\rho_2}(x, 1)\| < C W_2(\rho_1, \rho_2), \tag{A.3}$$

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

$$F_\rho(X, t) = \int_\theta f(X, t) \rho(\theta, t) d\theta. \tag{A.4}$$

Then, the ODE (A.2) becomes

$$\dot{X}_\rho(x, t) = F_\rho(X_\rho(x, t), t), \tag{A.5}$$

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

$$\|F_\rho(X, t)\| \le C_f^r \left| \int_\theta \rho(\theta, t) d\theta \right| < C_f^r C_\rho. \tag{A.6}$$

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 (A.5) 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

$$\begin{aligned}
\Delta(x,t) &= \left\| \int_0^t \dot{X}_{\rho_1}(x,s) - \dot{X}_{\rho_2}(x,s) ds \right\| \\
&= \left\| \int_0^t F_{\rho_1}(X_{\rho_1}, s) - F_{\rho_1}(X_{\rho_2}, s) ds + \int_0^t F_{\rho_1}(X_{\rho_2}, s) - F_{\rho_2}(X_{\rho_2}, s) ds \right\| \\
&\le \int_0^t \|F_{\rho_1}(X_{\rho_1}, s) - F_{\rho_1}(X_{\rho_2}, s)\| ds + \left\| \int_0^t F_{\rho_1}(X_{\rho_2}, s) - F_{\rho_2}(X_{\rho_2}, s) ds \right\|. \qquad \text{(A.7)}
\end{aligned}$$

Let $C_m = \max\{C_{\rho_1}, C_{\rho_2}\}$. For the first term in (A.7), 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

$$\|F_{\rho_1}(X_{\rho_1}, s) - F_{\rho_1}(X_{\rho_2}, s)\| \le (C_1 + C_2 X_0 + C_2 C_f^r C_m) C_m \Delta(x,s). \qquad \text{(A.8)}$$

For the second term of (A.7), we have

$$\left\| \int_0^t F_{\rho_1}(X_{\rho_2}, s) - F_{\rho_2}(X_{\rho_2}, s) ds \right\| = \left\| \int_0^t \int_\theta f(X_{\rho_2}, \theta)(\rho_1(\theta, s) - \rho_2(\theta, s)) d\theta ds \right\|. \qquad \text{(A.9)}$$

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, $f(X,\theta)$ 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

$$\left\| \int_0^t \int_\theta f(X_{\rho_2}, \theta)(\rho_1(\theta, s) - \rho_2(\theta, s)) d\theta ds \right\| \le C W_2(\rho_1, \rho_2). \qquad \text{(A.10)}$$

Finally, by defining

$$\hat{C} = \max\{(C_1 + C_2 X_0 + C_2 C_f^r C_m) C_m, C\}, \qquad \text{(A.11)}$$

we have by (A.7)

$$\Delta(x,t) \le \int_0^t \hat{C} \Delta(x,t) + \hat{C} W_2(\rho_1, \rho_2). \qquad \text{(A.12)}$$

Applying the Gronwall's inequality gives

$$\Delta(x,t) \le \hat{C} e^{\hat{C}t} W_2(\rho_1, \rho_2), \qquad \text{(A.13)}$$

and specifically for $t = 1$ we have

$$\|X_{\rho_1}(x,1) - X_{\rho_2}(x,1)\| \le \hat{C} e^{\hat{C}} W_2(\rho_1, \rho_2). \qquad \text{(A.14)}$$

□

Theorem 2. *(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 $\nu \in \mathcal{P}^2$ 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 1 (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

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

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

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

Corollary 2.1. *(Ambrosio et al., 2008) 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*

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

B Detailed Proofs For Landscape Analysis.

Theorem 3. *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 (Bartlett et al., 2018, 2019) 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

$$\dot{p}_\rho(x,t) = -p_\rho(x,t) \int \nabla_X f(X_\rho(x,t),\theta)\rho(\theta,t)d\theta \tag{B.1}$$

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 1, 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 (Micchelli et al., 2006), 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

$$\|g(x) - \int_\theta f(x,\theta)\delta\hat{\nu}(\theta)d\theta\|_{\hat{\mu}} \leq \epsilon, \tag{B.2}$$

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

$$\delta\nu = \left(\delta\hat{\mu}(\theta) - \frac{\rho(\theta,t_*)}{\int_\theta \rho(\theta,t_*)d\theta} \right) \phi(t), \tag{B.3}$$

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*

$$\left\langle \frac{\delta E}{\delta \rho}, (\nu - \rho) \right\rangle \leq -\frac{\delta r}{2} e^{-2(C_1+C_2 r)} E(\rho) < 0. \tag{B.4}$$

Proof. An application of the Gronwall inequality to (B.1) implies that

$$p_\rho(x,t_1)p_\rho(x,t_2) \geq e^{-|t_1-t_2|(C_1+C_2r)}\big(p_\rho(x,t_1)^2 \vee p_\rho(x,t_2)^2\big) \tag{B.5}$$

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

As a result of (B.3),

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

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

Thanks to (B.2), 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

$$I_1 \leq C\epsilon\delta r \int \|p_\rho(\cdot,t)\|_\mu \phi(t)dt. \tag{B.6}$$

Thanks to (B.5), one has

$$\begin{aligned}
I_2 &\leq -\delta r \int e^{-|t-t_*|(C_1+C_2r)}\|\hat{p}_\rho(\cdot,t)\|^2_{\hat{\mu}_{\rho,t}}\phi(t)dt \\
&= -\delta r \int e^{-|t-t_*|(C_1+C_2r)}\|p_\rho(\cdot,t)\|^2_{\mu}\phi(t)dt \\
&\leq -e^{-(C_1+C_2r)}\delta r \int \|p_\rho(\cdot,t)\|^2_\mu\phi(t)dt.
\end{aligned} \tag{B.7}$$

Combining the above together, and choosing ϵ sufficiently small that the right-hand-side of (B.6) is bounded by a half of the right-hand-side of (B.7) (note that the constants and the integral in the right-hand-side of (B.6) and (B.7) do not depend on ϵ), we arrive at

$$\begin{aligned} I_1 + I_2 &\leq -\frac{1}{2}e^{-(C_1+C_2r)}\delta r \int \|p_\rho(\cdot,t)\|_\mu^2 \phi(t)dt \\ &\leq -\frac{1}{2}e^{-(C_1+C_2r)}\delta r \int e^{-(C_1+C_2r)} E(\rho)\phi(t)dt \\ &= -\delta r \frac{1}{2}e^{-2(C_1+C_2r)}E(\rho), \end{aligned}$$

where the last inequality follows from Lemma 1. □

Now we go back to the proof of Theorem 3, as Lemma 2 illustrates, if the loss $E(\rho)$ is not equal to zero, then we can always find a direction to decrease the loss, this complete the proof. □

C 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 (Ambrosio et al., 2008). First we will prove that our objective function shares the same regularity with the two-layer neural network as shown in (Chizat et al., 2019). Then we will analyze the stationary solution of the gradient flow, and show that they are given by global minima.

Regularity in the Wasserstein Space

To address the regularity of the Wasserstein gradient flow, following (Chizat & Bach, 2018), we first analyze the regularity of E restricted to the set $\{\rho \mid \rho \in \mathcal{P}^2, \rho(Q_r) = 1\}$, to make this explicit, we denote the functional F_r as

$$F_r(\rho) = \begin{cases} E(\rho), & \text{if } \rho(Q_r) = 1; \\ \infty, & \text{otherwise.} \end{cases}$$

Theorem. *(Geodesically semiconvex property 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 the 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.

$$\sup_{\theta \in Q_r} \left\| \frac{\delta E}{\delta \rho}(\theta, t) \right\| = \sup_{\theta \in Q_r} \|\mathbb{E}_{x\sim\mu} f(X_\rho(x,t),\theta)) p_\rho(x,t)\| \leq e^{(C_1+C_2r)}\sigma_3(\sigma_2 R_1 + R_2 + C_f^r).$$

Proof. First the output of the neural network satisfies

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

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

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

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

Lemma 4. *The gradient of the objective function with respect to the feature $X_\rho(x,t)$ is Lipschitz 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

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

Introduce the short hand $M := (C_1 + C_2 r)e^{(C_1+C_2 r)}\sigma_3(\sigma_2 R_1 + R_2 + C_f^r)$ and applying the Gronwall

inequality, we obtain

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

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 particular the assumption that the Hessian is bounded. □

Now we show the continuity of the objective function in the Wasserstein space. By denoting $h(\tau) = F_r(\mu_\tau^\gamma)$

$$\begin{aligned}
h'(\tau) &= \frac{d}{d\tau}F_r(\mu_\tau^\gamma) \\
&= \left\langle \frac{\delta E}{\delta \rho}[\mu_\tau^\gamma], \frac{d}{d\tau}\mu_\tau^\gamma \right\rangle \\
&= \int d\frac{\delta E}{\delta \rho}[\mu_\tau^\gamma]((1-\tau)(\theta_1,t_1)+\tau(\theta_2,t_2))((\theta_1,t_1)-(\theta_2,t_2))d\gamma((\theta_1,t_1),(\theta_2,t_2)). \qquad \text{(C.2)}
\end{aligned}$$

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

$$\begin{aligned}
I &= \int d\frac{\delta E}{\delta \rho}[\mu_{\tau_1}^\gamma]((1-\tau_1)(\theta_1,t_1)+\tau_1(\theta_2,t_2))((\theta_1,t_1)-(\theta_2,t_2))d\gamma((\theta_1,t_1),(\theta_2,t_2)) \\
&\quad - \int d\frac{\delta E}{\delta \rho}[\mu_{\tau_2}^\gamma]((1-\tau_1)(\theta_1,t_1)+\tau_1(\theta_2,t_2))((\theta_1,t_1)-(\theta_2,t_2))d\gamma((\theta_1,t_1),(\theta_2,t_2)), \qquad \text{(C.3)} \\
J &= \int d\frac{\delta E}{\delta \rho}[\mu_{\tau_2}^\gamma]((1-\tau_1)(\theta_1,t_1)+\tau_1(\theta_2,t_2))((\theta_1,t_1)-(\theta_2,t_2))d\gamma((\theta_1,t_1),(\theta_2,t_2)) \\
&\quad - \int d\frac{\delta E}{\delta \rho}[\mu_{\tau_2}^\gamma]((1-\tau_2)(\theta_1,t_1)+\tau_2(\theta_2,t_2))((\theta_1,t_1)-(\theta_2,t_2))d\gamma((\theta_1,t_1),(\theta_2,t_2)). \qquad \text{(C.4)}
\end{aligned}$$

For I, we have

$$\begin{aligned}
|I| &\leq L_{g_1}\cdot 2r\|\mu_{\tau_1}^\gamma - \mu_{\tau_2}^\gamma\| \\
&\leq 2rL_{g_1}C_2(\gamma)|\tau_1-\tau_2|. \qquad \text{(C.5)}
\end{aligned}$$

Similarly, for J we have

$$|J| \leq L_{g_1}|\tau_1 - \tau_2| \int ((\theta_1, t_1) - (\theta_2, t_2))^2 d\gamma$$
$$= L_{g_1} C_2^2(\gamma)|\tau_1 - \tau_2|. \tag{C.6}$$

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

With the proved regularity, the short time well-posedness of Wasserstein gradient flow is a corollary of Theorem 11.2.1 of (Ambrosio et al., 2008).

Corollary 3.1. *There exists a $T_{\max}$ such that there exists a unique solution $\{\rho_s\}_{s\in[0,T_{\max}]}$ to 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 stationary point of the Wasserstein gradient flow achieves the global optimum with a support related assumption. Following (Chizat & Bach, 2018), we introduce an assumption of the homogeneity of the activation function which is a central requirement 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 (Chizat & Bach, 2018) 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 4. *When the residual block $f(X, \theta)$ 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 (A.2). Consider a stationary solution to the gradient flow ρ_∞ which concentrates in one of the nested sets Q_r 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 a global minimum satisfies $E(\rho_\infty) = 0$.*

Proof. First we use the conclusion of (Nitanda & Suzuki, 2017) which characterize the condition of the stationary points in the Wasserstein space, which concludes that the steady state ρ_∞ of the Wasserstein gradient flow

$$\frac{\partial_{(\theta,t)}\rho}{\partial s} = \text{div}_{(\rho,t)}(\rho\nabla_{(\rho,t)}\frac{\delta E}{\delta \rho})$$

must satisfy $\nabla_{(\theta,t)}\frac{\delta E}{\delta \rho}|_{\rho_\infty} = 0, \rho_\infty$-a.e.

We will use the homogeneity of the activation function and the separation property of the support of ρ_∞ to further prove that $\nabla_{(\theta,t)}\frac{\delta E}{\delta \rho}|_{\rho=\rho_\infty} = 0$, a.e. (i.e., it also vanishes outside the support of ρ_∞, which might not be the full parameter space).

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 exists $r > 0$ such that $(r\theta, t) \in \text{supp}(\rho_\infty)$. Due to the homogeneity assumption, we have

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

which leads to $\nabla_{(\theta,t)}\frac{\delta E}{\delta \rho}(r\theta, t) = r^p\nabla_{(\theta,t)}\frac{\delta E}{\delta \rho}(\theta, t)$. Thus, since $\nabla_{(\theta,t)}\frac{\delta E}{\delta \rho}|_{\rho=\rho_\infty} = 0, \rho_\infty$-a.e., we know that $\nabla_{(\theta,t)}\frac{\delta E}{\delta \rho}|_{\rho=\rho_\infty} = 0$, a.e. This further implies that the differential is a constant $\frac{\delta E}{\delta \rho}|_{\rho=\rho_\infty} \equiv c$.

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

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

However $\left\langle \frac{\delta E}{\delta \rho}|_{\rho=\rho_\infty}, (\rho-\nu) \right\rangle = c\left(\int \rho(\theta,t)d\theta dt - \int \nu(\theta,t)d\theta dt\right) = 0$ due to the normalization of the probability measure. This leads to a contradiction. Thus the stationary solution measure must satisfy $E(\rho_\infty) = 0$, which means that it is a global optimum. □

References

Ambrosio, L., Gigli, N., and Savaré, G. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

Bartlett, P. L., Evans, S. N., and Long, P. M. Representing smooth functions as compositions of near-identity functions with implications for deep network optimization. *arXiv preprint arXiv:1804.05012*, 2018.

Bartlett, P. L., Helmbold, D. P., and Long, P. M. Gradient descent with identity initialization efficiently learns positive-definite linear transformations by deep residual networks. *Neural computation*, 31(3): 477–502, 2019.

Chizat, L. and Bach, F. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in neural information processing systems*, pp. 3036–3046, 2018.

Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. 2019.

Micchelli, C. A., Xu, Y., and Zhang, H. Universal kernels. *Journal of Machine Learning Research*, 7(Dec): 2651–2667, 2006.

Nitanda, A. and Suzuki, T. Stochastic particle gradient descent for infinite ensembles. *arXiv preprint arXiv:1712.05438*, 2017.