TRANSFORMERS LEARN NONLINEAR FEATURES IN CONTEXT

Juno Kim[1,2]* **Taiji Suzuki**[1,2]
[1]The University of Tokyo, Tokyo, Japan [2]Center for Advanced Intelligence Project, RIKEN
*junokim@g.ecc.u-tokyo.ac.jp

ABSTRACT

Existing theoretical studies on how in-context learning arises are limited to the dynamics of a single layer of attention trained on linear regression tasks. In this paper, we study the optimization of a Transformer consisting of a fully connected layer followed by a linear attention layer. The MLP acts as a common nonlinear representation or feature map, greatly enhancing the power of in-context learning. We prove in the mean-field and two-timescale limit that the infinite-dimensional loss landscape for the parameter distribution becomes quite benign. We also analyze the second-order stability of mean-field dynamics and show that Wasserstein gradient flow almost always avoids saddle points. Furthermore, we establish novel methods for obtaining concrete improvement rates both away from and near critical points.

1 INTRODUCTION

In-context learning (ICL) refers to the capacity of a pretrained model to solve previously unseen tasks based on example prompts without further tuning. A line of research initiated by Garg et al. (2022) has sought to understand the theory behind ICL from a function class perspective. Studies have shown that certain Transformers are capable of implementing statistical learning algorithms such as gradient descent (GD) in context (von Oswald et al., 2023; Akyürek et al., 2023; Bai et al., 2023). In particular, Guo et al. (2023) analyze learning *with representations* where MLP layers act as transformations on top of which ICL is performed, achieving near-optimal performance. Other works have analyzed how ICL emerges from the training dynamics of Transformers (Zhang et al., 2023a; Huang et al., 2023; Ahn et al., 2023). However, they are limited to models consisting of only a single attention layer due to the complicated dynamics and thus can only explain ICL of linear functions. Hence the following question at the intersection of the two approaches remains unsolved:

How does in-context learning with nonlinear representations (features) arise in Transformers with MLP layers, optimized via gradient descent?

In this paper, we study a Transformer consisting of a 2-layer MLP followed by a linear self-attention (LSA) layer trained on linear transformations of a feature representation. Contrary to existing approaches which attempt to solve for exact attention dynamics, we focus on the loss landscape faced by the MLP and show in the mean-field limit that all critical points are either global minima or saddle points. We also formally prove that mean-field dynamics (MFD) 'almost always' avoids saddle points, explaining how the MLP can find globally optimal features, and further derive concrete improvement rates. This is also of technical interest as the first analysis of *nonconvex* MFD around saddle points. We develop many results for general functionals and derive an application to 3-layer networks. Theoretical preliminaries and related works are given in Appendix A, proofs and additional results for Sections 2-5 are given throughout Appendices B-E, and experiments are detailed in Appendix F.

2 IN-CONTEXT FEATURE LEARNING

We formally define our simplified MLP-LSA Transformer. We switch the ordering as in Guo et al. (2023) to view attention as a mechanism to exchange feature information encoded into the MLP layer. Let $\mathcal{D}_{\mathscr{X}}$ be a distribution over the input space $\mathscr{X} \subseteq \mathbb{R}^d$ and $\mathscr{T}$ a class of functions $\mathscr{X} \to \mathbb{R}$ with distribution $\mathcal{D}_{\mathscr{T}}$. For each prompt, we generate a new task $f \sim \mathcal{D}_{\mathscr{T}}$ and a batch of n example and one query input-output pairs $(\boldsymbol{x}_i, y_i)_{i=1}^n, (\boldsymbol{x}_{\text{qr}}, y_{\text{qr}})$ where $\boldsymbol{x}_i, \boldsymbol{x}_{\text{qr}} \sim \mathcal{D}_{\mathscr{X}}$ are i.i.d. and $y_i = f(\boldsymbol{x}_i)$.

MLP layer. A vector-valued neuron with parameter $\theta = (\boldsymbol{a}, \boldsymbol{w})^\top \in \Theta \subseteq \mathbb{R}^k \times \mathbb{R}^d$ and activation $\sigma : \mathbb{R} \to \mathbb{R}$ is defined as $h_\theta(\boldsymbol{x}) = \boldsymbol{a}\sigma(\boldsymbol{w}^\top \boldsymbol{x})$. While the original Transformer takes $k = d$, we allow any $k \leq d$ representing the number of features. The mean-field network corresponding to $\mu \in \mathscr{P}(\Theta)$, the space of probability measures on Θ, is defined as $h_\mu(\boldsymbol{x}) = \int_\Theta h_\theta(\boldsymbol{x})\mu(\mathrm{d}\theta)$. We also denote $\boldsymbol{\Sigma}_{\mu,\nu} = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_\mathscr{X}}[h_\mu(\boldsymbol{x})h_\nu(\boldsymbol{x})^\top] \in \mathbb{R}^{k\times k}$. To extract features from the input tokens, the MLP is applied to only the covariates $\boldsymbol{x}_i, \boldsymbol{x}_{\mathrm{qr}}$ so that the following prompt embedding $\mathbf{E}$ is transformed as

$$\mathbf{E} = \begin{bmatrix} \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_n & \boldsymbol{x}_{\mathrm{qr}} \\ y_1 & \cdots & y_n & 0 \end{bmatrix} \quad \mapsto \quad \mathrm{MLP}(\mathbf{E}) = \begin{bmatrix} h_\mu(\boldsymbol{x}_1) & \cdots & h_\mu(\boldsymbol{x}_n) & h_\mu(\boldsymbol{x}_{\mathrm{qr}}) \\ y_1 & \cdots & y_n & 0 \end{bmatrix}.$$

LSA layer. We reparametrize query, key and value matrices $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V \in \mathbb{R}^{(k+1)(k+1)}$ with $\mathbf{W} \in \mathbb{R}^{k\times k}$, $v \in \mathbb{R}$ in the usual manner (Zhang et al., 2023a; Mahankali et al., 2023) and set

$$\mathrm{LSA}(\mathbf{E}) = \mathbf{W}^V\mathbf{E} \cdot \frac{1}{n}(\mathbf{W}^K\mathbf{E})^\top(\mathbf{W}^Q\mathbf{E}), \quad \mathbf{W}^V = \begin{bmatrix} * & * \\ 0_d^\top & v \end{bmatrix}, \quad (\mathbf{W}^K)^\top\mathbf{W}^Q = \begin{bmatrix} \mathbf{W} & 0_d \\ 0_d^\top & * \end{bmatrix}.$$

We further absorb v into $\mathbf{W}$ and fix $v = 1$ to focus on the more complex dynamics of the MLP. Corresponding to the position of y_{qr}, the $(k+1, n+1)$th element of the output is read out as the model prediction, yielding $\widehat{y}_{\mathrm{qr}} = \mathrm{LSA} \circ \mathrm{MLP}(\mathbf{E}) = \frac{1}{n}\sum_{i=1}^n y_i h_\mu(\boldsymbol{x}_i)^\top \mathbf{W} h_\mu(\boldsymbol{x}_{\mathrm{qr}})$. Hence $\widehat{y}_{\mathrm{qr}}$ can be seen as a linear smoother with kernel $k(\boldsymbol{x}, \boldsymbol{x}_{\mathrm{qr}}) = \frac{1}{n}h_\mu(\boldsymbol{x})^\top \mathbf{W} h_\mu(\boldsymbol{x}_{\mathrm{qr}})$ encoded by the MLP.

Regression over Features. In this paper, we study ICL of linear regression tasks over a common nonlinear transformation or feature map $f^\circ \in C(\mathscr{X}, \mathbb{R}^k)$, that is $\mathscr{T} = \{\boldsymbol{v}^\top f^\circ | \, \boldsymbol{v} \in \mathbb{R}^k\}$. We assume that during pretraining the tasks are suitably spread out with $\mathbb{E}_{\boldsymbol{v}}[\boldsymbol{v}\boldsymbol{v}^\top] = \mathbf{I}_k$. We also take the $n \to \infty$ (infinite prompt length) limit to disregard sampling error and let $\widehat{y}_{\mathrm{qr}} = \mathbb{E}_{\boldsymbol{x}}[f(\boldsymbol{x})h_\mu(\boldsymbol{x})^\top]\mathbf{W}h_\mu(\boldsymbol{x}_{\mathrm{qr}})$ for any task $f \in \mathscr{T}$; see Wu et al. (2024) for an analysis of finite task and prompt lengths in the linear case. Hence our Transformer is pretrained with the following mean squared risk,

$$\mathcal{L}_{\mathrm{TF}}(\mu, \mathbf{W}) = \frac{1}{2}\mathbb{E}_{\boldsymbol{x}_{\mathrm{qr}},\boldsymbol{v}}\left[(y_{\mathrm{qr}} - \widehat{y}_{\mathrm{qr}})^2\right] = \frac{1}{2}\mathbb{E}_{\boldsymbol{x}_{\mathrm{qr}}}\left[\left\|f^\circ(\boldsymbol{x}_{\mathrm{qr}}) - \mathbb{E}_{\boldsymbol{x}}[f^\circ(\boldsymbol{x})h_\mu(\boldsymbol{x})^\top]\mathbf{W}h_\mu(\boldsymbol{x}_{\mathrm{qr}})\right\|^2\right] \quad (1)$$

Our goal is to show that gradient dynamics converges to a global minimum such that $\mathcal{L}_{\mathrm{TF}} = 0$. Then the MLP layer has successfully learned the true representations f°, and even for a new or 'unseen' task $\boldsymbol{v}_{\mathrm{new}} \in \mathbb{R}^k$ the Transformer is able to return the correct regression output y_{qr}:

$$\widehat{y}_{\mathrm{qr}} = \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{v}_{\mathrm{new}}^\top f^\circ(\boldsymbol{x})h_\mu(\boldsymbol{x})^\top]\mathbf{W}h_\mu(\boldsymbol{x}_{\mathrm{qr}}) = \boldsymbol{v}_{\mathrm{new}}^\top f^\circ(\boldsymbol{x}_{\mathrm{qr}}) = y_{\mathrm{qr}}.$$

We call this behavior *in-context feature learning* (ICFL). In Appendix B, we show by applying classical analyses of overparametrized 2-layer networks that adding even a shallow MLP greatly increases the class of representations learnable in context from linear functions to the Barron class.

3 Benign Attention Landscape

In this Section, we characterize the infinite-dimensional landscape of the ICFL objective. To ensure regularity, we assume C^2 smoothness of σ and restrict the second layer to $\mathbb{D}^k = \{\boldsymbol{z} \in \mathbb{R}^k : \|\boldsymbol{z}\| \leq 1\}$.

Assumption 1. The parameter space is $\Theta = \mathbb{D}^k \times \mathbb{R}^d$. The nonlinearity is C^2 and $|\sigma| \leq R_1$, $|\sigma'| \leq R_2$, $|\sigma''| \leq R_3$. $\mathcal{D}_\mathscr{X}$ has finite 4th moment, $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_\mathscr{X}}[\|\boldsymbol{x}\|^j] = M_j < \infty$ for $j = 2, 4$.

Next, we suppose $f^\circ = h_{\mu^\circ}$ for some 'teacher' distribution μ°, ensuring learnability and allowing for a rich class of representations (Lemma B.2). These must be suitably spread out in the feature space in order to be learned effectively. To simplify computations, we assume:

Assumption 2. The true features $f^\circ = h_{\mu^\circ}$, $\mu^\circ \in \mathscr{P}_2(\Theta)$ satisfy $\boldsymbol{\Sigma}_{\mu^\circ,\mu^\circ} = r^\circ\mathbf{I}_k$ for $r^\circ > 0$.

We do not require the inputs to be Gaussian as in von Oswald et al. (2023); Akyürek et al. (2023); Zhang et al. (2023a) or orthonormal as in Huang et al. (2023). Note that $R_1^2 \geq \mathrm{tr}\,\boldsymbol{\Sigma}_{\mu^\circ,\mu^\circ} = kr^\circ$ and $\boldsymbol{\Sigma}_{\mu,\nu} \preceq R_1^2\mathbf{I}_k$ for all μ, ν. One implicit assumption is that the number of true features $\dim h_{\mu^\circ}$ is known and equal to k; experiments on a misspecified model are also conducted in Appendix F.

Fast Convergence of Attention. To isolate the dynamics of μ, we first note that minimizing $\mathcal{L}_{\mathrm{TF}}$ over $\mathbf{W}$ is a least-squares regression problem. In particular, $\mathcal{L}_{\mathrm{TF}}$ is convex with respect to $\mathbf{W}$ (strongly

convex unless $\mathbf{\Sigma}_{\mu^\circ,\mu}$ or $\mathbf{\Sigma}_{\mu,\mu}$ are singular) and thus is optimized potentially much more quickly. A possibility is that the MLP $\boldsymbol{x} \mapsto h_\mu(\boldsymbol{x})$ degenerates to completely lie within a low-dimensional linear subspace of $\mathbb{R}^k$. As the regression (1) is ill-conditioned in this case, we set $\mathscr{P}_2^0(\Theta) := \{\mu \in \mathscr{P}_2(\Theta) : \operatorname{rank} \mathbf{\Sigma}_{\mu,\mu} < k\}$ and restrict our attention to $\mathscr{P}_2^+(\Theta) = \mathscr{P}_2(\Theta) \setminus \mathscr{P}_2^0(\Theta)$. We show the singular set $\mathscr{P}_2^0(\Theta)$ is sparse in a strong sense in Proposition C.5, justifying subsequent calculations.

Lemma 3.1. *For any $\mu \in \mathscr{P}_2^+(\Theta)$ and initialization $\mathbf{W}_0$, the flow $\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{W}_t = -\nabla_{\mathbf{W}} \mathcal{L}_{\mathrm{TF}}(\mu, \mathbf{W}_t)$ converges linearly to some $\mathbf{W}_\mu \in \arg\min_{\mathbf{W}} \mathcal{L}_{\mathrm{TF}}(\mu, \mathbf{W})$ satisfying $\mathbf{\Sigma}_{\mu^\circ,\mu}\mathbf{W}_\mu = \mathbf{\Sigma}_{\mu^\circ,\mu}\mathbf{\Sigma}_{\mu,\mu}^{-1}$.*

Thus it is reasonable to suppose that $\mathbf{W}$ is updated sufficiently quickly and has already converged to $\mathbf{W}_\mu$ for each μ – formally by modeling as two-timescale dynamics – leading us to study the objective

$$\mathcal{L}(\mu) := \inf_{\mathbf{W}} \mathcal{L}_{\mathrm{TF}}(\mu, \mathbf{W}) = \tfrac{1}{2}\mathbb{E}_{\boldsymbol{x}}\left[\|\zeta_{\mu^\circ,\mu}(\boldsymbol{x})\|^2\right], \quad \zeta_{\mu^\circ,\mu}(\boldsymbol{x}) = h_{\mu^\circ}(\boldsymbol{x}) - \mathbf{\Sigma}_{\mu^\circ,\mu}\mathbf{\Sigma}_{\mu,\mu}^{-1}h_\mu(\boldsymbol{x}). \quad (2)$$

No Spurious Local Minima. We denote the orthogonal group and unit ball in $\mathbb{R}^{k\times k}$ as $\mathcal{O}(k)$ and $\mathcal{B}_1(k) = \{\mathbf{R} \in \mathbb{R}^{k\times k} : \|\mathbf{R}\| \le 1\}$. For $\mathbf{R} \in \mathcal{O}(k)$, define $\mathbf{R}\sharp\mu$ as the pushforward of μ along the rotation map $\mathbf{R} : (\boldsymbol{a}, \boldsymbol{w}) \mapsto (\mathbf{R}\boldsymbol{a}, \boldsymbol{w})$ so that $h_{\mathbf{R}\sharp\mu}(\boldsymbol{x}) = \int_\Theta \mathbf{R}h_\theta(\boldsymbol{x})\,\mathrm{d}\mu(\theta) = \mathbf{R}h_\mu(\boldsymbol{x})$. Since the convex hull of $\mathcal{O}(k)$ is equal to $\mathcal{B}_1(k)$, this can be extended to any $\mathbf{R} \in \mathcal{B}_1(k)$ by decomposing $\mathbf{R} = \sum_{j=1}^m \alpha_j \mathbf{R}_j$ and defining $\mathbf{R}\sharp\mu = \sum_{j=1}^m \alpha_j \mathbf{R}_j\sharp\mu$. See Lemma C.6 for details. Achieving zero loss implies we have learned the true features h_{μ° up to a linear transformation:

Lemma 3.2. *The pushforwards $\mathbf{R}\sharp\mu^\circ$ for any invertible $\mathbf{R} \in \mathcal{B}_1(k)$ are global minima of $\mathcal{L}$. Conversely, $\mathcal{L}(\mu) = 0$ implies $h_\mu = \mathbf{R}h_{\mu^\circ}$ for some invertible matrix $\mathbf{R}$.*

The following theorem is the main result of this Section. It states that for any μ that is not a global minimum, it is either (1) possible to move in a direction where $\mathcal{L}$ is strictly decreasing, or (2) $\mathcal{L}$ possesses an unstable direction. In particular, all local minima must also be global minima.

Theorem 3.3 (no spurious local minima)**.** *For any $\mu \in \mathscr{P}_2^+(\Theta)$ that is not a global minimum,*

(i) *There exists $\mathbf{R} \in \mathcal{B}_1(k)$ such that along $\bar{\mu}_s = (1-s)\mu + s\mathbf{R}\sharp\mu^\circ$ we have $\frac{\mathrm{d}}{\mathrm{d}s}\big|_{t=0} \mathcal{L}(\bar{\mu}_s) \le 0$.*

(ii) *If $\frac{\mathrm{d}}{\mathrm{d}s}\big|_{s=0} \mathcal{L}(\bar{\mu}_s) = 0$ for all $\mathbf{R} \in \mathcal{B}_1(k)$ above, then $\frac{r^\circ}{2} \le \mathcal{L}(\mu) \le \frac{kr^\circ}{2}$ and $\frac{\mathrm{d}^2}{\mathrm{d}s^2}\big|_{s=0} \mathcal{L}(\bar{\mu}_s) \le -\frac{4}{kR_1^2}\mathcal{L}(\mu)^2$ for some $\mathbf{R} \in \mathcal{B}_1(k)$.*

As a corollary of (ii), critical points cannot exist if $\mathcal{L} < \frac{r^\circ}{2}$, the minimum loss when h_μ is uninformative, i.e. the regression coefficient $\mathbf{\Sigma}_{\mu^\circ,\mu}\mathbf{\Sigma}_{\mu,\mu}^{-1}$ against the true features is singular. We show this leads to an acceleration phase when converging to global minima in Appendix C.3 and Theorem 5.2.

4 Mean-field Dynamics Avoids Saddle Points

Strict saddle properties such as Theorem 3.3 have powerful implications for nonconvex optimization. In finite dimensions, GD almost always avoids saddle points and converges to global optima (Lee et al., 2019); see Appendix D.1 for a recap. We develop the analogous result for general Wasserstein gradient flows (WGF) (3) via the elegant formalism of Otto calculus (Otto, 2001). See Appendix D as well as Ambrosio et al. (2005); Villani (2009) for expository details. Let $F : \mathscr{P}_2(\Omega) \to \mathbb{R}$ be a general C^2 functional with domain $\Omega \subseteq \mathbb{R}^m$. We derive a tangent space characterization of MFD:

Lemma 4.1. *The WGF (μ_t) in a neighborhood of a critical point $\mu^\dagger$ of F can be written as $\mu_t = (\mathrm{id}_\Omega + \epsilon\boldsymbol{v}_t)\sharp\mu^\dagger$ where the field $\boldsymbol{v}_t$ satisfies $\partial_t\boldsymbol{v}_t = -\int \mathbf{H}_{\mu^\dagger}(\theta,\theta')\boldsymbol{v}_t(\theta')\mu^\dagger(\mathrm{d}\theta') + o(1)$. Here, $\mathbf{H}_\mu : (\Omega\times\Omega, \mu\otimes\mu) \to \mathbb{R}^{(k+d)(k+d)}$ is the matrix-valued kernel $\mathbf{H}_\mu(\theta,\theta') := \nabla_\theta\nabla_{\theta'}\frac{\delta^2 F}{\delta\mu^2}(\mu,\theta,\theta')$.*

This facilitates stability analysis via the spectral theory of linear operators,

Lemma 4.2. *Suppose $\mathbf{H}_\mu$ is Hilbert-Schmidt for $\mu \in \mathscr{P}_2(\Omega)$, that is $\iint \|\mathbf{H}_\mu\|^2\,\mathrm{d}\mu\otimes\mathrm{d}\mu < \infty$. Then its integral operator $\mathcal{H}_\mu : f \mapsto \mathcal{H}_\mu f(\theta) = \int \mathbf{H}_\mu(\theta,\theta')f(\theta')\mu(\mathrm{d}\theta')$ is compact self-adjoint, hence there exists an orthonormal basis $\{\psi_j\}_{j\in\mathbb{Z}}$ for $L^2(\Omega,\mu;\mathbb{R}^{k+d})$ of eigenfunctions of $\mathcal{H}_\mu$.*

We thus define the set of *strict saddle* points as $\mathscr{G}^\dagger := \{\mu \in \mathscr{P}_2(\Omega) : \nabla\frac{\delta F}{\delta\mu} = 0,\ \lambda_{\min}(\mathcal{H}_\mu) < 0\}$. Near such points, we now apply the center-stable manifold theorem for Banach spaces (Theorem D.3). This tells us that all flows converging to $\mu^\dagger$ must be eventually contained in the graph of a smooth map in the tangent space defined near the origin. Denoting the reversed WGF for time t inverse to the forward flow as ω_t^- whenever it is defined, we conclude:

Theorem 4.3. *For any C^2 functional $F : \mathscr{P}_2(\Omega) \to \mathbb{R}$ with Hilbert-Schmidt kernel $\mathbf{H}_\mu$, the set $\mathscr{G}_0^\dagger = \{\mu_0 \in \mathscr{P}_2(\Omega) : \lim_{t\to\infty} \mu_t \in \mathscr{G}^\dagger\}$ of initializations which converge to strictly saddle points is contained in the countable union $\cup_{\ell\in\mathbb{N}} \cup_{j\in\mathbb{N}} \omega_\ell^-(\mathscr{V}_j)$ of images of submanifolds $\mathscr{V}_j$.*

For ICFL (2), Proposition E.6 and Theorem 3.3(ii) show that all critical points that are not global optima are strict saddle in $\mathscr{G}^\dagger$. Hence Theorem 4.3 applies to $\mathcal{L}$ with the domain of interest replaced by $\mathscr{P}_2^+(\Theta)$, and thus 'almost all' convergent flows in $\mathscr{P}_2^+(\Theta)$ must converge to global minima. We apply our theory to the training dynamics of three-layer fully connected networks in Appendix D.4.

5 Convergence Rates for ICFL

Theorem 4.3 is encouraging but only qualitative. In this Section, we develop quantitative improvement results for MFD (1) away from critical points; (2) near global minima; and (3) near saddle points. We present our main results for $F = \mathcal{L}$ where $\frac{\delta\mathcal{L}}{\delta\mu}(\mu,\theta) = -\mathbb{E}_{\boldsymbol{x}}[\zeta_{\mu^\circ,\mu}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_{\mu^\circ,\mu}\boldsymbol{\Sigma}_{\mu,\mu}^{-1} h_\theta(\boldsymbol{x})]$ in (3).

MFD with Birth-Death. To preserve the benign landscape, we do not add entropic regularization typical in mean-field analyses; nonetheless, a different modification will be beneficial. For a fixed $\pi \in \mathscr{P}_2(\Theta)$, if at any time $t \geq 0$ the density ratio $\inf_\Theta \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}$ is no larger then a small threshold γ, we perform the discrete update $\mu_t \leftarrow (1-\gamma)\mu_t + \gamma\pi$. This is easily implemented, see Appendix F.1. Similar perturbations have been studied for convex MFD in Wei et al. (2019); Rotskoff et al. (2019).

Assumption 3. π is spherically symmetric in the $\boldsymbol{a}$ component, that is $\pi(\boldsymbol{a},\boldsymbol{w}) = \pi(\boldsymbol{a}',\boldsymbol{w})$ if $\|\boldsymbol{a}\| = \|\boldsymbol{a}'\|$. Also, μ° has finite density w.r.t. π as $\|\mathrm{d}\mu^\circ/\,\mathrm{d}\pi\|_\infty \leq R_4$.

First-order Improvement. We first give a result which translates nonzero gradients along a direction of improvement into a first-order rate of decrease for the gradient flow. Unlike convex mean-field *Langevin* dynamics which relies on a log-Sobolev inequality to control dissipation (Nitanda et al., 2022), our idea is to exploit the mobility of the second layer mass. The argument works for any objective built on top of the MLP layer h_μ; see Proposition E.1 in the Appendix for the general result.

Proposition 5.1. *Suppose MFD with birth-death on $\mathcal{L}$ at time t satisfies Theorem 3.3(i) with $\frac{\mathrm{d}}{\mathrm{d}s}\big|_{s=0}\mathcal{L}(\bar{\mu}_s) \leq -\delta$. Then $\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}(\mu_t) \leq -R_4^{-1}\gamma\delta^2$.*

We further establish convergence near global minima. The rate is quadratic in the *feature* dimensions k and independent of d; we take $r^\circ = \Theta(\frac{1}{k})$. O hides polynomial dependency on constants R_j, M_j.

Theorem 5.2. *Once $\mathcal{L}(\mu_t) \leq 0.49r^\circ$, MFD with birth-death converges with loss $\leq \epsilon$ in time $O(\frac{k^2}{\gamma\epsilon})$.*

Second-order Improvement. We now arrive at the main difficulty of our analysis: the behavior of MFD near critical points. In the finite-dimensional case, local stability is determined by the Hessian. We show that the mean-field analogue for a smooth functional $F : \mathscr{P}_2(\Theta) \to \mathbb{R}$ is

Lemma 5.3. *The velocity field $\nabla\frac{\delta F}{\delta\mu}$ of (3) evolves as $\partial_t\big[\nabla\frac{\delta F}{\delta\mu}(\mu_t)\big] = -\,\mathcal{H}_{\mu_t}\big[\nabla\frac{\delta F}{\delta\mu}(\mu_t)\big]$.*

For the ICFL objective $\mathcal{L}$, we show $\mathcal{H}_\mu$ is Hilbert-Schmidt and derive regularity properties in Lemma E.3 and E.5. Next, we can translate second-order instability into a spectral bound for $\mathcal{H}_\mu$.

Proposition 5.4. *Suppose MFD with birth-death at time $t \geq 0$ satisfies Theorem 3.3(ii) with $\frac{\mathrm{d}^2}{\mathrm{d}s^2}\big|_{s=0}\mathcal{L}(\bar{\mu}_s) \leq -\Lambda$. Then the smallest eigenvalue λ_0 of $\mathcal{H}_{\mu_t}$ satisfies $\lambda_0 \leq -R_4^{-1}\gamma\Lambda$.*

Therefore we expect that even if the dynamics is close to a saddle point and Proposition 5.1 is not useful, as long as the L^2-component along the eigenfunction ψ_0 corresponding to λ_0 is not exactly zero, it will blow up exponentially in time until μ_t escapes and makes progress. In detail,

Theorem 5.5. *Suppose MFD with birth-death on $\mathcal{L}$ at time t satisfies Theorem 3.3(i) with $\frac{\mathrm{d}}{\mathrm{d}s}\big|_{s=0}\mathcal{L}(\bar{\mu}_s) \geq -O(k^{-1}\,\mathcal{L}(\mu_t)^2)$. Further suppose ψ_0 satisfies $|\int \psi_0^\top \nabla\frac{\delta\mathcal{L}}{\delta\mu}(\mu_t)\,\mathrm{d}\mu_t| \geq \alpha$ for some $\alpha \geq 0$. Then MFD within the region $\{\mu \in \mathscr{P}_2(\Theta) : \lambda_{\min}(\boldsymbol{\Sigma}_{\mu,\mu}) = \Omega(\frac{1}{k})\}$ decreases $\mathcal{L}$ in time $\tau = \widetilde{O}\left(\frac{k}{\gamma\,\mathcal{L}(\mu_t)^2}\log\frac{1}{\alpha}\right)$ as $\mathcal{L}(\mu_{t+\tau}) \leq \mathcal{L}(\mu_t) - \widetilde{\Omega}\left(\frac{\gamma^2\alpha\,\mathcal{L}(\mu_t)^4}{k^5 d}\right)$.*

Simply put, we make $\widetilde{\Omega}(\alpha)$ progress in $\widetilde{O}(\log\frac{1}{\alpha})$ time. The idea is to find a $\mathcal{W}_2$-ball where if μ_t does not escape in time τ, the blowup guarantees improvement; if μ_t does escape, $\mathcal{L}$ must have decreased to warrant such displacement. Again, we present general versions in Proposition E.6 and Theorem E.7. Unfortunately, this is not enough to establish global rates as the flow might be initialized at or pass very near multiple saddle points; this is unavoidable even in finite dimensions (Du et al., 2017). We propose a perturbative scheme to escape saddles based on Gaussian processes in Appendix E.3.

6 CONCLUSION

In this paper, we explored the training dynamics of a Transformer with one MLP and one attention layer, enabling in-context feature learning of regression tasks on a rich class of representations. We showed that the loss landscape becomes benign in the two-timescale and mean-field limit and developed instability and improvement guarantees for the Wasserstein gradient flow. To our knowledge, this represents both the first work to theoretically study how features are learned in context, and the first analysis of nonconvex mean-field dynamics for strict saddle objectives.

ACKNOWLEDGMENTS

JK was partially supported by JST CREST (JPMJCR2015). TS was partially supported by JSPS KAKENHI (20H00576) and JST CREST (JPMJCR2115).

REFERENCES

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. *arXiv preprint arXiv:2306.00297*, 2023.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? Investigations with linear models. In *International Conference on Learning Representations*, 2023.

Mauricio Álvarez, Lorenzo Rosasco, and Neil Lawrence. Kernels for vector-valued functions: a review. *Foundations and Trends in Machine Learning*, 4(3):195–266, 2012.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Lectures in Mathematics, ETH Zürich. Springer, 2005.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: provable in-context learning with in-context algorithm selection. In *ICML Workshop on Efficient Systems for Foundation Models*, 2023.

Andrew Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.

Andrew Barron. Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 14(1):115–133, 1994.

Raphael Berthier, Andrea Montanari, and Kangjie Zhou. Learning time-scales in two-layers neural networks. *arXiv preprint arXiv:2303.00055*, 2023.

Alberto Bietti, Joan Bruna, and Loucas Pillaud-Vivien. On learning Gaussian multi-index models with gradient flow. *arXiv preprint arXiv:2310.19793*, 2023.

Sergey G. Bobkov and Michel Ledoux. One-dimensional empirical measures, order statistics, and Kantorovich transport distances. *Memoirs of the American Mathematical Society*, 261, 2019.

Siwan Boufadène and François-Xavier Vialard. On the global convergence of Wasserstein gradient flow of the Coulomb discrepancy. *arXiv preprint arXiv:2312.00800*, 2024.

Fan Chen, Zhenjie Ren, and Songbo Wang. Uniform-in-time propagation of chaos for mean field Langevin dynamics. *arXiv preprint arXiv:2212.03050v2*, 2022.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*, 2018.

Simon Shaolei Du, Chi Jin, J. Lee, Michael I. Jordan, Aarti Singh, and Barnabás Póczos. Gradient descent can take exponential time to escape saddle points. In *Advances in Neural Information Processing Systems*, 2017.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162:707–738, 2015.

Thierry Gallay. A center-stable manifold theorem for differential equations in Banach spaces. *Communications in Mathematical Physics*, 152(2):249–268, 1993.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can Transformers learn in-context? A case study of simple function classes. In *Advances in Neural Information Processing Systems*, 2022.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points - online stochastic gradient for tensor decomposition. *JMLR*, 40:1–46, 2015.

Rong Ge, Jason D. Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, 2016.

Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: a unified geometric analysis. In *International Conference on Machine Learning*, 2017.

Tianyu Guo, Wei Hu, Song Mei, Huan Wang, Caiming Xiong, Silvio Savarese, and Yu Bai. How do Transformers learn in-context beyond simple functions? A case study on learning with representations. *arXiv preprint arXiv:2310.10616*, 2023.

Chi Han, Ziqi Wang, Han Zhao, and Heng Ji. Explaining emergent in-context learning as kernel regression. *arXiv preprint arXiv:2305.12766*, 2023.

Yu Huang, Yuan Cheng, and Yingbin Liang. In-context convergence of Transformers. *arXiv preprint arXiv:2310.05249*, 2023.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. How to escape saddle points efficiently. In *International Conference on Machine Learning*, 2017.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

Juno Kim, Kakei Yamamoto, Kazusato Oko, Zhuoran Yang, and Taiji Suzuki. Symmetric mean-field Langevin dynamics for distributional minimax problems. In *International Conference on Learning Representations*, 2024.

Jason Klusowski and Andrew Barron. Risk bounds for high-dimensional ridge function combinations including neural networks. *arXiv preprint arXiv:1607.01434*, 2016.

Jason Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael Jordan, and Benjamin Recht. First-order methods almost always avoid saddle points. *Mathematical Programming*, 2019.

Xingguo Li, Junwei Lu, Raman Arora, Jarvis Haupt, Han Liu, Zhaoran Wang, and Tuo Zhao. Symmetry, saddle points, and global optimization landscape of nonconvex matrix factorization. *IEEE Transactions on Information Theory*, 2019.

Zhong Li, Chao Ma, and Lei Wu. Complexity measures for neural networks with general activation fnctions using path-based norms. *arXiv preprint arXiv:2009.06132*, 2020.

Licong Lin, Yu Bai, and Song Mei. Transformers as decision makers: provable in-context reinforcement learning via supervised pretraining. *arXiv preprint arXiv:2310.08566*, 2023.

John Lott. Some geometric calculations on Wasserstein space. *Communications in Mathematical Physics*, 277:423–437, 2008.

Arvind Mahankali, Tatsunori B. Hashimoto, and Tengyu Ma. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. *arXiv preprint arXiv:2307.03576*, 2023.

Pierre Marion and Raphaël Berthier. Leveraging the two-timescale regime to demonstrate convergence of neural networks. In *Advances in Neural Information Processing Systems*, 2023.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 115:7665–7671, 2018.

Atsushi Nitanda, Denny Wu, and Taiji Suzuki. Convex analysis of the mean field Langevin dynamics. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022.

Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. *Communications in Partial Differential Equations*, 26:101–174, 2001.

Grant Rotskoff, Samy Jelassi, Joan Bruna, and Eric Vanden-Eijnden. Global convergence of neuron birth-death dynamics. In *International Conference on Machine Learning*, 2019.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Progress in Nonlinear Differential Equations and Their Applications. Springer International Publishing, 2015.

Michael Shub. *Global Stability of Dynamical Systems*. Springer New York, 2013.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2018.

Taiji Suzuki, Denny Wu, and Atsushi Nitanda. Convergence of mean-field Langevin dynamics: Time and space discretization, stochastic gradient, and variance reduction. In *Advances in Neural Information Processing Systems*, 2023.

Alain-Sol Sznitman. Topics in propagation of chaos. *École d'Été de Probabilités de Saint-Flour XIX-1989*, 1464:165–251, 1991.

Yuandong Tian, Yiping Wang, Zhenyu Zhang, Beidi Chen, and Simon Du. JoMA: demystifying multilayer Transformers via joint dynamics of MLP and attention, 2023.

Yao-Hung Tsai, Shaojie Bai, Makoto Yamada, Louis-Philippe Morency, and Ruslan Salakhutdinov. Transformer dissection: an unified understanding for Transformer's attention via the lens of kernel. In *Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, 2019.

Cédric Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin, 2009.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, 2023.

Colin Wei, Jason Lee, Qiang Liu, and Tengyu Ma. Regularization matters: generalization and optimization of neural nets v.s. their induced kernel. In *Advances in Neural Information Processing Systems*, 2019.

E Weinan and Stephan Wojtowytsch. Representation formulas and pointwise properties for Barron functions. *Calculus of Variations and Partial Differential Equations*, 61(2):1–37, 2022.

E Weinan, Chao Ma, and Lei Wu. A priori estimates of the population risk for two-layer neural networks. *Communications in Mathematical Sciences*, 17(5):1407–1425, 2019.

E Weinan, Chao Ma, Lei Wu, and Stephan Wojtowytsch. Towards a mathematical understanding of neural network-based machine learning: what we know and what we don't. *CSIAM Transactions on Applied Mathematics*, 1(4):561–615, 2020.

E Weinan, Chao Ma, and Lei We. The Barron space and the flow-induced function spaces for neural network models. *Constructive Approximation*, 55(1):369–406, 2022.

Jingfeng Wu, Difan Zou, Zixiang Chen, Vladimir Braverman, Quanquan Gu, and Peter L. Bartlett. How many pretraining tasks are needed for in-context learning of linear regression? In *International Conference on Learning Representations*, 2024.

Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained Transformers learn linear models in-context. *arXiv preprint arXiv:2306.09927*, 2023a.

Yufeng Zhang, Fengzhuo Zhang, Zhuoran Yang, and Zhaoran Wang. What and how does in-context learning learn? Bayesian model averaging, parameterization, and generalization. *arXiv preprint arXiv:2305.19420*, 2023b.

A PRELIMINARIES

A.1 BACKGROUND THEORY

We begin by providing some necessary background for mean-field dynamics. Let $\Omega \subseteq \mathbb{R}^m$ be a Euclidean domain with smooth boundary $\partial\Omega$. For $p \geq 1$, let $\mathscr{P}_p(\Omega)$ be the p-Wasserstein space of probability measures on Ω vanishing on $\partial\Omega$ with finite pth moment. We will mostly be concerned with the space $\mathscr{P}_2(\Omega)$.

Definition A.1 (functional derivative)**.** The functional derivative $\frac{\delta F}{\delta \mu}$ of a functional $F : \mathscr{P}_p(\Omega) \to \mathbb{R}$ is defined (if one exists) as a functional $\frac{\delta F}{\delta \mu} : \mathscr{P}_p(\Omega) \times \Omega \to \mathbb{R}$ satisfying for all $\nu \in \mathscr{P}_p(\Omega)$,

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon}\bigg|_{\epsilon=0} F(\mu + \epsilon(\nu - \mu)) = \int_\Omega \frac{\delta F}{\delta \mu}(\mu, \theta)(\nu - \mu)(\mathrm{d}\theta).$$

Note that the functional derivative is defined up to additive constants. We say a functional F is C^1 if $\nabla \frac{\delta F}{\delta \mu}(\mu, \theta)$ is well-defined and continuous, and C^2 if $\nabla_\theta \nabla_{\theta'} \frac{\delta^2 F}{\delta \mu^2}(\mu, \theta, \theta')$ is well-defined and continuous. Furthermore, the functional F is convex if for all $\nu \in \mathscr{P}_p(\Omega)$ it holds that

$$F(\nu) \geq F(\mu) + \int_\Omega \frac{\delta F}{\delta \mu}(\mu, \theta)(\nu - \mu)(\mathrm{d}\theta).$$

Definition A.2 (p-Wasserstein metric)**.** The p-Wasserstein distance between $\mu, \nu \in \mathscr{P}_p(\Omega)$ is defined as

$$\mathcal{W}_p(\mu, \nu) = \left(\inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}^m} \|x - y\|^p \, \mathrm{d}\gamma(x, y) \right)^{\frac{1}{p}}$$

where $\Pi(\mu, \nu)$ denotes the set of joint distributions on $\Omega \times \Omega$ whose first and second factors have marginal laws μ and ν, respectively.

We consider $\mathscr{P}_p(\Omega)$ as a metric space with respect to $\mathcal{W}_p$, which metrizes weak convergence on $\mathscr{P}_p(\Omega)$ (Villani, 2009, Theorem 6.9). By Hölder's inequality it always holds that $\mathscr{P}_2(\Omega) \subset \mathscr{P}_1(\Omega)$ and $\mathcal{W}_1(\mu, \nu) \leq \mathcal{W}_2(\mu, \nu)$. The $\mathcal{W}_1$ metric is also characterized via Kantorovich-Rubinstein duality as

$$\mathcal{W}_1(\mu, \nu) = \sup_{\|f\|_{\mathrm{Lip}} \leq 1} \int_\Omega f \, \mathrm{d}\mu - \int_\Omega f \, \mathrm{d}\nu,$$

where the supremum runs over all 1-Lipschitz functions $f : \Omega \to \mathbb{R}$, which makes it well-suited for perturbation analyses.

We develop more advanced theory concerning the local metric geometry and characterization of flows on $\mathscr{P}_2(\Omega)$ in Appendix D. As a consequence, one can show the following variational formulation of the $\mathcal{W}_2$ metric:

Proposition A.3 (Benamou-Brenier formula)**.** *For $\mu, \nu \in \mathscr{P}_2(\Omega)$ it holds that*

$$\mathcal{W}_2(\mu, \nu) = \inf \left\{ \int_0^1 \|\boldsymbol{v}_t\|^2_{L^2(\Omega, \mu_t; \mathbb{R}^m)} \, \mathrm{d}t : \partial_t \mu_t + \nabla \cdot (\boldsymbol{v}_t \mu_t) = 0, \ \mu_0 = \mu, \ \mu_1 = \nu \right\},$$

where the infimum runs over all unit time flows $(\mu_t)_{t \in [0,1]}$ from μ to ν.

Proof. See e.g. Ambrosio et al. (2005), Chapter 8 or Santambrogio (2015), Section 6.1. □

The formula can be used to bound the movement of Wasserstein flows in relation to the magnitude of the gradient field. For convenience, we will use the following time-rescaled version which is easily checked:

$$\mathcal{W}_2(\mu, \nu) = \inf \left\{ \tau \int_0^\tau \|\boldsymbol{v}_t\|^2_{L^2(\Omega, \mu_t; \mathbb{R}^m)} \, \mathrm{d}t : \partial_t \mu_t + \nabla \cdot (\boldsymbol{v}_t \mu_t) = 0, \ \mu_0 = \mu, \ \mu_\tau = \nu, \ \tau > 0 \right\}.$$

When the velocity field $\boldsymbol{v}_t = -\nabla \frac{\delta F}{\delta \mu}(\mu_t)$ is given as the functional derivative of a given functional F, the dynamics can be interpreted as the continuous-time limit of a discrete gradient descent process on

F w.r.t. the $\mathcal{W}_2$ metric via the celebrated JKO scheme (Jordan et al., 1998). Specifically, the implicit Euler scheme

$$\mu_\eta^{(k+1)} = \operatorname*{arg\,min}_{\mu \in \mathscr{P}_2(\Omega)} \frac{1}{2} \mathcal{W}_2(\mu, \mu^{(k)})^2 + \eta F(\mu), \quad \mu^{(0)} = \mu_0$$

converges weakly in the limit $\eta \to 0$ to the solution of the continuity or Fokker-Planck equation $\partial_t \mu_t = \nabla \cdot \left(\mu_t \nabla \frac{\delta F}{\delta \mu}(\mu_t) \right)$ in the sense that $\mu_\eta^{\lfloor t/\eta \rfloor} \rightharpoonup \mu_t$ for all time $t \geq 0$. Hence we refer to this process as the Wasserstein gradient flow on $\mathscr{P}_2(\Omega)$ with respect to F.

A.2 Related Works

In-context learning. A wide literature has developed around the various aspects of ICL; we only mention those most relevant to our setup. Akyürek et al. (2023); von Oswald et al. (2023); Mahankali et al. (2023) give a construction where a single linear attention layer is equivalent to one step of GD or ridge regression. Transformers are also capable of implementing statistical (Bai et al., 2023) and reinforcement learning algorithms (Lin et al., 2023) and model averaging (Zhang et al., 2023b). The attention-over-representation viewpoint has been studied by Guo et al. (2023) and also Tsai et al. (2019); Han et al. (2023) from a kernel regression perspective. Zhang et al. (2023a) analyze the optimization of a linear attention-only Transformer and show global convergence; a relationship to preconditioned GD is established in Ahn et al. (2023). Also, Huang et al. (2023) give a stage-wise analysis for the softmax attention-only model. Finally, a joint dynamic framework for MLP and attention has been proposed in Tian et al. (2023).

Mean-field dynamics. Let h_θ denote a single neuron with parameter $\theta \in \Omega \subseteq \mathbb{R}^m$ and $\mathscr{P}(\Omega)$ the space of probability measures over Ω.[1] In the infinite-width limit, a 2-layer neural network $h_N(\boldsymbol{x}) = \frac{1}{N} \sum_{j=1}^N h_{\theta^{(j)}}(\boldsymbol{x})$ can be written as an expectation $h_\mu(\boldsymbol{x}) = \mathbb{E}_{\theta \sim \mu}[h_\theta(\boldsymbol{x})]$ over a distribution $\mu \in \mathscr{P}(\Omega)$. The corresponding mean-field limit of gradient flow (GF) w.r.t. an objective function $F : \mathscr{P}(\Omega) \to \mathbb{R}$, the Wasserstein gradient flow, is given by the continuity equation

$$\partial_t \mu_t = \nabla \cdot \left(\mu_t \nabla \frac{\delta F}{\delta \mu}(\mu_t) \right), \quad t \geq 0. \tag{3}$$

Networks in this regime are capable of dynamic feature learning and convergence, compared to the NTK regime where the underlying kernel is essentially frozen. Works such as Chizat & Bach (2018); Mei et al. (2018); Nitanda et al. (2022) exploit the linearity of h_μ in μ and the convexity of the loss to lift to a convex optimization problem on $\mathscr{P}(\Omega)$ and obtain convergence results. In contrast, the ICL loss is inherently nonconvex due to the additional attention layer.

Landscape analyses. Certain nonconvex objectives such as matrix completion, sensing and factorization have been proved to be benign via directional analysis (Ge et al., 2016; 2017; Li et al., 2019). Recently, Gaussian k-index models have been shown to possess benign landscapes w.r.t. the projection matrix after factoring out the link function via a similar two-timescale limit (Bietti et al., 2023). However, our work focuses on the optimization of the *infinite*-dimensional variable $\mu \in \mathscr{P}(\Theta)$, and the ICL objective (2) has a novel, more complex structure compared to these problems. Boufadène & Vialard (2024) study a certain energy functional and prove benignity via flow interchange techniques; however, they do not discuss its implications for general gradient flow.

B Results and Proofs for Section 2

B.1 Expressivity of Representations

Multivariate Barron class. Barron-type spaces have been well established as the natural function classes for analyzing approximation and generalization of shallow networks (Barron, 1994; Weinan et al., 2020; Weinan & Wojtowytsch, 2022). Here, we extend the theory to our vector-valued setting. We focus on the ReLU case for ease of presentation which does not satisfy Assumption 1, but many results extend to more general activations (Klusowski & Barron, 2016; Li et al., 2020).

[1] We will also consider the space $\mathscr{P}_2(\Omega)$ of the space of probability measures on Θ with bounded second moment that vanish on the boundary of Θ, equipped with the 2-Wasserstein metric.

Set $\Theta = \mathbb{R}^k \times \mathbb{R}^d$, $\sigma(z) = \max\{0, z\}$ and suppose $M_2 = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\mathscr{X}}}[\|\boldsymbol{x}\|^2] < \infty$. The Barron space $\mathscr{B}_p$ of order $p \in [1, \infty]$ is defined as the set of functions $f = h_\mu$, $\mu \in \mathscr{P}(\Theta)$ with finite Barron norm

$$\|f\|_{\mathscr{B}_p} := \inf_{\mu: f = h_\mu} \left[\int \|\boldsymbol{a}\|^p \|\boldsymbol{w}\|^p \mu(\mathrm{d}\theta)\right]^{1/p}.$$

This turns out to not depend on p (Lemma B.6), so we refer to *the* Barron space and norm as $(\mathscr{B}, \|\cdot\|_{\mathscr{B}})$. This space contains a rich variety of functions. The following is an application of the classical Fourier analysis (Barron, 1993).

Proposition B.1. *Suppose h_μ includes a bias term, i.e. $\mathscr{X} \subseteq \mathscr{X}_0 \times \{1\}$. If $f = (f_j)_{j=1}^k \in C(\mathscr{X}_0, \mathbb{R}^k)$ such that each f_j satisfies $\inf_{\widehat{f}_j} \int_{\mathbb{R}^{d-1}} \|\omega\|_1^2 |\widehat{f}_j(\omega)| < \infty$ for $\widehat{f}_j$ the Fourier transform of an extension of f_j to $\mathbb{R}^{d-1}$, then $f \in \mathscr{B}$. In particular, the Sobolev space $H^s(\mathscr{X}_0)^k \subset \mathscr{B}$ for $s > \frac{d+1}{2}$.*

Furthermore, $\mathscr{B}$ is exactly the class of representations that can be learned in context, demonstrating the expressive power gained by incorporating the MLP layer:

Lemma B.2. *$\mathcal{L}_{\mathrm{TF}}(\mu, \mathbf{W}) = 0$ has a solution such that $\operatorname{ess\,sup}_\mu \|\boldsymbol{a}\|\|\boldsymbol{w}\| < \infty$ if and only if $f^\circ \in \mathscr{B}$.*

In contrast, Mahankali et al. (2023) show that the optimal LSA-only Transformer implements one step of GD for the *linear* regression problem $(\boldsymbol{x}_i, y_i)_{i=1}^n$ even when $y_i|\boldsymbol{x}_i$ is nonlinear; thus we establish a clear gap in learning ability.

Generalization to unseen tasks. If the Transformer has successfully learned f°, it will achieve perfect accuracy on any new linear task $\boldsymbol{v}_{\mathrm{new}}^\top f^\circ$ as discussed. On the other hand, if the test task is an arbitrary function $g \in C(\mathscr{X})$, we cannot hope to do better than the projection to the linear span of learned features $f_1^\circ, \cdots, f_k^\circ$ since (1) is a regression loss. We show this lower bound is optimal:

Proposition B.3. *Suppose $\mathcal{L}_{\mathrm{TF}}(\mu, \mathbf{W}) \le \epsilon$ for $f^\circ \in \mathscr{B}$ and $\|h_\mu\|_{\mathscr{B}}, \|\mathbf{W}\| \lesssim 1$. Then for any new task $g \in C(\mathscr{X})$ with $\|g\|_{L^2(\mathcal{D}_{\mathscr{X}})} \lesssim 1$, the test ICL error satisfies*

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{x}_{\mathrm{qr}}}\left[\left\|g(\boldsymbol{x}_{\mathrm{qr}}) - \mathbb{E}_{\boldsymbol{x}}[g(\boldsymbol{x})h_\mu(\boldsymbol{x})^\top]\mathbf{W}h_\mu(\boldsymbol{x}_{\mathrm{qr}})\right\|^2\right] \\ &\lesssim \epsilon + \inf_{\boldsymbol{v}\in\mathbb{R}^k} \|g - \boldsymbol{v}^\top f^\circ\|_{L^2(\mathcal{D}_{\mathscr{X}})}^2.\end{aligned}$$

This extends the LSA-only case where the optimal output was shown to be the near-optimal linear model in Zhang et al. (2023a). This also raises an important question: if the task g depends nonlinearly on h_{μ°, is it still beneficial to have learned the relevant features μ°? Clearly this depends on both g and the initialization μ_0; however, we present experiments supporting this intuition in Appendix F.

B.2 From Finite to Infinite Width

Continuing the above discussion, elements of the Barron space are effectively approximated by finite-width networks, which can be seen as an adaptive kernel method. The proof of the following is essentially due to Weinan et al. (2022).

Proposition B.4. *For any integer N and $f^\circ \in \mathscr{B}$, there exists a width N network $\widehat{h}_N$ given by the discrete measure $\widehat{\mu}_N = \frac{1}{N}\sum_{j=1}^N \delta_{\theta^{(j)}}$ with path norm $\|\widehat{h}_N\|_{\mathcal{P}} := \frac{1}{N}\sum_{j=1}^N \|\boldsymbol{a}^{(j)}\|\|\boldsymbol{w}^{(j)}\| \le 3\|f^\circ\|_{\mathscr{B}}$ and*

$$\inf_{\mathbf{W}} \mathcal{L}_{\mathrm{TF}}(\widehat{\mu}_N, \mathbf{W}) \le \tfrac{1}{2}\|\widehat{h}_N - f^\circ\|_{L^2(\mathcal{D}_{\mathscr{X}})}^2 \le \tfrac{M_2\|f^\circ\|_{\mathscr{B}}^2}{N}.$$

Using the low Rademacher complexity of Barron spaces, we can also simultaneously bound the generalization gap for a finite number of tasks T as $\widetilde{O}(T^{-1/2})$ which is nearly minimax optimal (Weinan et al., 2019, Theorem 4.1).

Moreover from a dynamical perspective, a propagation of chaos argument (Sznitman, 1991) shows that gradient descent indeed converges to (3) in the infinite-width limit. Let $F : \mathscr{P}_2(\Omega) \to \mathbb{R}$ be any C^1 functional such that $\|\nabla \frac{\delta F}{\delta \mu}\| \le L_1$, $\nabla \frac{\delta F}{\delta \mu}(\mu, \theta)$ is L_2-Lipschitz w.r.t. θ and L_3-Lipschitz w.r.t. μ in the $\mathcal{W}_1$ metric. Denote the initial measure as $\mu_0 \in \mathscr{P}_2(\Omega)$, let $\theta_0^{(1)}, \cdots, \theta_0^{(N)}$ be i.i.d. samples from μ_0 and consider the empirical GF trajectories

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\theta_t^{(j)} = -\nabla \tfrac{\delta F}{\delta \mu}(\widehat{\mu}_{t,N}, \theta_t^{(j)}), \quad \widehat{\mu}_{t,N} = \tfrac{1}{N}\sum_{j=1}^N \delta_{\theta_t^{(j)}}.$$

Proposition B.5. *For any $T \geq 0$, the N-particle empirical measure $\widehat{\mu}_{t,N}$ converges to μ_t as $\mathbb{E}[\mathcal{W}_1(\widehat{\mu}_{t,N}, \mu_t)] \to 0$ uniformly for all $t \in [0,T]$ as $N \to \infty$.*

See Remark B.9 for the case of the ICFL objective. Hence it is natural to analyze optimization in the mean-field or extremely overparametrized regime.

B.3 Proofs for Appendix B.1

Lemma B.6. *For any $p > 1$ it holds that $\|f\|_{\mathscr{B}_1} = \|f\|_{\mathscr{B}_p} = \|f\|_{\mathscr{B}_\infty}$, where*

$$\|f\|_{\mathscr{B}_\infty} := \inf_{\mu : f = h_\mu} \operatorname*{ess\,sup}_{(\boldsymbol{a},\boldsymbol{w})\sim\mu} \|\boldsymbol{a}\|\|\boldsymbol{w}\|.$$

Proof. Note $\|f\|_{\mathscr{B}_1} \leq \|f\|_{\mathscr{B}_p} \leq \|f\|_{\mathscr{B}_\infty}$ trivially by Hölder's inequality. For $f \in \mathscr{B}_1$, choose a measure μ such that $f = h_\mu$ and $\int \|\boldsymbol{a}\|\|\boldsymbol{w}\|\mu(\mathrm{d}\theta) \leq \|f\|_{\mathscr{B}_1} + \epsilon$ and define the nonnegative measure $\underline{\mu}$ on $\mathbb{S}^{k-1} \times \mathbb{S}^{d-1}$ as

$$\underline{\mu}(A,B) = \int_{\underline{\boldsymbol{a}}\in A, \underline{\boldsymbol{w}}\in B} \|\boldsymbol{a}\|\|\boldsymbol{w}\|\mu(\mathrm{d}\boldsymbol{a},\mathrm{d}\boldsymbol{w}), \quad \underline{\boldsymbol{a}} = \frac{\boldsymbol{a}}{\|\boldsymbol{a}\|}, \quad \underline{\boldsymbol{w}} = \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$$

for Borel sets $A \subseteq \mathbb{S}^{k-1}$, $B \subseteq \mathbb{S}^{d-1}$. Then we can rewrite f via the 'projected' measure $\underline{\mu}$ as

$$f = \int h_\theta(\boldsymbol{x})\mu(\mathrm{d}\theta) = \int \|\boldsymbol{a}\|\|\boldsymbol{w}\| \cdot \underline{\boldsymbol{a}}\sigma(\underline{\boldsymbol{w}}^\top \boldsymbol{x})\mu(\mathrm{d}\boldsymbol{a},\mathrm{d}\boldsymbol{w}) = \int \underline{\boldsymbol{a}}\sigma(\underline{\boldsymbol{w}}^\top \boldsymbol{x})\underline{\mu}(\mathrm{d}\boldsymbol{a},\mathrm{d}\boldsymbol{w}).$$

Factoring out the total mass of $\underline{\mu}$ to form a probability distribution on $\mathscr{P}(\mathbb{S}^{k-1} \times \mathbb{S}^{d-1})$, we obtain a representation of f such that the ∞-Barron norm becomes bounded as

$$\|f\|_{\mathscr{B}_\infty} \leq \underline{\mu}(\mathbb{S}^{k-1}, \mathbb{S}^{d-1}) \operatorname*{ess\,sup}_{(\boldsymbol{a},\boldsymbol{w})\sim\underline{\mu}} \|\boldsymbol{a}\|\|\boldsymbol{w}\| \leq \|f\|_{\mathscr{B}_1} + \epsilon.$$

Taking $\epsilon \to 0$ shows the reverse inequality. □

Proof of Proposition B.1. If $f_j \in C(\mathscr{X}, \mathbb{R})$ satisfies $\inf_{\widehat{f}_j} \int_{\mathbb{R}^{d-1}} \|\omega\|_1^2 |\widehat{f}_j(\omega)| < \infty$ for some transform $\widehat{f}_j$, it admits a representation

$$f_j(\boldsymbol{x}) = \int a_j \sigma(\boldsymbol{w}^\top \boldsymbol{x})\mu_j(\mathrm{d}a_j, \mathrm{d}\boldsymbol{w})$$

for a probability distribution μ_j on $\mathbb{R} \times \mathbb{R}^d$ (Barron, 1993; Weinan et al., 2022). Consider the scaled inclusion map

$$\iota_j : \mathbb{R} \times \mathbb{R}^d \hookrightarrow \mathbb{R}^k \times \mathbb{R}^d, \quad \iota_j(a_j, \boldsymbol{w}) = (k a_j e_j, \boldsymbol{w}),$$

where e_j is the unit vector with all zeros except for a single 1 at the jth coordinate. Then for the averaged pushforward measure $\bar{\mu} = \frac{1}{k}\sum_{j=1}^k \iota_j \sharp \mu_j \in \mathscr{P}(\Theta)$ it holds that

$$\begin{aligned} h_{\bar{\mu}}(\boldsymbol{x}) &= \frac{1}{k}\sum_{j=1}^k \int \boldsymbol{a}_j \sigma(\boldsymbol{w}^\top \boldsymbol{x}) \iota_j \sharp \mu_j(\mathrm{d}\boldsymbol{a}_j, \mathrm{d}\boldsymbol{w}) \\ &= \sum_{j=1}^k \int a_j e_j \sigma(\boldsymbol{w}^\top \boldsymbol{x})\mu_j(\mathrm{d}\boldsymbol{a}_j, \mathrm{d}\boldsymbol{w}) \\ &= \sum_{j=1}^k e_j f_j(\boldsymbol{x}) = f(\boldsymbol{x}), \end{aligned}$$

and therefore $f \in \mathscr{B}$. □

Proof of Lemma B.2. Let $\mathbf{A}^\dagger$ denote the pseudoinverse of a matrix $\mathbf{A}$. If $f^\circ = h_{\mu^\circ}$ for some distribution $\mu^\circ \in \mathscr{P}(\Theta)$ with $\|f^\circ\|_{\mathscr{B}} < \infty$, then setting $\mathbf{W}^\circ = \mathbf{\Sigma}^\dagger_{\mu^\circ,\mu^\circ} = \mathbb{E}_{\boldsymbol{x}}[f^\circ(\boldsymbol{x})f^\circ(\boldsymbol{x})^\top]^\dagger$,

$$\begin{aligned}\mathcal{L}_{\mathrm{TF}}(\mu^\circ, \mathbf{W}^\circ) &= \frac{1}{2}\mathbb{E}_{\boldsymbol{x}_{\mathrm{qr}}}\left[\|f^\circ(\boldsymbol{x}_{\mathrm{qr}}) - \mathbf{\Sigma}_{\mu^\circ,\mu^\circ}\mathbf{W}^\circ f^\circ(\boldsymbol{x}_{\mathrm{qr}})\|^2\right] \\ &= \frac{1}{2}\operatorname{tr}\mathbf{\Sigma}_{\mu^\circ,\mu^\circ} - \operatorname{tr}\left(\mathbf{\Sigma}_{\mu^\circ,\mu^\circ}\mathbf{W}^\circ\mathbf{\Sigma}_{\mu^\circ,\mu^\circ}\right) + \frac{1}{2}\operatorname{tr}\left(\mathbf{\Sigma}_{\mu^\circ,\mu^\circ}\mathbf{W}^\circ\mathbf{\Sigma}_{\mu^\circ,\mu^\circ}\mathbf{W}^\circ\mathbf{\Sigma}_{\mu^\circ,\mu^\circ}\right) \\ &= 0.\end{aligned}$$

Conversely, $\mathcal{L}_{\mathrm{TF}}(\mu, \mathbf{W}) = 0$ implies that $f^\circ(\boldsymbol{x}_{\mathrm{qr}}) = \mathbb{E}_{\boldsymbol{x}}[f^\circ(\boldsymbol{x})h_\mu(\boldsymbol{x})^\top]\mathbf{W}h_\mu(\boldsymbol{x}_{\mathrm{qr}})$ or $f^\circ = \mathbf{A}h_\mu$ for some $\mathbf{A} \in \mathbb{R}^{k\times k}$. Then the pushforward measure $\mathbf{A}\sharp\mu$ of μ along the map $(\boldsymbol{a}, \boldsymbol{w}) \mapsto (\mathbf{A}\boldsymbol{a}, \boldsymbol{w})$ satisfies

$$h_{\mathbf{A}\sharp\mu}(\boldsymbol{x}) = \int \mathbf{A}\boldsymbol{a}\sigma(\boldsymbol{w}^\top\boldsymbol{x})\mu(\mathrm{d}\theta) = \mathbf{A}h_\mu(\boldsymbol{x}) = f^\circ(\boldsymbol{x})$$

and $\operatorname{ess\,sup}_{\mathbf{A}\sharp\mu}\|\boldsymbol{a}\|\|\boldsymbol{w}\| \le \|\mathbf{A}\|\operatorname{ess\,sup}_\mu\|\boldsymbol{a}\|\|\boldsymbol{w}\| < \infty$, thus $f^\circ = h_{\mathbf{A}\sharp\mu} \in \mathscr{B}$. □

Proof of Proposition B.3. Since the minimization problem is standard linear regression, we can explicitly set

$$\boldsymbol{v} = \mathbb{E}_{\boldsymbol{x}}[f^\circ(\boldsymbol{x})f^\circ(\boldsymbol{x})^\top]^\dagger\mathbb{E}_{\boldsymbol{x}}[g(\boldsymbol{x})f^\circ(\boldsymbol{x})], \quad \|\boldsymbol{v}\| \le \|\mathbb{E}_{\boldsymbol{x}}[f^\circ(\boldsymbol{x})f^\circ(\boldsymbol{x})^\top]^\dagger\| \cdot \|f^\circ\|_{L^2(\mathcal{D}_{\mathscr{X}})}\|g\|_{L^2(\mathcal{D}_{\mathscr{X}})}.$$

Writing $g_0 = g - \boldsymbol{v}^\top f^\circ$, we can bound

$$\begin{aligned}&\frac{1}{2}\mathbb{E}_{\boldsymbol{x}_{\mathrm{qr}}}\left[\left\|g(\boldsymbol{x}_{\mathrm{qr}}) - \mathbb{E}_{\boldsymbol{x}}[g(\boldsymbol{x})h_\mu(\boldsymbol{x})^\top]\mathbf{W}h_\mu(\boldsymbol{x}_{\mathrm{qr}})\right\|^2\right] \\ &\le \|\boldsymbol{v}\|^2\mathbb{E}_{\boldsymbol{x}_{\mathrm{qr}}}\left[\left\|f^\circ(\boldsymbol{x}_{\mathrm{qr}}) - \mathbb{E}_{\boldsymbol{x}}[f^\circ(\boldsymbol{x})h_\mu(\boldsymbol{x})^\top]\mathbf{W}h_\mu(\boldsymbol{x}_{\mathrm{qr}})\right\|^2\right] \\ &\quad + \mathbb{E}_{\boldsymbol{x}_{\mathrm{qr}}}\left[\left\|g_0(\boldsymbol{x}_{\mathrm{qr}}) - \mathbb{E}_{\boldsymbol{x}}[g_0(\boldsymbol{x})h_\mu(\boldsymbol{x})^\top]\mathbf{W}h_\mu(\boldsymbol{x}_{\mathrm{qr}})\right\|^2\right] \\ &\le 2\|\boldsymbol{v}\|^2\epsilon + 2\|g_0\|^2_{L^2(\mathcal{D}_{\mathscr{X}})} + 2\|g_0\|^2_{L^2(\mathcal{D}_{\mathscr{X}})}\|h_\mu\|^4_{L^2(\mathcal{D}_{\mathscr{X}})}\|\mathbf{W}\|^2.\end{aligned}$$

The statement follows by noting that

$$\begin{aligned}\|h_\mu\|_{L^2(\mathcal{D}_{\mathscr{X}})} &\le \left(\int \mathbb{E}_{\boldsymbol{x}}[\|h_\theta(\boldsymbol{x})\|^2]\mu(\mathrm{d}\theta)\right)^{1/2} \\ &\le \left(\int \|\boldsymbol{a}\|^2\|\boldsymbol{w}\|^2\mathbb{E}_{\boldsymbol{x}}[\|\boldsymbol{x}\|^2]\mu(\mathrm{d}\theta)\right)^{1/2} = M_2^{1/2}\|h_\mu\|_{\mathscr{B}}\end{aligned}$$

from the limiting argument in Lemma B.6. □

B.4 Proofs for Appendix B.2

Proof of Proposition B.4. For any network μ, we may take $\mathbf{W} = \mathbb{E}_{\boldsymbol{x}}[f^\circ(\boldsymbol{x})h_\mu(\boldsymbol{x})^\top]^\dagger$ so that

$$\begin{aligned}\inf_{\mathbf{W}}\mathcal{L}_{\mathrm{TF}}(\mu, \mathbf{W}) &\le \frac{1}{2}\mathbb{E}_{\boldsymbol{x}_{\mathrm{qr}}}\left[\left\|f^\circ(\boldsymbol{x}_{\mathrm{qr}}) - \mathbb{E}_{\boldsymbol{x}}[f^\circ(\boldsymbol{x})h_\mu(\boldsymbol{x})^\top]\mathbf{W}h_\mu(\boldsymbol{x}_{\mathrm{qr}})\right\|^2\right] \\ &= \frac{1}{2}\mathbb{E}_{\boldsymbol{x}}[\|f^\circ(\boldsymbol{x})\|^2] - \operatorname{tr}\left(\mathbb{E}_{\boldsymbol{x}}[f^\circ(\boldsymbol{x})h_\mu(\boldsymbol{x})^\top]\mathbf{W}\mathbb{E}_{\boldsymbol{x}}[h_\mu(\boldsymbol{x})f^\circ(\boldsymbol{x})^\top]\right) \\ &\quad + \frac{1}{2}\operatorname{tr}\left(\mathbf{W}^\top\mathbb{E}_{\boldsymbol{x}}[h_\mu(\boldsymbol{x})f^\circ(\boldsymbol{x})^\top]\mathbb{E}_{\boldsymbol{x}}[f^\circ(\boldsymbol{x})h_\mu(\boldsymbol{x})^\top]\mathbf{W}\mathbf{\Sigma}_{\mu,\mu}\right) \\ &\le \frac{1}{2}\mathbb{E}_{\boldsymbol{x}}[\|f^\circ(\boldsymbol{x})\|^2] - \operatorname{tr}\left(\mathbb{E}_{\boldsymbol{x}}[f^\circ(\boldsymbol{x})h_\mu(\boldsymbol{x})^\top]\right) + \frac{1}{2}\operatorname{tr}\mathbf{\Sigma}_{\mu,\mu} \\ &= \frac{1}{2}\mathbb{E}_{\boldsymbol{x}}[\|f^\circ(\boldsymbol{x}) - h_\mu(\boldsymbol{x})\|^2].\end{aligned}$$

Now let $\mu^\circ \in \mathscr{P}(\Theta)$ be a distribution such that $f^\circ = h_{\mu^\circ}$ and $\int\|\boldsymbol{a}\|^2\|\boldsymbol{w}\|^2\mu^\circ(\mathrm{d}\theta) \le (1+\epsilon)\|f^\circ\|^2_{\mathscr{B}}$. Let $\theta^{(1)}, \cdots, \theta^{(N)}$ be an i.i.d. sample from μ°. Then from $\mathbb{E}_{\theta\sim\mu^\circ}[h_\theta(\boldsymbol{x})] = h_{\mu^\circ}(\boldsymbol{x})$, it holds on average that

$$\mathbb{E}_{\widehat{\mu}_N}\mathbb{E}_{\boldsymbol{x}}[\|\widehat{h}_N(\boldsymbol{x}) - f^\circ(\boldsymbol{x})\|^2]$$

$$
\begin{aligned}
&= \mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\widehat{\mu}_N} \left[\left\| \frac{1}{N} \sum_{j=1}^{N} h_{\theta^{(j)}}(\boldsymbol{x}) - f^\circ(\boldsymbol{x}) \right\|^2 \right] \\
&= \frac{1}{N^2} \sum_{j=1}^{N} \mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\widehat{\mu}_N} \left[\| h_{\theta^{(j)}}(\boldsymbol{x}) - h_{\mu^\circ}(\boldsymbol{x}) \|^2 \right] \\
&\quad + \frac{1}{N^2} \sum_{j \neq \ell} \mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\widehat{\mu}_N} \left[(h_{\theta^{(j)}}(\boldsymbol{x}) - h_{\mu^\circ}(\boldsymbol{x}))^\top (h_{\theta^{(\ell)}}(\boldsymbol{x}) - h_{\mu^\circ}(\boldsymbol{x})) \right] \\
&\leq \frac{1}{N^2} \sum_{j=1}^{N} \mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\theta^{(j)} \sim \mu^\circ} [\| h_{\theta^{(j)}}(\boldsymbol{x}) \|^2] \\
&\leq \frac{1}{N} \int \|\boldsymbol{a}\|^2 \mathbb{E}_{\boldsymbol{x}} [(\boldsymbol{w}^\top \boldsymbol{x})^2] \mu^\circ(\mathrm{d}\theta) \\
&\leq \frac{(1+\epsilon) M_2}{N} \|f^\circ\|_{\mathscr{B}}^2.
\end{aligned}
$$

Moreover, the path norm is bounded on average as $\mathbb{E}_{\widehat{\mu}_N}[\|\widehat{h}_N\|_{\mathcal{P}}] \leq (1+\epsilon)\|f^\circ\|_{\mathscr{B}}$. Then by Markov's inequality, the event $\|\widehat{h}_N - f^\circ\|_{L^2(\mathcal{D}_{\mathscr{X}})}^2 > \frac{2M_2\|f^\circ\|_{\mathscr{B}}^2}{N}$ has probability at most $\frac{1+\epsilon}{2}$, and the event $\|\widehat{h}_N\|_{\mathcal{P}} > 3\|f^\circ\|_{\mathscr{B}}$ has probability at most $\frac{1+\epsilon}{3}$. Hence the stated bounds hold with positive probability as $\epsilon \to 0$, thus for some size N network $\widehat{\mu}_N$. □

For the propagation of chaos result, we require the following bounds.

Lemma B.7. *The second moment* $m_2(\mu_t) = \int \|\theta\|^2 \mu_t(\mathrm{d}\theta)$ *satisfies* $m_2(\mu_t) \leq e^{2L_1^2} m_2(\mu_0)$.

Proof. The assertion follows immediately from

$$
\frac{\mathrm{d}}{\mathrm{d}t} m_2(\mu_t) = \int \|\theta\|^2 \partial_t \mu_t(\mathrm{d}\theta) = -2 \int \theta^\top \nabla \frac{\delta F}{\delta \mu}(\mu_t, \theta) \mu_t(\mathrm{d}\theta) \leq 2L_1^2 m_2(\mu_t).
$$

□

Lemma B.8. *Let* $\mu \in \mathscr{P}_2(\Omega)$ *and* $\theta^{(1)}, \cdots, \theta^{(N)}$ *be an i.i.d. sample from* μ *with corresponding empirical distribution* $\widehat{\mu}_N = \frac{1}{N}\sum_{j=1}^{N} \delta_{\theta^{(j)}}$. *Then for dimension* $m \geq 3$ *it holds that* $\mathbb{E}[\mathcal{W}_1(\mu, \widehat{\mu}_N)] \leq C_m \cdot m_2(\mu)^{1/2} N^{-1/m}$. *The rate is replaced by* $N^{-1/2} \log N$ *if* $m = 2$ *and* $N^{-1/2}$ *if* $m = 1$.

Proof. See e.g. Fournier & Guillin (2015) for the case $m \geq 2$ and Bobkov & Ledoux (2019) for $m = 1$. □

Proof of Proposition B.5. Consider the coupled process

$$
\frac{\mathrm{d}}{\mathrm{d}t} \tilde{\theta}_t^{(j)} = -\nabla \frac{\delta F}{\delta \mu}(\mu_t, \tilde{\theta}_t^{(j)}), \quad \tilde{\theta}_0^{(j)} = \theta_0^{(j)}, \quad j \in [N]
$$

and write the corresponding empirical distribution as $\tilde{\mu}_{t,N} = \frac{1}{N}\sum_{j=1}^{N} \delta_{\tilde{\theta}_t^{(j)}}$. For any finite time horizon $T \geq 0$, it holds that

$$
\begin{aligned}
\frac{1}{N}\sum_{j=1}^{N} \|\theta_T^{(j)} - \tilde{\theta}_T^{(j)}\| &= \frac{1}{N}\sum_{j=1}^{N} \left\| \int_0^T \nabla \frac{\delta F}{\delta \mu}(\widehat{\mu}_{t,N}, \theta_t^{(j)}) - \nabla \frac{\delta F}{\delta \mu}(\mu_t, \tilde{\theta}_t^{(j)}) \, \mathrm{d}t \right\| \\
&\leq \int_0^T \frac{L_2}{N} \sum_{j=1}^{N} \|\theta_t^{(j)} - \tilde{\theta}_t^{(j)}\| + L_3 \, \mathcal{W}_1(\widehat{\mu}_{t,N}, \mu_t) \, \mathrm{d}t.
\end{aligned}
$$

Then applying Gronwall's inequality and taking the expectation over random initialization, we have for all $t \in [0, T]$

$$
\mathbb{E}[\mathcal{W}_1(\widehat{\mu}_{t,N}, \tilde{\mu}_{t,N})] \leq \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N} \|\theta_t^{(j)} - \tilde{\theta}_t^{(j)}\| \right] \leq L_3 e^{L_2 T} \int_0^t \mathbb{E}[\mathcal{W}_1(\widehat{\mu}_{s,N}, \mu_s)] \, \mathrm{d}s.
$$

Since each trajectory $\tilde{\theta}_t^{(j)}$ of the coupled process is an independent sample from the true distribution μ_t, by Lemma B.7 and B.8 it moreover holds that

$$\begin{aligned}\mathbb{E}[\mathcal{W}_1(\widehat{\mu}_{t,N}, \mu_t)] &\le \mathbb{E}[\mathcal{W}_1(\widehat{\mu}_{t,N}, \tilde{\mu}_{t,N})] + \mathbb{E}[\mathcal{W}_1(\tilde{\mu}_{t,N}, \mu_t)] \\ &\le L_3 e^{L_2 T} \int_0^t \mathbb{E}[\mathcal{W}_1(\widehat{\mu}_{s,N}, \mu_s)] \,\mathrm{d}s + C_m e^{L_1^2} m_2(\mu_0)^{1/2} N^{-1/m}\end{aligned}$$

with the appropriate modification when $m = 1, 2$. Hence another application of Gronwall's inequality yields

$$\mathbb{E}[\mathcal{W}_1(\widehat{\mu}_{t,N}, \mu_t)] \le C_m m_2(\mu_0)^{1/2} N^{-1/m} \exp(L_1^2 + L_3 T e^{L_2 T}) \to 0$$

as $N \to \infty$. The convergence is uniform for any finite horizon T. □

Remark B.9. When $F = \mathcal{L}$, we rely on the Lipschitz constants obtained in Lemma E.4 to obtain the same statement, with the caveat that the flow must not reach the singular set $\mathscr{P}_2^0(\Theta)$ in order to ensure existence and regularity of the flow; this will be a recurring issue. The result is clearly still valid for mean-field dynamics incorporating birth-death by the ordinary law of large numbers, assuming the update happens at the same instant for $\widehat{\mu}_t$ and μ_t. See also Rotskoff et al. (2019) for a more involved study of birth-death dynamics.

Remark B.10. The above bounds are not optimized; compare for example Berthier et al. (2023). Explicit *uniform-in-time* propagation of chaos bounds have recently been proved for convex mean-field Langevin dynamics (Chen et al., 2022; Suzuki et al., 2023) and convex-concave descent-ascent dynamics (Kim et al., 2024). It remains an open problem to prove such results for general nonconvex mean-field dynamics, with or without the entropic regularization framework.

C Results and Proofs for Section 3

C.1 Auxiliary Results

We will use the following elementary results from linear algebra without proof.

Lemma C.1. *The spectral norm of a block matrix* $\mathbf{A} = \begin{bmatrix} \mathbf{A}_{1,1} & \mathbf{A}_{1,2} \\ \mathbf{A}_{2,1} & \mathbf{A}_{2,2} \end{bmatrix}$ *is bounded as* $\|\mathbf{A}\| \le \sum_{i,j=1}^2 \|\mathbf{A}_{i,j}\|$.

Lemma C.2. *The spectral and nuclear norms are dual:* $\|\mathbf{A}\|_* = \max_{\|\mathbf{B}\| \le 1} \langle \mathbf{A}, \mathbf{B} \rangle$ *and* $\|\mathbf{A}\| = \max_{\|\mathbf{B}\|_* \le 1} \langle \mathbf{A}, \mathbf{B} \rangle$ *for any* $\mathbf{A} \in \mathbb{R}^{m \times m}$, $m \ge 1$. *In particular,* $\operatorname{tr}(\mathbf{A}^\top \mathbf{B}) \le \|\mathbf{A}\| \|\mathbf{B}\|_*$ *for any* $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times m}$.

Lemma C.3. *For a positive semi-definite matrix* $\mathbf{A} \in \mathbb{R}^{k \times k}$ *it holds that* $\frac{1}{k}(\operatorname{tr} \mathbf{A})^2 \le \operatorname{tr} \mathbf{A}^2 \le (\operatorname{tr} \mathbf{A})^2$.

The neural network output is continuous and well-behaved in the following sense:

Lemma C.4. *The map* $\theta \mapsto h_\theta(\boldsymbol{x})$ *on* Θ *is* $(R_1^2 + R_2^2 \|\boldsymbol{x}\|^2)^{1/2}$*-Lipschitz for each* $\boldsymbol{x} \in \mathscr{X}$. *Also, the map* $\mu \mapsto h_\mu(\boldsymbol{x})$ *on* $\mathscr{P}_2(\Theta)$ *is* $(kR_1^2 + kR_2^2 \|\boldsymbol{x}\|^2)^{1/2}$*-Lipschitz w.r.t. 1-Wasserstein distance for each* $\boldsymbol{x} \in \mathscr{X}$.

Proof. For $\theta_1 = (\boldsymbol{a}_1, \boldsymbol{w}_1), \theta_2 = (\boldsymbol{a}_2, \boldsymbol{w}_2)$ we have

$$\begin{aligned}\|h_{\theta_1}(\boldsymbol{x}) - h_{\theta_2}(\boldsymbol{x})\| &= \|\boldsymbol{a}_1 \sigma(\boldsymbol{w}_1^\top \boldsymbol{x}) - \boldsymbol{a}_2 \sigma(\boldsymbol{w}_2^\top \boldsymbol{x})\| \\ &\le \|\boldsymbol{a}_1 - \boldsymbol{a}_2\| \cdot |\sigma(\boldsymbol{w}_1^\top \boldsymbol{x})| + \|\boldsymbol{a}_2\| \cdot |\sigma(\boldsymbol{w}_1^\top \boldsymbol{x}) - \sigma(\boldsymbol{w}_2^\top \boldsymbol{x})| \\ &\le R_1 \|\boldsymbol{a}_1 - \boldsymbol{a}_2\| + R_2 \|\boldsymbol{w}_1 - \boldsymbol{w}_2\| \cdot \|\boldsymbol{x}\| \\ &\le (R_1^2 + R_2^2 \|\boldsymbol{x}\|^2)^{1/2} \|\theta_1 - \theta_2\|.\end{aligned}$$

The difference of each coordinate $|h_{\theta_1}(\boldsymbol{x})_j - h_{\theta_2}(\boldsymbol{x})_j|$ satisfies the same bound for $1 \le j \le k$, implying that

$$|h_\mu(\boldsymbol{x})_j - h_\nu(\boldsymbol{x})_j| = \left| \int_\Theta h_\theta(\boldsymbol{x})_j \mu(\mathrm{d}\theta) - \int_\Theta h_\theta(\boldsymbol{x})_j \nu(\mathrm{d}\theta) \right| \le (R_1^2 + R_2^2 \|\boldsymbol{x}\|^2)^{1/2} \, \mathcal{W}_1(\mu, \nu)$$

and hence $\|h_\mu(\boldsymbol{x}) - h_\nu(\boldsymbol{x})\| \le (kR_1^2 + kR_2^2 \|\boldsymbol{x}\|^2)^{1/2} \, \mathcal{W}_1(\mu, \nu)$. □

Proof of Lemma 3.1. The gradient flow equation for $\mathbf{W}$ is given as

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{W}_t &= -\frac{1}{2}\nabla_{\mathbf{W}}\,\mathrm{tr}\left(-2\mathbf{\Sigma}_{\mu^\circ,\mu}\mathbf{W}\mathbf{\Sigma}_{\mu,\mu^\circ} + \mathbf{\Sigma}_{\mu^\circ,\mu}\mathbf{W}\mathbf{\Sigma}_{\mu,\mu}\mathbf{W}^\top\mathbf{\Sigma}_{\mu,\mu^\circ}\right)\\ &= -\mathbf{\Sigma}_{\mu,\mu^\circ}\mathbf{\Sigma}_{\mu^\circ,\mu}(\mathbf{W}_t\mathbf{\Sigma}_{\mu,\mu} - \mathbf{I}_k).\end{aligned}$$

Denote the singular value decomposition of $\mathbf{\Sigma}_{\mu,\mu^\circ}$ as $\mathbf{U}_1\mathbf{D}_1\mathbf{V}_1^\top$ and the spectral decomposition of $\mathbf{\Sigma}_{\mu,\mu}$ as $\mathbf{U}_2\mathbf{D}_2\mathbf{U}_2^\top$ where $\mathbf{U}_1,\mathbf{U}_2,\mathbf{V}_1 \in \mathcal{O}(k)$ and $\mathbf{D}_j = \mathrm{diag}(d_{j,1},\cdots,d_{j,k})$. Since we assume $\mathbf{\Sigma}_{\mu,\mu} = \mathbb{E}_{\boldsymbol{x}}[h_\mu(\boldsymbol{x})h_\mu(\boldsymbol{x})^\top]$ is positive definite, we also have $b_{2,i} > 0$ for all i. Further defining the auxiliary matrix $\mathbf{Z}_t = \mathbf{U}_1^\top\mathbf{W}_t\mathbf{U}_2$, the dynamics for $\mathbf{Z}_t$ is expressed as

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{Z}_t &= -\mathbf{U}_1^\top(\mathbf{U}_1\mathbf{D}_1\mathbf{V}_1^\top)(\mathbf{V}_1\mathbf{D}_1\mathbf{U}_1^\top)(\mathbf{U}_1\mathbf{Z}_t\mathbf{D}_2\mathbf{U}_2^\top - \mathbf{I}_k)\mathbf{U}_2\\ &= -\mathbf{D}_1^2\mathbf{Z}_t\mathbf{D}_2 + \mathbf{D}_1^2\mathbf{U}_1^\top\mathbf{U}_2.\end{aligned}$$

Writing $\mathbf{U}_1^\top\mathbf{U}_2 = (u_{i,j})_{1\le i,j\le k}$, for each entry $z_{i,j}(t) := (\mathbf{Z}_t)_{i,j}$ we obtain that $z'_{i,j}(t) = -d_{1,i}^2(d_{2,j}z_{i,j}(t) - u_{i,j})$ and therefore

$$\lim_{t\to\infty} z_{i,j}(t) = \begin{Bmatrix} d_{2,j}^{-1}u_{i,j} & d_{1,i}\neq 0\\ z_{i,j}(0) & d_{1,i}=0\end{Bmatrix} = \mathbf{1}_{\{d_{1,i}\neq 0\}}d_{2,j}^{-1}u_{i,j} + \mathbf{1}_{\{d_{1,i}=0\}}z_{i,j}(0).$$

This can be recast in matrix form as $\lim_{t\to\infty}\mathbf{Z}_t = \mathbf{D}_1^\dagger\mathbf{D}_1\mathbf{U}_1^\top\mathbf{U}_2\mathbf{D}_2^{-1} + (\mathbf{I}_k - \mathbf{D}_1^\dagger\mathbf{D}_1)\mathbf{Z}_0$, and the convergence rate is exponential. We conclude for the limit $\mathbf{W}_\mu := \lim_{t\to\infty}\mathbf{W}_t$ that

$$\begin{aligned}\mathbf{\Sigma}_{\mu^\circ,\mu}\mathbf{W}_\mu &= (\mathbf{V}_1\mathbf{D}_1\mathbf{U}_1^\top)\mathbf{U}_1\big(\mathbf{D}_1^\dagger\mathbf{D}_1\mathbf{U}_1^\top\mathbf{U}_2\mathbf{D}_2^{-1} + (\mathbf{I}_k - \mathbf{D}_1^\dagger\mathbf{D}_1)\mathbf{Z}_0\big)\mathbf{U}_2^\top\\ &= (\mathbf{V}_1\mathbf{D}_1\mathbf{U}_1^\top)(\mathbf{U}_2\mathbf{D}_2^{-1}\mathbf{U}_2^\top)\\ &= \mathbf{\Sigma}_{\mu^\circ,\mu}\mathbf{\Sigma}_{\mu,\mu}^{-1}.\end{aligned}$$

□

Proposition C.5. *For any $\mu \in \mathscr{P}_2(\Theta)$, $\nu \in \mathscr{P}_2^+(\Theta)$, there are at most k values $t \in [0,1]$ such that $(1-t)\mu + t\nu \in \mathscr{P}_2^0(\Theta)$. Consequently, $\mathscr{P}_2^+(\Theta)$ is dense in $\mathscr{P}_2(\Theta)$.*

Note in particular that $\mathbf{R}\sharp\mu^\circ \in \mathscr{P}_2^+(\Theta)$ for any invertible $\mathbf{R} \in \mathcal{B}_1(k)$ as $\mathbf{\Sigma}_{\mathbf{R}\sharp\mu^\circ,\mathbf{R}\sharp\mu^\circ} = r^\circ\mathbf{R}\mathbf{R}^\top$. This justifies the computations which appear in the statement and proof of Theorem 3.3.

Proof. Suppose there exist $k+1$ distinct $t_j \in [0,1]$, $j = 0,1,\cdots,k$ such that $(1-t_j)\mu + t_j\nu \in \mathscr{P}_2^0(\Theta)$; note that $t_j \neq 1$ since $\nu \in \mathscr{P}_2^+(\Theta)$. Then there exist nonzero vectors $\boldsymbol{z}_j$ such that $(1-t_j)\boldsymbol{z}_j^\top h_\mu(\boldsymbol{x}) + t_j\boldsymbol{z}_j^\top h_\nu(\boldsymbol{x}) \equiv 0$, and which must be linearly dependent. Without loss of generality, let $\{\boldsymbol{z}_j\}_{j=0}^\ell$ be a minimally dependent subset of $\{\boldsymbol{z}_j\}_{j=0}^k$ so that $\sum_{j=0}^\ell b_j\boldsymbol{z}_j = 0$ for constants b_j not all zero. Suppose $b_0 \neq 0$. Then the equality

$$0 \equiv \sum_{j=0}^\ell b_j\boldsymbol{z}_j^\top h_\mu(\boldsymbol{x}) + \frac{t_jb_j}{1-t_j}\boldsymbol{z}_j^\top h_\nu(\boldsymbol{x}) = \left(\sum_{j=0}^\ell \frac{t_jb_j}{1-t_j}\boldsymbol{z}_j^\top\right)h_\nu(\boldsymbol{x})$$

implies that

$$\sum_{j=0}^{\ell-1}\left(\frac{t_j}{1-t_j} - \frac{t_\ell}{1-t_\ell}\right)b_j\boldsymbol{z}_j = \sum_{j=0}^\ell \frac{t_jb_j}{1-t_j}\boldsymbol{z}_j - \frac{t_\ell}{1-t_\ell}\sum_{j=0}^\ell b_j\boldsymbol{z}_j = 0,$$

which contradicts the minimality of $\{\boldsymbol{z}_j\}_{j=0}^\ell$ since the coefficient of $\boldsymbol{z}_0$ is nonzero. This proves the first claim. Denseness of $\mathscr{P}_2^+(\Theta)$ immediately follows: for any $\mu \in \mathscr{P}_2(\Theta)$, all but finitely many mixture distributions $(1-t)\mu + t\mu^\circ$ lie in $\mathscr{P}_2^+(\Theta)$, so there exists a subsequence weakly converging to μ in $\mathscr{P}_2^+(\Theta)$. □

Lemma C.6. *Any element $\mathbf{R} \in \mathcal{B}_1(k)$ can be expressed as a convex combination of finitely many elements $\mathbf{R}_1,\cdots,\mathbf{R}_m$ of $\mathcal{O}(k)$. In particular, the pushforward can be defined for any $\mathbf{R} \in \mathcal{B}_1(k)$.*

Proof. Denote the singular value decomposition of $\mathbf{R}$ as $\mathbf{U}\mathbf{D}\mathbf{V}^\top$ and denote by $\mathbf{D}_1, \cdots, \mathbf{D}_{2^k}$ all diagonal matrices with every diagonal element equal to ± 1. Since every diagonal element of $\mathbf{D}$ has absolute value at most 1, $\mathbf{D}$ is contained in the convex hull of $\mathbf{D}_1, \cdots, \mathbf{D}_{2^k}$ and hence $\mathbf{R}$ can be written a convex combination of $\mathbf{U}\mathbf{D}_1\mathbf{V}^\top, \cdots, \mathbf{U}\mathbf{D}_{2^k}\mathbf{V}^\top \in \mathcal{O}(k)$.

Furthermore, writing $\mathbf{R} = \sum_{j=1}^m \alpha_j \mathbf{R}_j$ for $\alpha_j \in (0,1)$, $\sum_{j=1}^m \alpha_j = 1$ we may define for all $\mu \in \mathscr{P}_2(\Theta)$ the pushforward measure $\mathbf{R}\sharp\mu := \sum_{j=1}^m \alpha_j \mathbf{R}_j \sharp \mu$ so that

$$h_{\mathbf{R}\sharp\mu}(\boldsymbol{x}) = \sum_{j=1}^m \alpha_j \int_\Theta h_\theta(\boldsymbol{x}) \,\mathrm{d}\mathbf{R}_j\sharp\mu(\theta) = \sum_{j=1}^m \alpha_j \int_\Theta \mathbf{R}_j h_\theta(\boldsymbol{x}) \,\mathrm{d}\mu(\theta) = \mathbf{R}h_\mu(\boldsymbol{x}).$$

We remark that simply defining $\mathbf{R}\sharp\mu$ as the pushforward along the map $\mathbf{R} : (\boldsymbol{a}, \boldsymbol{w}) \mapsto (\mathbf{R}\boldsymbol{a}, \boldsymbol{w})$ would not preserve the bounded density condition (Assumption 3) for pushforwards of μ°. □

Proof of Lemma 3.2. It is straightforward to check that

$$\mathcal{L}(\mathbf{R}\sharp\mu^\circ) = \frac{1}{2}\mathbb{E}_{\boldsymbol{x}}\left[\|h_{\mu^\circ}(\boldsymbol{x}) - (\mathbf{\Sigma}_{\mu^\circ,\mu^\circ}\mathbf{R}^\top)(\mathbf{R}\mathbf{\Sigma}_{\mu^\circ,\mu^\circ}\mathbf{R}^\top)^{-1}\mathbf{R}h_{\mu^\circ}(\boldsymbol{x})\|^2\right] = 0.$$

Conversely, $\mathcal{L}(\mu) = 0$ implies that $h_{\mu^\circ}(\boldsymbol{x}) = \mathbf{\Sigma}_{\mu^\circ,\mu}\mathbf{\Sigma}_{\mu,\mu}^{-1}h_\mu(\boldsymbol{x})$ a.e. Since $\boldsymbol{x} \mapsto h_\mu(\boldsymbol{x})$ is always continuous, equality holds for all $\boldsymbol{x} \in \mathscr{X}$. Finally, $\mathbf{\Sigma}_{\mu^\circ,\mu}\mathbf{\Sigma}_{\mu,\mu}^{-1}$ cannot be singular since the image of h_{μ° is not constrained on a lower-dimensional subspace by Assumption 2. □

C.2 Proof of Theorem 3.3

We study the first- and second-order properties of the optimization landscape for the functional $\mathcal{L}$. First note that $\mathcal{L}(\mu) \le \mathcal{L}_{\mathrm{TF}}(\mu, 0_{k\times k}) = \frac{kr^\circ}{2}$. We denote

$$\mathbf{L}_\mu = \frac{1}{2}\mathbb{E}_{\boldsymbol{x}}\left[\zeta_{\mu^\circ,\mu}(\boldsymbol{x})\zeta_{\mu^\circ,\mu}(\boldsymbol{x})^\top\right] = \frac{1}{2}r^\circ\mathbf{I}_k - \frac{1}{2}\mathbf{\Sigma}_{\mu^\circ,\mu}\mathbf{\Sigma}_{\mu,\mu}^{-1}\mathbf{\Sigma}_{\mu,\mu^\circ}$$

so that $\mathbf{L}_\mu$ is positive semi-definite and $\operatorname{tr}\mathbf{L}_\mu = \mathcal{L}(\mu)$. Let $\mathbf{R} \in \mathcal{B}_1(k)$ and $\bar\mu_s = (1-s)\mu + s\mathbf{R}\sharp\mu^\circ$ for $s \in [0,1]$. By linearity of the mean-field mapping $\mu \mapsto h_\mu$,

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}s}h_{\bar\mu_s}(\boldsymbol{x}) = \mathbf{R}h_{\mu^\circ}(\boldsymbol{x}) - h_\mu(\boldsymbol{x}), \quad \frac{\mathrm{d}}{\mathrm{d}s}\mathbf{\Sigma}_{\mu^\circ,\bar\mu_s} = r^\circ\mathbf{R}^\top - \mathbf{\Sigma}_{\mu^\circ,\mu},\\
&\frac{\mathrm{d}}{\mathrm{d}s}\mathbf{\Sigma}_{\bar\mu_s,\bar\mu_s} = 2r^\circ s\mathbf{I}_k + (1-2s)(\mathbf{R}\mathbf{\Sigma}_{\mu^\circ,\mu} + \mathbf{\Sigma}_{\mu,\mu^\circ}\mathbf{R}^\top) - 2(1-s)\mathbf{\Sigma}_{\mu,\mu}.
\end{aligned}$$

Then the time derivative of $\mathcal{L}(\bar\mu_s)$ for $s \in [0,1]$ is obtained as

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}s}\mathcal{L}(\bar\mu_s) &= -\mathbb{E}_{\boldsymbol{x}}\left[\zeta_{\mu^\circ,\bar\mu_s}(\boldsymbol{x})^\top\frac{\mathrm{d}}{\mathrm{d}s}\left(\mathbf{\Sigma}_{\mu^\circ,\bar\mu_s}\mathbf{\Sigma}_{\bar\mu_s,\bar\mu_s}^{-1}h_{\bar\mu_s}(\boldsymbol{x})\right)\right]\\
&= -\mathbb{E}_{\boldsymbol{x}}\left[\zeta_{\mu^\circ,\bar\mu_s}(\boldsymbol{x})^\top\mathbf{\Sigma}_{\mu^\circ,\bar\mu_s}\mathbf{\Sigma}_{\bar\mu_s,\bar\mu_s}^{-1}(\mathbf{R}h_{\mu^\circ}(\boldsymbol{x}) - h_\mu(\boldsymbol{x}))\right],
\end{aligned}$$

where we have used that

$$\mathbb{E}_{\boldsymbol{x}}\left[h_{\bar\mu_s}(\boldsymbol{x})\zeta_{\mu^\circ,\bar\mu_s}(\boldsymbol{x})^\top\right] = \mathbb{E}_{\boldsymbol{x}}\left[h_{\bar\mu_s}(\boldsymbol{x})(h_{\mu^\circ}(\boldsymbol{x})^\top - h_{\bar\mu_s}(\boldsymbol{x})^\top\mathbf{\Sigma}_{\bar\mu_s,\bar\mu_s}^{-1}\mathbf{\Sigma}_{\bar\mu_s,\mu^\circ})\right] = 0.$$

In particular, the derivative at $s = 0$ is equal to

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}s}\bigg|_{s=0}\mathcal{L}(\bar\mu_s) &= -\mathbb{E}_{\boldsymbol{x}}\left[\zeta_{\mu^\circ,\mu}(\boldsymbol{x})^\top\mathbf{\Sigma}_{\mu^\circ,\mu}\mathbf{\Sigma}_{\mu,\mu}^{-1}(\mathbf{R}h_{\mu^\circ}(\boldsymbol{x}) - h_\mu(\boldsymbol{x}))\right]\\
&= -\mathbb{E}_{\boldsymbol{x}}\left[\zeta_{\mu^\circ,\mu}(\boldsymbol{x})^\top\mathbf{\Sigma}_{\mu^\circ,\mu}\mathbf{\Sigma}_{\mu,\mu}^{-1}\mathbf{R}\zeta_{\mu^\circ,\mu}(\boldsymbol{x})\right]\\
&= -2\operatorname{tr}\left(\mathbf{R}\mathbf{L}_\mu\mathbf{\Sigma}_{\mu^\circ,\mu}\mathbf{\Sigma}_{\mu,\mu}^{-1}\right).
\end{aligned}$$

We may choose the pushforward $\mathbf{R}$ so that this quantity is minimized over $\mathbf{R} \in \mathcal{B}_1(k)$. Via duality of the spectral and nuclear norms, this yields

$$\frac{\mathrm{d}}{\mathrm{d}s}\bigg|_{s=0}\mathcal{L}(\bar\mu_s) = \min_{\|\mathbf{R}\|\le 1} -2\operatorname{tr}\left(\mathbf{R}\mathbf{L}_\mu\mathbf{\Sigma}_{\mu^\circ,\mu}\mathbf{\Sigma}_{\mu,\mu}^{-1}\right) = -2\left\|\mathbf{L}_\mu\mathbf{\Sigma}_{\mu^\circ,\mu}\mathbf{\Sigma}_{\mu,\mu}^{-1}\right\|_* \le 0, \tag{4}$$

proving the first claim.

Now if the above first order analysis does not yield a direction of improvement (strict decrease) for $\mathcal{L}$, it must be the case that $\mathbf{L}_\mu \mathbf{\Sigma}_{\mu^\circ,\mu} \mathbf{\Sigma}_{\mu,\mu}^{-1} = 0$. If μ is not a global minimum then $\mathbf{L}_\mu \neq 0$ and hence $\operatorname{rank} \mathbf{\Sigma}_{\mu^\circ,\mu} \mathbf{\Sigma}_{\mu,\mu}^{-1} < k$, so that the linear regression predictions $\mathbf{\Sigma}_{\mu^\circ,\mu} \mathbf{\Sigma}_{\mu,\mu}^{-1} h_\mu(\boldsymbol{x})$ are contained in a lower-dimensional subspace $\{\boldsymbol{z}\}^\perp$ for some $\boldsymbol{z} \in \mathbb{S}^{k-1}$. This further implies that

$$\mathcal{L}(\mu) \geq \frac{1}{2} \mathbb{E}_{\boldsymbol{x}} \left[(\boldsymbol{z}^\top h_{\mu^\circ}(\boldsymbol{x}))^2 \right] = \frac{1}{2} \boldsymbol{z}^\top \mathbf{\Sigma}_{\mu^\circ,\mu^\circ} \boldsymbol{z} = \frac{r^\circ}{2},$$

confirming the critical point lower bound.

We proceed to analyze the second-order stability of critical points. The second derivative along any pushforward $\mathbf{R} \in \mathcal{B}_1(k)$ is computed as

$$\begin{aligned}
\left. \frac{\mathrm{d}^2}{\mathrm{d}s^2} \right|_{s=0} \mathcal{L}(\bar{\mu}_s) &= -\left. \frac{\mathrm{d}}{\mathrm{d}s} \right|_{s=0} \mathbb{E}_{\boldsymbol{x}} \left[\zeta_{\mu^\circ,\bar{\mu}_s}(\boldsymbol{x})^\top \mathbf{\Sigma}_{\mu^\circ,\bar{\mu}_s} \mathbf{\Sigma}_{\bar{\mu}_s,\bar{\mu}_s}^{-1} (\mathbf{R} h_{\mu^\circ}(\boldsymbol{x}) - h_\mu(\boldsymbol{x})) \right] \\
&= \mathbb{E}_{\boldsymbol{x}} \left[\left. \frac{\mathrm{d}}{\mathrm{d}s} \right|_{s=0} \left(\mathbf{\Sigma}_{\mu^\circ,\bar{\mu}_s} \mathbf{\Sigma}_{\bar{\mu}_s,\bar{\mu}_s}^{-1} h_{\bar{\mu}_s}(\boldsymbol{x}) \right)^\top \mathbf{\Sigma}_{\mu^\circ,\bar{\mu}_s} \mathbf{\Sigma}_{\bar{\mu}_s,\bar{\mu}_s}^{-1} (\mathbf{R} h_{\mu^\circ}(\boldsymbol{x}) - h_\mu(\boldsymbol{x})) \right] \\
&\quad - \mathbb{E}_{\boldsymbol{x}} \left[\zeta_{\mu^\circ,\bar{\mu}_s}(\boldsymbol{x})^\top \left. \frac{\mathrm{d}}{\mathrm{d}s} \right|_{s=0} \mathbf{\Sigma}_{\mu^\circ,\bar{\mu}_s} \mathbf{\Sigma}_{\bar{\mu}_s,\bar{\mu}_s}^{-1} (\mathbf{R} h_{\mu^\circ}(\boldsymbol{x}) - h_\mu(\boldsymbol{x})) \right].
\end{aligned}$$

The first term can be expanded as

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}} \Big[(\mathbf{R} h_{\mu^\circ}(\boldsymbol{x}) - h_\mu(\boldsymbol{x}))^\top \mathbf{\Sigma}_{\mu,\mu}^{-1} \mathbf{\Sigma}_{\mu,\mu^\circ} \mathbf{\Sigma}_{\mu^\circ,\mu} \mathbf{\Sigma}_{\mu,\mu}^{-1} (\mathbf{R} h_{\mu^\circ}(\boldsymbol{x}) - h_\mu(\boldsymbol{x})) \\
&\quad - h_\mu(\boldsymbol{x})^\top \mathbf{\Sigma}_{\mu,\mu}^{-1} (\mathbf{R} \mathbf{\Sigma}_{\mu^\circ,\mu} + \mathbf{\Sigma}_{\mu,\mu^\circ} \mathbf{R}^\top - 2\mathbf{\Sigma}_{\mu,\mu}) \mathbf{\Sigma}_{\mu,\mu}^{-1} \mathbf{\Sigma}_{\mu,\mu^\circ} \mathbf{\Sigma}_{\mu^\circ,\mu} \mathbf{\Sigma}_{\mu,\mu}^{-1} (\mathbf{R} h_{\mu^\circ}(\boldsymbol{x}) - h_\mu(\boldsymbol{x})) \\
&\quad + h_\mu(\boldsymbol{x})^\top \mathbf{\Sigma}_{\mu,\mu}^{-1} (r^\circ \mathbf{R} - \mathbf{\Sigma}_{\mu,\mu^\circ}) \mathbf{\Sigma}_{\mu^\circ,\mu} \mathbf{\Sigma}_{\mu,\mu}^{-1} (\mathbf{R} h_{\mu^\circ}(\boldsymbol{x}) - h_\mu(\boldsymbol{x})) \Big] \\
&= \mathbb{E}_{\boldsymbol{x}} \Big[(\mathbf{R} h_{\mu^\circ}(\boldsymbol{x}) + h_\mu(\boldsymbol{x}))^\top \mathbf{\Sigma}_{\mu,\mu}^{-1} \mathbf{\Sigma}_{\mu,\mu^\circ} \mathbf{\Sigma}_{\mu^\circ,\mu} \mathbf{\Sigma}_{\mu,\mu}^{-1} (\mathbf{R} h_{\mu^\circ}(\boldsymbol{x}) - h_\mu(\boldsymbol{x})) \\
&\quad - h_\mu(\boldsymbol{x})^\top \mathbf{\Sigma}_{\mu,\mu}^{-1} (\mathbf{R} \mathbf{\Sigma}_{\mu^\circ,\mu} \mathbf{\Sigma}_{\mu,\mu}^{-1} \mathbf{\Sigma}_{\mu,\mu^\circ} + \mathbf{\Sigma}_{\mu,\mu^\circ} \mathbf{R}^\top \mathbf{\Sigma}_{\mu,\mu}^{-1} \mathbf{\Sigma}_{\mu,\mu^\circ}) \mathbf{\Sigma}_{\mu^\circ,\mu} \mathbf{\Sigma}_{\mu,\mu}^{-1} (\mathbf{R} h_{\mu^\circ}(\boldsymbol{x}) - h_\mu(\boldsymbol{x})) \\
&\quad + h_\mu(\boldsymbol{x})^\top \mathbf{\Sigma}_{\mu,\mu}^{-1} (r^\circ \mathbf{R} - \mathbf{\Sigma}_{\mu,\mu^\circ}) \mathbf{\Sigma}_{\mu^\circ,\mu} \mathbf{\Sigma}_{\mu,\mu}^{-1} (\mathbf{R} h_{\mu^\circ}(\boldsymbol{x}) - h_\mu(\boldsymbol{x})) \Big] \\
&= r^\circ \operatorname{tr} \left(\mathbf{\Sigma}_{\mu,\mu}^{-1} \mathbf{\Sigma}_{\mu,\mu^\circ} \mathbf{\Sigma}_{\mu^\circ,\mu} \mathbf{\Sigma}_{\mu,\mu}^{-1} \mathbf{R} \mathbf{R}^\top \right) - \operatorname{tr} \left(\mathbf{\Sigma}_{\mu^\circ,\mu} \mathbf{\Sigma}_{\mu,\mu}^{-1} \mathbf{\Sigma}_{\mu,\mu^\circ} \right) \\
&\quad + \operatorname{tr} \left((r^\circ \mathbf{R} - \mathbf{R} \mathbf{\Sigma}_{\mu^\circ,\mu} \mathbf{\Sigma}_{\mu,\mu}^{-1} \mathbf{\Sigma}_{\mu,\mu^\circ}) \mathbf{\Sigma}_{\mu^\circ,\mu} \mathbf{\Sigma}_{\mu,\mu}^{-1} (\mathbf{R} \mathbf{\Sigma}_{\mu^\circ,\mu} \mathbf{\Sigma}_{\mu,\mu}^{-1} - \mathbf{I}_k) \right) \\
&\quad - \operatorname{tr} \left((\mathbf{\Sigma}_{\mu,\mu^\circ} + \mathbf{\Sigma}_{\mu,\mu^\circ} \mathbf{R}^\top \mathbf{\Sigma}_{\mu,\mu}^{-1} \mathbf{\Sigma}_{\mu,\mu^\circ}) \mathbf{\Sigma}_{\mu^\circ,\mu} \mathbf{\Sigma}_{\mu,\mu}^{-1} (\mathbf{R} \mathbf{\Sigma}_{\mu^\circ,\mu} \mathbf{\Sigma}_{\mu,\mu}^{-1} - \mathbf{I}_k) \right) \\
&= r^\circ \operatorname{tr} \left(\mathbf{\Sigma}_{\mu,\mu}^{-1} \mathbf{\Sigma}_{\mu,\mu^\circ} \mathbf{\Sigma}_{\mu^\circ,\mu} \mathbf{\Sigma}_{\mu,\mu}^{-1} \mathbf{R} \mathbf{R}^\top \right) - kr^\circ + 2 \operatorname{tr} \mathbf{L}_\mu \\
&\quad + 2 \operatorname{tr} \left(\mathbf{R} \mathbf{L}_\mu \mathbf{\Sigma}_{\mu^\circ,\mu} \mathbf{\Sigma}_{\mu,\mu}^{-1} (\mathbf{R} \mathbf{\Sigma}_{\mu^\circ,\mu} \mathbf{\Sigma}_{\mu,\mu}^{-1} - \mathbf{I}_k) \right) \\
&\quad - \operatorname{tr} \left((r^\circ \mathbf{I}_k - 2\mathbf{L}_\mu) (\mathbf{R}^\top \mathbf{\Sigma}_{\mu,\mu}^{-1} \mathbf{\Sigma}_{\mu,\mu^\circ} + \mathbf{I}_k) (\mathbf{\Sigma}_{\mu^\circ,\mu} \mathbf{\Sigma}_{\mu,\mu}^{-1} \mathbf{R} - \mathbf{I}_k) \right) \\
&= 2 \operatorname{tr} \left(\mathbf{L}_\mu (\mathbf{\Sigma}_{\mu^\circ,\mu} \mathbf{\Sigma}_{\mu,\mu}^{-1} \mathbf{R} + \mathbf{R}^\top \mathbf{\Sigma}_{\mu,\mu}^{-1} \mathbf{\Sigma}_{\mu,\mu^\circ} - \mathbf{I}_k) \mathbf{\Sigma}_{\mu^\circ,\mu} \mathbf{\Sigma}_{\mu,\mu}^{-1} \mathbf{R} \right),
\end{aligned}$$

where we have taken advantage of the symmetry of $\mathbf{L}_\mu$ to cancel out various terms. The second term can be expanded as

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}} \Big[-\zeta_{\mu^\circ,\mu}(\boldsymbol{x})^\top (r^\circ \mathbf{R}^\top - \mathbf{\Sigma}_{\mu^\circ,\mu}) \mathbf{\Sigma}_{\mu,\mu}^{-1} (\mathbf{R} h_{\mu^\circ}(\boldsymbol{x}) - h_\mu(\boldsymbol{x})) \\
&\qquad + \zeta_{\mu^\circ,\mu}(\boldsymbol{x})^\top \mathbf{\Sigma}_{\mu^\circ,\mu} \mathbf{\Sigma}_{\mu,\mu}^{-1} (\mathbf{R} \mathbf{\Sigma}_{\mu^\circ,\mu} + \mathbf{\Sigma}_{\mu,\mu^\circ} \mathbf{R}^\top - 2\mathbf{\Sigma}_{\mu,\mu}) \mathbf{\Sigma}_{\mu,\mu}^{-1} (\mathbf{R} h_{\mu^\circ}(\boldsymbol{x}) - h_\mu(\boldsymbol{x})) \Big] \\
&= 2 \operatorname{tr} \left(\mathbf{L}_\mu (-r^\circ \mathbf{R}^\top - \mathbf{\Sigma}_{\mu^\circ,\mu} + \mathbf{\Sigma}_{\mu^\circ,\mu} \mathbf{\Sigma}_{\mu,\mu}^{-1} \mathbf{R} \mathbf{\Sigma}_{\mu^\circ,\mu} + \mathbf{\Sigma}_{\mu^\circ,\mu} \mathbf{\Sigma}_{\mu,\mu}^{-1} \mathbf{\Sigma}_{\mu,\mu^\circ} \mathbf{R}^\top) \mathbf{\Sigma}_{\mu,\mu}^{-1} \mathbf{R} \right) \\
&= -4 \operatorname{tr} \left(\mathbf{L}_\mu^2 \mathbf{R}^\top \mathbf{\Sigma}_{\mu,\mu}^{-1} \mathbf{R} \right) + 2 \operatorname{tr} \left(\mathbf{L}_\mu (\mathbf{\Sigma}_{\mu^\circ,\mu} \mathbf{\Sigma}_{\mu,\mu}^{-1} \mathbf{R} - \mathbf{I}_k) \mathbf{\Sigma}_{\mu^\circ,\mu} \mathbf{\Sigma}_{\mu,\mu}^{-1} \mathbf{R} \right).
\end{aligned}$$

Combining the above, we obtain

$$\left. \frac{\mathrm{d}^2}{\mathrm{d}s^2} \right|_{s=0} \mathcal{L}(\bar{\mu}_s)$$

$$= -4 \operatorname{tr}\left(\mathbf{L}_\mu^2 \mathbf{R}^\top \boldsymbol{\Sigma}_{\mu,\mu}^{-1} \mathbf{R}\right) + 2 \operatorname{tr}\left(\mathbf{L}_\mu (2\boldsymbol{\Sigma}_{\mu^\circ,\mu} \boldsymbol{\Sigma}_{\mu,\mu}^{-1} \mathbf{R} + \mathbf{R}^\top \boldsymbol{\Sigma}_{\mu,\mu}^{-1} \boldsymbol{\Sigma}_{\mu,\mu^\circ} - 2\mathbf{I}_k) \boldsymbol{\Sigma}_{\mu^\circ,\mu} \boldsymbol{\Sigma}_{\mu,\mu}^{-1} \mathbf{R}\right). \tag{5}$$

When $\mathbf{L}_\mu \boldsymbol{\Sigma}_{\mu^\circ,\mu} \boldsymbol{\Sigma}_{\mu,\mu}^{-1} = 0$, we may take $\mathbf{R} \in \mathcal{O}(k)$ such that $\boldsymbol{\Sigma}_{\mu^\circ,\mu} \boldsymbol{\Sigma}_{\mu,\mu}^{-1} \mathbf{R}$ is symmetric, i.e. $\mathbf{R} = \mathbf{V}\mathbf{U}^\top$ where $\mathbf{U}\mathbf{D}\mathbf{V}^\top$ is the singular value decomposition of $\boldsymbol{\Sigma}_{\mu^\circ,\mu} \boldsymbol{\Sigma}_{\mu,\mu}^{-1}$. Then the second trace term vanishes since $\boldsymbol{\Sigma}_{\mu^\circ,\mu} \boldsymbol{\Sigma}_{\mu,\mu}^{-1} \mathbf{R} \mathbf{L}_\mu = (\mathbf{L}_\mu^\top \boldsymbol{\Sigma}_{\mu^\circ,\mu} \boldsymbol{\Sigma}_{\mu,\mu}^{-1} \mathbf{R})^\top = (\mathbf{L}_\mu \boldsymbol{\Sigma}_{\mu^\circ,\mu} \boldsymbol{\Sigma}_{\mu,\mu}^{-1} \mathbf{R})^\top = 0$ and we have that

$$\left.\frac{\mathrm{d}^2}{\mathrm{d}s^2}\right|_{s=0} \mathcal{L}(\bar{\mu}_s) = -4 \operatorname{tr}\left(\mathbf{L}_\mu^2 \mathbf{R}^\top \boldsymbol{\Sigma}_{\mu,\mu}^{-1} \mathbf{R}\right) \le -\frac{4}{R_1^2} \operatorname{tr} \mathbf{L}_\mu^2 \le -\frac{4}{kR_1^2} \mathcal{L}(\mu)^2,$$

which moreover implies the constant bound $\left.\frac{\mathrm{d}^2}{\mathrm{d}s^2}\right|_{s=0} \mathcal{L}(\bar{\mu}_s) \le -\frac{r^{\circ 2}}{kR_1^2}$. This concludes the second claim. □

C.3 Accelerated Convergence Phase

Proposition C.7. *Let $\delta \in [0, \frac{r^{\circ 2}}{4R_1^2}]$. For any $\mu \in \mathscr{P}_2^+(\Theta)$ such that*

$$r^\circ - \sqrt{r^{\circ 2} - 4R_1^2\delta} \le 4\,\mathcal{L}(\mu) \le r^\circ + \sqrt{r^{\circ 2} - 4R_1^2\delta},$$

there exists $\mathbf{R} \in \mathcal{B}_1(k)$ such that along $\bar{\mu}_s = (1-s)\mu + s\mathbf{R}\sharp\mu^\circ$ we have $\left.\frac{\mathrm{d}}{\mathrm{d}s}\right|_{s=0} \mathcal{L}(\bar{\mu}_s) \le -\delta$.

In other words, once in the band $(0, \frac{r^\circ}{2})$ we are guaranteed a non-vanishing gradient which moreover becomes steeper closer to the center of the band, proportional to $\mathcal{L}(\mu)(r^\circ - 2\,\mathcal{L}(\mu))$.

Proof. Observe that the term $\mathbf{L}_\mu \boldsymbol{\Sigma}_{\mu^\circ,\mu} \boldsymbol{\Sigma}_{\mu,\mu}^{-1}$ lower bounding the first order decrease of $\mathcal{L}$ in the proof of Theorem 3.3 also appears in the expansion

$$\mathbf{L}_\mu^2 = \frac{1}{2} r^\circ \mathbf{L}_\mu - \frac{1}{2} \mathbf{L}_\mu \boldsymbol{\Sigma}_{\mu^\circ,\mu} \boldsymbol{\Sigma}_{\mu,\mu}^{-1} \boldsymbol{\Sigma}_{\mu,\mu^\circ}.$$

Supposing $\|\mathbf{L}_\mu \boldsymbol{\Sigma}_{\mu^\circ,\mu} \boldsymbol{\Sigma}_{\mu,\mu}^{-1}\|_* < \frac{\delta}{2}$ then allows us to construct the following inequality,

$$\begin{aligned}
\mathcal{L}(\mu)^2 = (\operatorname{tr} \mathbf{L}_\mu)^2 &\ge \operatorname{tr} \mathbf{L}_\mu^2 \\
&= \frac{r^\circ}{2} \operatorname{tr} \mathbf{L}_\mu - \frac{1}{2} \operatorname{tr}\left(\mathbf{L}_\mu \boldsymbol{\Sigma}_{\mu^\circ,\mu} \boldsymbol{\Sigma}_{\mu,\mu}^{-1} \boldsymbol{\Sigma}_{\mu,\mu^\circ}\right) \\
&\ge \frac{r^\circ}{2} \mathcal{L}(\mu) - \frac{1}{2} \|\mathbf{L}_\mu \boldsymbol{\Sigma}_{\mu^\circ,\mu} \boldsymbol{\Sigma}_{\mu,\mu}^{-1}\|_* \|\boldsymbol{\Sigma}_{\mu,\mu^\circ}\| \\
&> \frac{r^\circ}{2} \mathcal{L}(\mu) - \frac{R_1^2 \delta}{4},
\end{aligned}$$

which implies either $4\,\mathcal{L}(\mu) < r^\circ - \sqrt{r^{\circ 2} - 4R_1^2\delta}$ or $4\,\mathcal{L}(\mu) > r^\circ + \sqrt{r^{\circ 2} - 4R_1^2\delta}$. The bounds are non-vacuous only when $\delta \le \frac{r^{\circ 2}}{4R_1^2}$ and are strictly tighter for larger δ. Taking the contrapositive yields the desired statement. □

D Results and Proofs for Section 4

D.1 Recap: Finite-dimensional Dynamics

To help gain intuition, we draw parallels with the ordinary GF for a C^2 nonconvex function $f : \mathbb{R}^m \to \mathbb{R}$,

$$\mathrm{d}\boldsymbol{z}_t = -\nabla_{\boldsymbol{z}} f(\boldsymbol{z}_t)\,\mathrm{d}t.$$

A strict saddle point $\boldsymbol{z}^\dagger$ is defined as a critical point such that $\lambda_{\min}(\mathrm{Hess}_f(\boldsymbol{z}^\dagger)) < 0$, where Hess_f is the local curvature or Hessian matrix of f. Lee et al. (2019) show that the set of initial values $\boldsymbol{z}_0$ for which $\lim_{t\to\infty} \boldsymbol{z}_t$ converges to a strict saddle point has measure zero.[2] If every saddle point of f is strict and all local minima are also global minima, $\boldsymbol{z}_t$ converges to global minima for almost all initializations. The result follows easily from the center-stable manifold theorem (Shub, 2013, Theorem III.7), which states that all stable local orbits must be contained in a local embedded disk tangent to the stable eigenspace of Hess_f at $\boldsymbol{z}^\dagger$.

[2] More precisely, this is shown for iterates of discrete gradient descent, but the proof is easily adapted to the continuous-time flow.

D.2 Local Geometry of Wasserstein Space

We present some background theory on the metric geometry of Wasserstein spaces. The following result characterizes absolutely continuous curves in $\mathscr{P}_2(\Omega)$.

Theorem D.1 (Ambrosio et al. (2005), Theorem 8.3.1 and Proposition 8.4.5)**.** *Let $I \subset \mathbb{R}$ be an open interval and $\mu_t : I \to \mathscr{P}_2(\Omega)$ an absolutely continuous curve with metric derivative $|\mu'| \in L^1(I)$. Then among all Borel vector fields $\boldsymbol{v}_t \in L^2(\Omega, \mu_t)$ satisfying the continuity equation $\partial_t \mu_t + \nabla \cdot (\boldsymbol{v}_t \mu_t) = 0$, there exists an $L^1(I)$-a.e. unique minimal norm velocity field $(\boldsymbol{v}_t)$ such that*

$$\|\boldsymbol{v}_t\|_{L^2(\Omega,\mu_t)} \leq |\mu'|(t).$$

The field $(\boldsymbol{v}_t)$ is also uniquely characterized by the condition that $\boldsymbol{v}_t$ is $L^1(I)$-a.e. contained in the $L^2(\Omega, \mu_t)$-closure of the subspace $\{\nabla\psi : \psi \in C_c^\infty(\Omega)\}$. Conversely, a narrowly continuous curve given by the continuity equation for some square-integrable Borel velocity field $\boldsymbol{v}_t$ with $\|\boldsymbol{v}_t\|_{L^2(\Omega,\mu_t)} \in L^1(I)$ satisfies $|\mu'|(t) \leq \|\boldsymbol{v}_t\|_{L^2(\Omega,\mu_t)}$ a.e.

This motivates the formal definition of the tangent space to $\mathscr{P}_2(\Omega)$ at μ as

$$\operatorname{Tan}_\mu \mathscr{P}_2(\Omega) := \overline{\{\boldsymbol{v} = \nabla\psi : \psi \in C_c^\infty(\Omega)\}}^{L^2(\Omega,\mu)} \tag{6}$$

with the inherited inner product. The space can also be retrieved by the following variational principle: a vector field $\boldsymbol{v} \in L^2(\Omega, \mu)$ belongs to $\operatorname{Tan}_\mu \mathscr{P}_2(\Omega)$ if and only if $\|\boldsymbol{v} + \boldsymbol{w}\|_{L^2(\Omega,\mu)} \geq \|\boldsymbol{v}\|_{L^2(\Omega,\mu)}$ for all divergence-free fields $\boldsymbol{w} \in L^2(\Omega, \mu)$ such that $\nabla \cdot (\boldsymbol{w}\mu) = 0$. Moreover, for every $\boldsymbol{v} \in L^2(\Omega, \mu)$ there exists a unique representative $\Pi\boldsymbol{v} \in \operatorname{Tan}_\mu \mathscr{P}_2(\Omega)$ equivalent to $\boldsymbol{v}$ modulo divergence-free fields. Geometrically, this allows us to describe infinitesimal transport along curves μ_t by pushing forward along their tangent fields, analogously to the exponential map.

Proposition D.2 (Ambrosio et al. (2005), Theorem 8.3.1 and Proposition 8.4.6)**.** *Let $\mu_t : I \to \mathscr{P}_2(\Omega)$ be an absolutely continuous curve with velocity field $\boldsymbol{v}_t \in \operatorname{Tan}_{\mu_t} \mathscr{P}_2(\Omega)$ determined as in Theorem D.1. Then for a.e. $t \in I$ we have*

$$\mathcal{W}_2(\mu_{t+\epsilon}, (\mathrm{id}_\Omega + \epsilon\boldsymbol{v}_t)\sharp\mu_t) = o(\epsilon).$$

In light of Proposition D.2, the tangent space can alternatively be defined using optimal transport plans. Denote by $\Gamma_o(\mu, \nu) \subset \mathscr{P}_2(\Omega \times \Omega)$ the set of optimal transport plans from μ to ν with cost function the 2-norm and let

$$\operatorname{Tan}_\mu \mathscr{P}_2(\Omega) = \overline{\{\lambda(\boldsymbol{r} - \mathrm{id}_\Omega) : (\mathrm{id}_\Omega \times \boldsymbol{r})\sharp\mu \in \Gamma_o(\mu, \boldsymbol{r}\sharp\mu),\ \lambda > 0\}}^{L^2(\Omega,\mu)}; \tag{7}$$

this construction is equivalent to (6) (Ambrosio et al., 2005, Theorem 8.5.1).

D.3 Stability of Wasserstein Gradient Flow

We now proceed with the proofs.

Proof of Lemma 4.1. Let $\mu^\dagger$ be a critical point of F, that is $\frac{\delta F}{\delta \mu}(\mu^\dagger) = 0$. From the description (7) for the tangent space at $\mu^\dagger$, we write a local WGF (μ_t) as $\mu_t = (\mathrm{id}_\Omega + \epsilon\boldsymbol{v}_t)\sharp\mu^\dagger$ for a velocity field $\boldsymbol{v}_t \in \operatorname{Tan}_{\mu^\dagger} \mathscr{P}_2(\Omega)$. The evolution of $\boldsymbol{v}_t$ is derived as follows: for any smooth integrable function $g : \Omega \to \mathbb{R}$, the identity $\int g \,\mathrm{d}\mu_t = \int g \circ (\mathrm{id}_\Omega + \epsilon\boldsymbol{v}_t) \,\mathrm{d}\mu^\dagger$ implies that

$$\begin{aligned}\int \nabla g \cdot \nabla \frac{\delta F}{\delta \mu}(\mu_t) \,\mathrm{d}\mu_t = -\int g \,\mathrm{d}(\partial_t \mu_t) &= -\epsilon \int \nabla g \circ (\mathrm{id}_\Omega + \epsilon\boldsymbol{v}_t) \cdot \partial_t \boldsymbol{v}_t \,\mathrm{d}\mu^\dagger \\ &= -\epsilon \int \nabla g \cdot \partial_t \boldsymbol{v}_t \circ (\mathrm{id}_\Omega + \epsilon\boldsymbol{v}_t)^{-1} \,\mathrm{d}\mu_t,\end{aligned}$$

and hence $\partial_t \boldsymbol{v}_t = -\epsilon^{-1} \nabla \frac{\delta F}{\delta \mu}(\mu_t) \circ (\mathrm{id}_\Omega + \epsilon\boldsymbol{v}_t)$. On the other hand, by Proposition D.2 we can locally approximate the pushforward displacement by the absolutely continuous curve defined by $\partial_s \tilde{\mu}_s + \nabla \cdot (\boldsymbol{v}_t \mu_s) = 0$ initialized at $\tilde{\mu}_0 = \mu^\dagger$:

$$\nabla \frac{\delta F}{\delta \mu}(\tilde{\mu}_\epsilon, \theta) - \nabla \frac{\delta F}{\delta \mu}(\mu^\dagger, \theta) = \nabla \frac{\delta F}{\delta \mu}(\tilde{\mu}_s, \theta)\Big|_{s=0}^{\epsilon}$$

$$
\begin{aligned}
&= \nabla_\theta \int_0^\epsilon \int \frac{\delta^2 F}{\delta \mu^2}(\tilde{\mu}_s, \theta, \theta') \partial_s \tilde{\mu}_s(\mathrm{d}\theta')\, \mathrm{d}s \\
&= \nabla_\theta \int_0^\epsilon \int \nabla_{\theta'} \frac{\delta^2 F}{\delta \mu^2}(\tilde{\mu}_s, \theta, \theta') \boldsymbol{v}_t(\theta') \tilde{\mu}_s(\mathrm{d}\theta')\, \mathrm{d}s \\
&= \int_0^\epsilon \int \mathbf{H}_{\mu^\dagger}(\theta, \theta') \boldsymbol{v}_t(\theta') \mu^\dagger(\mathrm{d}\theta') + O(\mathcal{W}_2(\tilde{\mu}_s, \mu^\dagger))\, \mathrm{d}s \\
&= \epsilon\, \mathcal{H}_{\mu^\dagger}\, \boldsymbol{v}_t + o(\epsilon)
\end{aligned}
$$

so that

$$
\begin{aligned}
\partial_t \boldsymbol{v}_t = -\frac{1}{\epsilon} \Bigg(& \underbrace{\nabla \frac{\delta F}{\delta \mu}(\mu_t) \circ (\mathrm{id}_\Omega + \epsilon \boldsymbol{v}_t) - \nabla \frac{\delta F}{\delta \mu}(\mu_t)}_{=o(\epsilon)} + \underbrace{\nabla \frac{\delta F}{\delta \mu}(\mu_t) - \nabla \frac{\delta F}{\delta \mu}(\tilde{\mu}_\epsilon)}_{=o(\epsilon)} \\
& + \nabla \frac{\delta F}{\delta \mu}(\tilde{\mu}_\epsilon) - \nabla \frac{\delta F}{\delta \mu}(\mu^\dagger) + \underbrace{\nabla \frac{\delta F}{\delta \mu}(\mu^\dagger)}_{=0} \Bigg) \\
= & - \mathcal{H}_{\mu^\dagger}\, \boldsymbol{v}_t + o(1).
\end{aligned}
$$

Here, we see that the $o(1)$ perturbation term is more precisely of order $O(\mathcal{W}_2(\mu_t, \mu^\dagger))$ and vanishes when the L^2-norm of the velocity field $\boldsymbol{v}_t$ goes to zero. □

Proof of Lemma 4.2. It will suffice to show $\mathbf{H}_\mu$ is symmetric in the sense that $\mathbf{H}_\mu(\theta, \theta')^\top = \mathbf{H}_\mu(\theta', \theta)$ for all $\theta, \theta' \in \Omega$. We appeal directly to Definition A.1: for any μ, ν_1, ν_2,

$$
\begin{aligned}
&\frac{\mathrm{d}^2}{\mathrm{d}\epsilon_1\, \mathrm{d}\epsilon_2}\bigg|_{\epsilon_1=\epsilon_2=0} F(\mu + \epsilon_1(\nu_1 - \mu) + \epsilon_2(\nu_2 - \mu)) \\
&= \frac{\mathrm{d}}{\mathrm{d}\epsilon_2}\bigg|_{\epsilon_2=0} \int \frac{\delta F}{\delta \mu}(\mu + \epsilon_2(\nu_2 - \mu), \theta)(\nu_1 - \mu)(\mathrm{d}\theta) \\
&= \iint \frac{\delta^2 F}{\delta \mu^2}(\mu, \theta, \theta')(\nu_1 - \mu)(\mathrm{d}\theta)(\nu_2 - \mu)(\mathrm{d}\theta'),
\end{aligned}
$$

and comparing with the same computation with the indices swapped yields that $\frac{\delta^2 F}{\delta \mu^2}$ is symmetric in θ, θ'. Therefore the Hessian matrix satisfies $\nabla_\theta \nabla_{\theta'} \frac{\delta^2 F}{\delta \mu^2}(\mu, \theta, \theta') = \nabla_{\theta'} \nabla_\theta \frac{\delta^2 F}{\delta \mu^2}(\mu, \theta', \theta)^\top$. Then for any functions $f, g \in L^2(\Omega, \mu; \mathbb{R}^{k+d})$ it holds that

$$
\begin{aligned}
\langle f, \mathcal{H}_\mu g \rangle_{L^2(\Omega, \mu; \mathbb{R}^{k+d})} &= \iint f(\theta)^\top \mathbf{H}_\mu(\theta, \theta') g(\theta') \mu(\mathrm{d}\theta) \mu(\mathrm{d}\theta') \\
&= \iint g(\theta)^\top \mathbf{H}_\mu(\theta, \theta') f(\theta') \mu(\mathrm{d}\theta) \mu(\mathrm{d}\theta') \\
&= \langle \mathcal{H}_\mu f, g \rangle_{L^2(\Omega, \mu; \mathbb{R}^{k+d})},
\end{aligned}
$$

thus $\mathcal{H}_\mu$ is self-adjoint. Since the kernel is Hilbert-Schmidt by assumption, $\mathcal{H}_\mu$ is also compact, and we can invoke the spectral theorem to conclude the statement. □

Theorem D.3 (Gallay (1993), Theorem 1.1)**.** *Let $\mathscr{E}$ be a Banach space, $\mathbf{A}$ a linear operator on $\mathscr{E}$, and $f : \mathscr{E} \to \mathscr{E}$ a C^k perturbation with $f(0) = 0$, $Df(0) = 0$, where $k > 1$. Consider the differential equation*

$$
\frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{z}_t = \mathbf{A} \boldsymbol{z}_t + f(\boldsymbol{z}_t), \quad t \geq 0. \tag{8}
$$

Assume that $\mathscr{E}$ is the direct sum of two closed, $\mathbf{A}$-invariant subspaces $\mathscr{E}^s, \mathscr{E}^u$. The corresponding restrictions $\mathbf{A}^s = \mathbf{A}|_{\mathscr{E}^s}$, $\mathbf{A}^u = \mathbf{A}|_{\mathscr{E}^u}$ generate strongly continuous semigroups $e^{\mathbf{A}^s t}$, $e^{-\mathbf{A}^u t}$ for $t \geq 0$ which moreover satisfy for real numbers $0 \leq \lambda^s < \lambda^u$,

$$
\sup_{t \geq 0} \| e^{\mathbf{A}^s t} \| e^{-\lambda^s t} < \infty, \quad \sup_{t \geq 0} \| e^{-\mathbf{A}^u t} \| e^{\lambda^u t} < \infty.
$$

Further assume there exists a spectral gap of $\lambda^u > k\lambda^s$ and that $\mathscr{E}^s$ has the C^k extension property. Let $\mathcal{B}_r, \mathcal{B}_r^s, \mathcal{B}_r^u$ denote the balls of radius r around the origin in $\mathscr{E}, \mathscr{E}^s, \mathscr{E}^u$, respectively. Then for sufficiently small $r > 0$, there exists a C^k map $h : \mathcal{B}_r^s \to \mathcal{B}_r^u$ with $h(0) = 0$, $Dh(0) = 0$ whose graph $\mathscr{V} \subset \mathcal{B}_r$ (the local center-stable manifold) has the following properties.

(i) *(Invariance) For all initial values $\boldsymbol{z}_0 \in \mathscr{V}$ there exists a C^1 curve $\boldsymbol{z}_t : \mathbb{R}_{\geq 0} \to \mathscr{E}$ such that as long as $\boldsymbol{z}_t \in \mathcal{B}_r$, then $\boldsymbol{z}_t \in \mathscr{V}$ and (8) holds.*

(ii) *(Uniqueness) If $\boldsymbol{z}_t$ is any solution of (8) such that $\boldsymbol{z}_t \in \mathcal{B}_r$ for all $t \geq 0$, then $\boldsymbol{z}_t \in \mathscr{V}$ for all $t \geq 0$.*

Proof of Theorem 4.3. Let $\mu^\dagger \in \mathscr{G}^\dagger$ be a strict saddle point. We apply the local center-stable manifold theorem to the system of Lemma 4.1 on $L^2(\Omega, \mu^\dagger; \mathbb{R}^{k+d})$. By the spectral theorem, the operator $\mathcal{H}_{\mu^\dagger}$ has a complete set of eigenvalues λ_j and corresponding eigenfunctions ψ_j for $j \in \mathbb{Z}$, ordered such that

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq 0, \quad \lambda_0 = \cdots = \lambda_{-(p-1)} < \lambda_{-p} \leq \cdots \leq 0.$$

Since the spectrum may possess a limit point at 0, we cannot separate $\mathcal{H}_{\mu^\dagger}$ into absolutely convergent and divergent components. Instead, we set the cutoff at the largest negative eigenvalue $\lambda_0 = \lambda_{\min}(\mathcal{H}_{\mu^\dagger})$, taking all possibly multiple eigenvalues, and defining the subspace $\mathscr{E}^u$ as the span of the corresponding $\psi_0, \cdots, \psi_{-(p-1)}$. Then we are guaranteed a jump $\lambda_{-(p-1)} < \lambda_{-p}$ since the spectrum is discrete, and we choose $\lambda_s = -\lambda_{-p}, \lambda^u = -\lambda_0$ and $k \in (1, |\lambda_0/\lambda_{-p}|)$ so that the spectral gap condition is satisfied – we only need continuity (i.e. $k \geq 0$) for our argument. Moreover, the C^k extension property for $\mathscr{E}^s$ holds automatically as $L^2(\Omega, \mu^\dagger; \mathbb{R}^{k+d})$ is a Hilbert space. Therefore, any convergent local flow $(\boldsymbol{v}_t)$ defined in an open neighborhood $\mathcal{B}_{\mu^\dagger}$ must be contained in a graph $\mathscr{V}_{\mu^\dagger} \subset \mathcal{B}_{\mu^\dagger}$ containing $\mu^\dagger$.

The rest of the proof is similar to Lee et al. (2019). Since the collection $\{\mathcal{B}_{\mu^\dagger} : \mu^\dagger \in \mathscr{G}^\dagger\}$ forms an open cover of $\mathscr{G}^\dagger$ and $\mathscr{P}_2(\Omega)$ is separable with respect to 2-Wasserstein distance (Ambrosio et al., 2005, Proposition 7.1.5), we can extract a countable subcover $\{\mathcal{B}_j : j \in \mathbb{N}\}$ containing $\mathscr{G}^\dagger$. If the WGF $(\mu_t)_{t \geq 0}$ converges to a strict saddle point, there exists an index j and an integer threshold ℓ such that $\mu_t \in \mathcal{B}_j$ for $t \geq \ell$. In particular, μ_t must be contained in the corresponding center-stable manifold $\mathscr{V}_j$ for $t \geq \ell$.

Let $\omega_t^-(\nu)$ denote the result of running the reversed gradient flow $\partial \nu_{-t} = -\nabla \cdot (\nu_{-t} \nabla \frac{\delta F}{\delta \mu}(\nu_{-t}))$, $\nu_0 = \nu$ for time t whenever it exists; time inversion $t \mapsto -t$ shows that $\omega_t^-(\mu_t) = \mu_0$ for the forward flow $(\mu_t)_{t \geq 0}$. Since $\mu_\ell \in \mathscr{V}_j$ for some integer time ℓ and $\mathscr{V}_j$, it holds that

$$\mathscr{G}_0^\dagger \subseteq \bigcup_{j \in \mathbb{N}} \bigcup_{\ell \in \mathbb{N}} \omega_\ell^-(\mathscr{V}_j),$$

hence $\mathscr{G}_0^\dagger$ must be contained in the countable union of images of graphs. □

Remark D.4. We point out that Otto calculus is only formal in the sense that existence and regularity issues are ignored, so it is difficult to rigorously turn the above into a meaningful measure-theoretic statement as in Lee et al. (2019). This is compounded by the fact that there is no well-behaved canonical measure on $\mathscr{P}_2(\Omega)$. A possible justification is to restrict to the subspace of measures with smooth positive Lebesgue density whose geometry is well-behaved (Lott, 2008; Villani, 2009), but this is outside of the scope of our paper.

D.4 Application to three-layer networks

The problem (2) can also be motivated by the training dynamics of a three-layer neural network. We construct the first two layers identically to our MLP layer h_μ and consider a linear third layer given by the transformation $\mathbf{T} \in \mathbb{R}^{k \times k}$. Then the L^2 loss with respect to a teacher network $\boldsymbol{x} \mapsto \mathbf{T}^\star h_{\mu^\star}(\boldsymbol{x})$ is

$$\mathcal{L}_{\text{NN}}(\mu, \mathbf{T}) = \mathbb{E}_{\boldsymbol{x}}[\|\mathbf{T}^\star h_{\mu^\star}(\boldsymbol{x}) - \mathbf{T} h_\mu(\boldsymbol{x})\|^2].$$

By setting $\mu^\circ = \mathbf{T}^\star \sharp \mu^\star$ and taking the two-timescale limit where the last layer updates infinitely quickly, we see that $\mathbf{T}$ must converge to $\boldsymbol{\Sigma}_{\mu^\circ, \mu} \boldsymbol{\Sigma}_{\mu,\mu}^{-1}$ and we end up with the regression objective (2), hence Sections 3-5 also directly apply to this problem. We remark that the two-timescale regime has been leveraged to show convergence of SGD for *two*-layer networks in Marion & Berthier (2023).

E Results and Proofs for Section 5

E.1 First-order Improvement

Proposition E.1. *Let F be a functional depending on μ only through the MLP layer h_μ. Suppose MFD (3) at time t admits a distribution $\bar{\mu} \in \mathscr{P}_2(\Theta)$ with $\chi^2(\bar{\mu}, \mu_t) \leq \bar{\chi}^2$ such that along the linear homotopy $\bar{\mu}_s = (1-s)\mu_t + s\bar{\mu}$ we have $\frac{\mathrm{d}}{\mathrm{d}s}\big|_{s=0} F(\bar{\mu}_s) \leq -\delta \leq 0$. Then $\frac{\mathrm{d}}{\mathrm{d}t} F(\mu_t) \leq -\bar{\chi}^{-2}\delta^2$.*

Proof. We may express F as $F(\mu) = J(h_\mu)$ for an auxiliary functional $h \mapsto J(h)$ defined on $C(\mathscr{X}, \mathbb{R}^d)$, which implies that

$$\frac{\delta F}{\delta \mu}(\mu, \theta) = \int \frac{\delta J}{\delta h}(h_\mu, \boldsymbol{x})^\top h_\theta(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}.$$

In particular, since the dependency on the second layer $\boldsymbol{a}$ is linear, it always holds that $\boldsymbol{a}^\top \nabla_{\boldsymbol{a}} \frac{\delta F}{\delta \mu} = \frac{\delta F}{\delta \mu}$. We can then directly lower bound the decrease rate of the objective under (3) by isolating the gradient provided by $\boldsymbol{a}$:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} F(\mu_t) &= \int \frac{\delta F}{\delta \mu}(\mu_t, \theta) \partial_t \mu_t(\mathrm{d}\theta) \\
&= -\int \left\| \nabla_\theta \frac{\delta F}{\delta \mu}(\mu_t, \theta) \right\|^2 \mu_t(\mathrm{d}\theta) \\
&\leq -\int \left\| \nabla_{\boldsymbol{a}} \frac{\delta F}{\delta \mu}(\mu_t, \theta) \right\|^2 \mu_t(\mathrm{d}\theta) \\
&\leq -\int \left(\boldsymbol{a}^\top \nabla_{\boldsymbol{a}} \frac{\delta F}{\delta \mu}(\mu_t, \theta) \right)^2 \mu_t(\mathrm{d}\theta) \\
&= -\int \left(\frac{\delta F}{\delta \mu}(\mu_t, \theta) \right)^2 \mu_t(\mathrm{d}\theta).
\end{aligned}$$

Starting from the first-order condition, by the Cauchy-Schwarz inequality we can also bound

$$\left(\frac{\mathrm{d}}{\mathrm{d}s}\bigg|_{s=0} F(\bar{\mu}_s) \right)^2 = \left(\int \frac{\delta F}{\delta \mu}(\mu_t, \theta)(\bar{\mu} - \mu_t)(\mathrm{d}\theta) \right)^2 \leq \chi^2(\bar{\mu}, \mu_t) \int \left(\frac{\delta F}{\delta \mu}(\mu_t, \theta) \right)^2 \mu_t(\mathrm{d}\theta).$$

Joining the two inequalities gives the desired bound. □

Proof of Proposition 5.1. Recall that the functional derivative is computed as

$$\frac{\delta \mathcal{L}}{\delta \mu}(\mu, \theta) = -\mathbb{E}_{\boldsymbol{x}} \left[\zeta_{\mu^\circ, \mu}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_{\mu^\circ, \mu} \boldsymbol{\Sigma}_{\mu, \mu}^{-1} h_\theta(\boldsymbol{x}) \right], \tag{9}$$

where the additive constant has been implicitly normalized such that the integral with respect to the current measure μ is zero, i.e. $\int \frac{\delta \mathcal{L}}{\delta \mu}(\mu) \, \mathrm{d}\mu = 0$ as shown in the proof of Theorem 3.3. Due to the spherical symmetry of π in the first component, it is also immediate that

$$\int \frac{\delta \mathcal{L}}{\delta \mu}(\mu) \, \mathrm{d}\pi = -\mathbb{E}_{\boldsymbol{x}} \left[\zeta_{\mu^\circ, \mu}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_{\mu^\circ, \mu} \boldsymbol{\Sigma}_{\mu, \mu}^{-1} h_\pi(\boldsymbol{x}) \right] = 0.$$

The chi-square divergence between $\bar{\mu} = \mathbf{R}\sharp\mu^\circ$ and μ_t can be bounded as

$$\int \left(\frac{\mathrm{d}\bar{\mu}}{\mathrm{d}\mu_t} - 1 \right)^2 \mathrm{d}\mu_t \leq \left\| \frac{\mathrm{d}\bar{\mu}}{\mathrm{d}\mu_t} \right\|_\infty - 1 \leq \gamma^{-1} \left\| \frac{\mathrm{d}\bar{\mu}}{\mathrm{d}\pi} \right\|_\infty$$

where the birth-death mechanism prevents the density ratio $\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}$ from falling below the threshold γ at any point. Writing the convex decomposition of $\mathbf{R}$ in the sense of Lemma C.6 as $\sum_{j=1}^m \alpha_j \mathbf{R}_j$ with $\mathbf{R}_j \in \mathcal{O}(k)$, the density of $\bar{\mu}$ relative to π is further bounded as

$$\left\| \frac{\mathrm{d}\bar{\mu}}{\mathrm{d}\pi} \right\|_\infty = \left\| \frac{\mathrm{d}\mathbf{R}\sharp\mu^\circ}{\mathrm{d}\pi} \right\|_\infty \leq \sum_{j=1}^m \alpha_j \left\| \frac{\mathrm{d}\mu^\circ}{\mathrm{d}\pi} \right\|_\infty \leq R_4$$

by the spherical symmetry of π. Hence we may apply Proposition E.1 with $\bar{\chi}^2 = \gamma^{-1} R_4$, showing that the objective decreases along MFD by a rate of at least $\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}(\mu_t) \le -R_4^{-1}\gamma\delta^2$.

Moreover, whenever the discrete linear update is performed, along the homotopy $\hat{\mu}_s := (1 - s\gamma)\mu_t + s\gamma\pi$ we have

$$\frac{\mathrm{d}}{\mathrm{d}s}\mathcal{L}(\hat{\mu}_s) = \gamma \int \frac{\delta \mathcal{L}}{\delta \mu}(\hat{\mu}_s, \theta)(\pi - \mu_t)(\mathrm{d}\theta) = 0.$$

Hence $t \mapsto \mathcal{L}(\mu_t)$ is unaffected by the discrete updates, justifying the inequality for all time $t \ge 0$. □

As we mentioned briefly, the proof can also be easily modified to handle unbounded second layer $\boldsymbol{a}$ by invoking the Cauchy-Schwarz inequality to lower bound the gradient

$$\int \left\| \nabla_{\boldsymbol{a}} \frac{\delta \mathcal{L}}{\delta \mu}(\mu_t, \theta) \right\|^2 \mu_t(\mathrm{d}\theta) \ge \left(\int \|\boldsymbol{a}\|^2 \mu_t(\mathrm{d}\theta) \right)^{-1} \int \left(\boldsymbol{a}^\top \nabla_{\boldsymbol{a}} \frac{\delta \mathcal{L}}{\delta \mu}(\mu_t, \theta) \right)^2 \mu_t(\mathrm{d}\theta)$$

and bounding the second moment uniformly in time with the following result,

Lemma E.2. *Denote the second moment of $\mu \in \mathscr{P}_2(\Theta)$ along the $\boldsymbol{a}$ component as $m_{\boldsymbol{a}}(\mu) = \int \|\boldsymbol{a}\|^2 \mu(\mathrm{d}\theta)$. Then the mean-field dynamics μ_t for all time $t \ge 0$ satisfies $m_{\boldsymbol{a}}(\mu_t) \le m_{\boldsymbol{a}}(\mu_0) \vee m_{\boldsymbol{a}}(\pi)$.*

Proof. In fact, $m_{\boldsymbol{a}}(\cdot)$ remains unchanged by gradient flow:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} m_{\boldsymbol{a}}(\mu_t) &= \int \|\boldsymbol{a}\|^2 \partial_t \mu_t(\mathrm{d}\theta) \\
&= -2 \int (\boldsymbol{a} \;\; 0_d)^\top \nabla_\theta \frac{\delta \mathcal{L}}{\delta \mu}(\mu_t, \theta) \mu_t(\mathrm{d}\theta) \\
&= -2 \int \boldsymbol{a}^\top \nabla_{\boldsymbol{a}} \frac{\delta \mathcal{L}}{\delta \mu}(\mu_t, \theta) \mu_t(\mathrm{d}\theta) \\
&= -2 \int \frac{\delta \mathcal{L}}{\delta \mu}(\mu_t, \theta) \mu_t(\mathrm{d}\theta) = 0.
\end{aligned}$$

Also if the discrete update is performed, the output satisfies $m_{\boldsymbol{a}}((1-\gamma)\mu_t + \gamma\pi) = (1-\gamma) m_{\boldsymbol{a}}(\mu_t) + \gamma m_{\boldsymbol{a}}(\pi)$ by linearity of the moment functional $\mu \mapsto m_{\boldsymbol{a}}(\mu)$. Hence $m_{\boldsymbol{a}}(\mu_t)$ always interpolates between $m_{\boldsymbol{a}}(\mu_0)$ and $m_{\boldsymbol{a}}(\pi)$. □

Proof of Theorem 5.2. Suppose $\mathcal{L}(\mu_0) \le 0.49 r^\circ$. Then $\mathcal{L}(\mu_t) \le 0.49 r^\circ$ for all $t \ge 0$ and by Proposition C.7 we are guaranteed a direction of improvement $\bar{\mu}_s = (1-s)\mu_t + s\bar{\mu}$ with $\bar{\mu} = \mathbf{R}\sharp\mu$ for some $\mathbf{R} \in \mathcal{B}_1(k)$ such that

$$\left.\frac{\mathrm{d}}{\mathrm{d}s}\right|_{s=0} \mathcal{L}(\bar{\mu}_s) \le -\frac{4}{R_1^2}\mathcal{L}(\mu_t)\left(\frac{r^\circ}{2} - \mathcal{L}(\mu_t)\right).$$

Proposition 5.1 then ensures the objective decreases along the Wasserstein flow as

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}(\mu_t) \le -\frac{16\gamma}{R_1^4 R_4}\mathcal{L}(\mu_t)^2\left(\frac{r^\circ}{2} - \mathcal{L}(\mu_t)\right)^2, \quad 0 \le \mathcal{L}(\mu_t) \le \frac{r^\circ}{2}.$$

We now divide the band into two halves.

(i) $\frac{r^\circ}{4} \le \mathcal{L} \le \frac{r^\circ}{2}$ (acceleration band). By substituting $\mathcal{L}(\mu_t)^2 \ge \frac{r^{\circ 2}}{16}$ above and solving the differential inequality, we obtain

$$\mathcal{L}(\mu_t) \le \frac{r^\circ}{2} - \left(\frac{100}{r^\circ} - \frac{r^{\circ 2}\gamma t}{R_1^4 R_4}\right)^{-1}$$

and hence $\mathcal{L}(\mu_t)$ decreases below $\frac{r^\circ}{4}$ after time $t_1 \le \frac{96 R_1^4 R_4}{r^{\circ 3}\gamma}$.

(ii) $0 \le \mathcal{L} \le \frac{r^\circ}{4}$ (deceleration band). By substituting $(\frac{r^\circ}{2} - \mathcal{L}(\mu_t))^2 \ge \frac{r^{\circ 2}}{16}$ we likewise obtain

$$\mathcal{L}(\mu_t) \le \left(\frac{4}{r^\circ} + \frac{r^{\circ 2}\gamma(t - t_1)}{R_1^4 R_4}\right)^{-1}$$

and hence $\mathcal{L}(\mu_t)$ achieves loss $\le \epsilon$ after time $t_1 + \frac{R_1^4 R_4}{r^{\circ 2}\gamma} \cdot \frac{1}{\epsilon}$.

Finally, note that the second term dominates the first since $\epsilon = O(r^\circ)$. □

E.2 Second-order Improvement

Proof of Lemma 5.3. It is straightforward to show that

$$\begin{aligned}\partial_t\left[\nabla_\theta \frac{\delta F}{\delta \mu}(\mu_t,\theta)\right] &= \nabla_\theta \int \frac{\delta^2 F}{\delta \mu^2}(\mu_t,\theta,\theta')(\partial_t \mu_t)(\mathrm{d}\theta') \\ &= -\nabla_\theta \int \nabla_{\theta'}\frac{\delta^2 F}{\delta \mu^2}(\mu_t,\theta,\theta')\cdot \nabla_{\theta'}\frac{\delta F}{\delta \mu}(\mu_t,\theta')\mu_t(\mathrm{d}\theta') \\ &= -\int \mathbf{H}_{\mu_t}(\theta,\theta')\nabla_{\theta'}\frac{\delta F}{\delta \mu}(\mu_t,\theta')\mu_t(\mathrm{d}\theta').\end{aligned}$$

Each term is well-defined as soon as the kernel is assumed to be Hilbert-Schmidt, or due to Lemma E.3 for the case $F=\mathcal{L}$. □

Lemma E.3. *The kernel $\mathbf{H}_\mu$ for the functional $\mathcal{L}$ is Hilbert-Schmidt for all $\mu \in \mathscr{P}_2^+(\Theta)$. Moreover, the corresponding integral operator $\mathcal{H}_\mu f(\theta) = \int \mathbf{H}_\mu(\theta,\theta')f(\theta')\mu(\mathrm{d}\theta')$ is compact self-adjoint, hence there exists an orthonormal basis $\{\psi_j\}_{j\in\mathbb{Z}}$ for $L^2(\Theta,\mu;\mathbb{R}^{k+d})$ consisting of eigenfunctions of $\mathcal{H}_\mu$.*

Proof. We extend our notation to write for example $\mathbf{\Sigma}_{\mu,\theta} = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\mathscr{X}}}[h_\mu(\boldsymbol{x})h_\theta(\boldsymbol{x})^\top]$. From (9) the second order functional derivative can be derived as

$$\begin{aligned}\frac{\delta^2\mathcal{L}}{\delta\mu^2}(\mu,\theta,\theta') &= -\frac{\delta}{\delta\mu}\mathbb{E}_{\boldsymbol{x}}\left[(h_{\mu^\circ}(\boldsymbol{x}) - \mathbf{\Sigma}_{\mu^\circ,\mu}\mathbf{\Sigma}_{\mu,\mu}^{-1}h_\mu(\boldsymbol{x}))^\top \mathbf{\Sigma}_{\mu^\circ,\mu}\mathbf{\Sigma}_{\mu,\mu}^{-1}h_\theta(\boldsymbol{x})\right](\theta') \\ &= -\operatorname{tr}\left(\mathbf{\Sigma}_{\mu^\circ,\theta'}\mathbf{\Sigma}_{\mu,\mu}^{-1}\mathbf{\Sigma}_{\theta,\mu^\circ}\right) \\ &\quad + \operatorname{tr}\left(\mathbf{\Sigma}_{\mu^\circ,\mu}\mathbf{\Sigma}_{\mu,\mu}^{-1}(\mathbf{\Sigma}_{\theta',\mu}+\mathbf{\Sigma}_{\mu,\theta'})\mathbf{\Sigma}_{\mu,\mu}^{-1}\mathbf{\Sigma}_{\theta,\mu^\circ}\right) \\ &\quad + \operatorname{tr}\left(\mathbf{\Sigma}_{\mu,\mu}^{-1}\mathbf{\Sigma}_{\mu,\mu^\circ}\mathbf{\Sigma}_{\mu^\circ,\mu}\mathbf{\Sigma}_{\mu,\mu}^{-1}\mathbf{\Sigma}_{\theta,\theta'}\right) \\ &\quad - \operatorname{tr}\left(\mathbf{\Sigma}_{\mu,\mu}^{-1}(\mathbf{\Sigma}_{\theta',\mu}+\mathbf{\Sigma}_{\mu,\theta'})\mathbf{\Sigma}_{\mu,\mu}^{-1}\mathbf{\Sigma}_{\mu,\mu^\circ}\mathbf{\Sigma}_{\mu^\circ,\mu}\mathbf{\Sigma}_{\mu,\mu}^{-1}\mathbf{\Sigma}_{\theta,\mu}\right) \\ &\quad + \operatorname{tr}\left(\mathbf{\Sigma}_{\mu,\mu}^{-1}\mathbf{\Sigma}_{\theta',\mu^\circ}\mathbf{\Sigma}_{\mu^\circ,\mu}\mathbf{\Sigma}_{\mu,\mu}^{-1}\mathbf{\Sigma}_{\theta,\mu}\right) \\ &\quad + \operatorname{tr}\left(\mathbf{\Sigma}_{\mu,\mu}^{-1}\mathbf{\Sigma}_{\mu,\mu^\circ}\mathbf{\Sigma}_{\mu^\circ,\theta'}\mathbf{\Sigma}_{\mu,\mu}^{-1}\mathbf{\Sigma}_{\theta,\mu}\right) \\ &\quad - \operatorname{tr}\left(\mathbf{\Sigma}_{\mu,\mu}^{-1}\mathbf{\Sigma}_{\mu,\mu^\circ}\mathbf{\Sigma}_{\mu^\circ,\mu}\mathbf{\Sigma}_{\mu,\mu}^{-1}(\mathbf{\Sigma}_{\theta',\mu}+\mathbf{\Sigma}_{\mu,\theta'})\mathbf{\Sigma}_{\mu,\mu}^{-1}\mathbf{\Sigma}_{\theta,\mu}\right).\end{aligned}$$

It is tedious but straightforward to check that this expression is symmetric in θ,θ' (which would otherwise follow directly if we had a priori second order regularity estimates for $\mathcal{L}$). We then have that

$$[\mathbf{H}_\mu(\theta,\theta')]_{i,j} = \partial_{\theta_i}\partial_{\theta'_j}\frac{\delta^2\mathcal{L}}{\delta\mu^2}(\mu,\theta,\theta') = \partial_{\theta'_j}\partial_{\theta_i}\frac{\delta^2\mathcal{L}}{\delta\mu^2}(\mu,\theta',\theta) = [\mathbf{H}_\mu(\theta',\theta)]_{j,i}$$

which implies $\mathcal{H}_\mu$ is self-adjoint as before. For the proof of the first claim, we refer to the uniform spectral bound for $\mathbf{H}_\mu$ obtained in Lemma E.5; this also shows that $\mathcal{H}_\mu$ is compact. □

In Lemma E.4 and E.5, we derive various regularity bounds of the ICFL objective $\mathcal{L}$. The constants $C_1,\cdots,C_5$, numbered as to be consistent with Theorem E.7, are explicitly defined during the proofs and have at most polynomial dependency on all problem constants.

Lemma E.4. *The gradients of the functional derivative of $\mathcal{L}$ at any $\mu \in \mathscr{P}_2^+(\Theta)$ such that $\lambda_{\min}(\mathbf{\Sigma}_{\mu,\mu}) \geq \lambda$ uniformly satisfy $\|\nabla_{\boldsymbol{a}}\frac{\delta\mathcal{L}}{\delta\mu}\| \leq C_{\boldsymbol{a}}$, $\|\nabla_{\boldsymbol{w}}\frac{\delta\mathcal{L}}{\delta\mu}\| \leq C_{\boldsymbol{w}}$ and $\|\nabla\frac{\delta\mathcal{L}}{\delta\mu}\| \leq C_1$. Moreover, $\nabla\frac{\delta\mathcal{L}}{\delta\mu}$ is C_2-Lipschitz on Θ, where $C_{\boldsymbol{a}}, C_2 = O(\frac{1}{(k\lambda)^{1/2}})$ and $C_{\boldsymbol{w}}, C_1 = O(\frac{1}{k\lambda})$.*

Proof. The gradient with respect to each component is given by

$$\begin{aligned}\nabla_{\boldsymbol{a}}\frac{\delta\mathcal{L}}{\delta\mu}(\mu,\theta) &= -\mathbb{E}_{\boldsymbol{x}}\left[\zeta_{\mu^\circ,\mu}(\boldsymbol{x})^\top \mathbf{\Sigma}_{\mu^\circ,\mu}\mathbf{\Sigma}_{\mu,\mu}^{-1}\sigma(\boldsymbol{w}^\top\boldsymbol{x})\right]^\top, \\ \nabla_{\boldsymbol{w}}\frac{\delta\mathcal{L}}{\delta\mu}(\mu,\theta) &= -\mathbb{E}_{\boldsymbol{x}}\left[\zeta_{\mu^\circ,\mu}(\boldsymbol{x})^\top \mathbf{\Sigma}_{\mu^\circ,\mu}\mathbf{\Sigma}_{\mu,\mu}^{-1}\boldsymbol{a}\sigma'(\boldsymbol{w}^\top\boldsymbol{x})\boldsymbol{x}\right].\end{aligned}$$

Hence we can bound

$$\begin{aligned}\left\|\nabla_{\boldsymbol{a}}\frac{\delta\mathcal{L}}{\delta\mu}(\mu,\theta)\right\|^2 &\le \mathbb{E}_{\boldsymbol{x}}\left[\|\zeta_{\mu^\circ,\mu}(\boldsymbol{x})^\top\boldsymbol{\Sigma}_{\mu^\circ,\mu}\boldsymbol{\Sigma}_{\mu,\mu}^{-1}\sigma(\boldsymbol{w}^\top\boldsymbol{x})\|^2\right]\\ &\le R_1^2\cdot\mathbb{E}_{\boldsymbol{x}}\left[\zeta_{\mu^\circ,\mu}(\boldsymbol{x})^\top\boldsymbol{\Sigma}_{\mu^\circ,\mu}\boldsymbol{\Sigma}_{\mu,\mu}^{-2}\boldsymbol{\Sigma}_{\mu,\mu^\circ}\zeta_{\mu^\circ,\mu}(\boldsymbol{x})\right]\\ &\le \frac{R_1^2}{\lambda}\operatorname{tr}\left(\mathbf{L}_\mu\boldsymbol{\Sigma}_{\mu^\circ,\mu}\boldsymbol{\Sigma}_{\mu,\mu}^{-1}\boldsymbol{\Sigma}_{\mu,\mu^\circ}\right)\\ &\le \frac{r^\circ R_1^2}{\lambda}\mathcal{L}(\mu)-\frac{2R_1^2}{\lambda}\operatorname{tr}\mathbf{L}_\mu^2\\ &\le \frac{kr^{\circ 2}R_1^2}{2\lambda}=:C_{\boldsymbol{a}}^2,\end{aligned}$$

and also

$$\begin{aligned}&\left\|\nabla_{\boldsymbol{w}}\frac{\delta\mathcal{L}}{\delta\mu}(\mu,\theta)\right\|^2\\ &\le \mathbb{E}_{\boldsymbol{x}}\left[\|\zeta_{\mu^\circ,\mu}(\boldsymbol{x})^\top\boldsymbol{\Sigma}_{\mu^\circ,\mu}\boldsymbol{\Sigma}_{\mu,\mu}^{-1}\boldsymbol{a}\sigma'(\boldsymbol{w}^\top\boldsymbol{x})\boldsymbol{x}\|^2\right]\\ &\le R_2^2\cdot\mathbb{E}_{\boldsymbol{x}}\left[\|\zeta_{\mu^\circ,\mu}(\boldsymbol{x})^\top\boldsymbol{\Sigma}_{\mu^\circ,\mu}\boldsymbol{\Sigma}_{\mu,\mu}^{-1}\|^2\|\boldsymbol{x}\|^2\right]\\ &\le \frac{R_2^2}{\lambda}\cdot\mathbb{E}_{\boldsymbol{x}}\left[\|\zeta_{\mu^\circ,\mu}(\boldsymbol{x})^\top\boldsymbol{\Sigma}_{\mu^\circ,\mu}\boldsymbol{\Sigma}_{\mu,\mu}^{-1/2}\|^4\right]^{1/2}\mathbb{E}_{\boldsymbol{x}}[\|\boldsymbol{x}\|^4]^{1/2}\\ &\le \frac{R_2^2M_4^{1/2}}{\lambda}\left(\operatorname{tr}\mathbb{E}_{\boldsymbol{x}}\left[\left(\boldsymbol{\Sigma}_{\mu^\circ,\mu}\boldsymbol{\Sigma}_{\mu,\mu}^{-1}\boldsymbol{\Sigma}_{\mu,\mu^\circ}\zeta_{\mu^\circ,\mu}(\boldsymbol{x})\zeta_{\mu^\circ,\mu}(\boldsymbol{x})^\top\right)^2\right]\right)^{1/2}\\ &\le \frac{r^\circ R_2^2M_4^{1/2}}{\lambda}\left(\operatorname{tr}\mathbb{E}_{\boldsymbol{x}}\left[\zeta_{\mu^\circ,\mu}(\boldsymbol{x})\zeta_{\mu^\circ,\mu}(\boldsymbol{x})^\top\boldsymbol{\Sigma}_{\mu^\circ,\mu}\boldsymbol{\Sigma}_{\mu,\mu}^{-1}\boldsymbol{\Sigma}_{\mu,\mu^\circ}\zeta_{\mu^\circ,\mu}(\boldsymbol{x})\zeta_{\mu^\circ,\mu}(\boldsymbol{x})^\top\right]\right)^{1/2}\\ &\quad-\frac{2R_2^2M_4^{1/2}}{\lambda}\left(\operatorname{tr}\mathbf{L}_\mu\mathbb{E}_{\boldsymbol{x}}\left[\zeta_{\mu^\circ,\mu}(\boldsymbol{x})\zeta_{\mu^\circ,\mu}(\boldsymbol{x})^\top\boldsymbol{\Sigma}_{\mu^\circ,\mu}\boldsymbol{\Sigma}_{\mu,\mu}^{-1}\boldsymbol{\Sigma}_{\mu,\mu^\circ}\zeta_{\mu^\circ,\mu}(\boldsymbol{x})\zeta_{\mu^\circ,\mu}(\boldsymbol{x})^\top\right]\right)^{1/2}\\ &\le \frac{r^\circ R_2^2M_4^{1/2}}{\lambda}\left(r^\circ\operatorname{tr}\mathbb{E}_{\boldsymbol{x}}\left[\left(\zeta_{\mu^\circ,\mu}(\boldsymbol{x})\zeta_{\mu^\circ,\mu}(\boldsymbol{x})^\top\right)^2\right]-2\operatorname{tr}\mathbf{L}_\mu\mathbb{E}_{\boldsymbol{x}}\left[\left(\zeta_{\mu^\circ,\mu}(\boldsymbol{x})\zeta_{\mu^\circ,\mu}(\boldsymbol{x})^\top\right)^2\right]\right)^{1/2}\\ &\le \frac{r^{\circ 3/2}R_2^2M_4^{1/2}}{\lambda}\sup_{\boldsymbol{x}}\|\zeta_{\mu^\circ,\mu}(\boldsymbol{x})\|(2\operatorname{tr}\mathbf{L}_\mu)^{1/2}\\ &\le \frac{2k^{1/2}r^{\circ 5/2}R_1^3R_2^2M_4^{1/2}}{\lambda^2}=:C_{\boldsymbol{w}}^2,\end{aligned}$$

where for the last line we have used the coarser bounds $\|\zeta_{\mu^\circ,\mu}(\boldsymbol{x})\| \le R_1 + R_1^3\lambda^{-1}$ and $\lambda \le \frac{1}{k}\operatorname{tr}\boldsymbol{\Sigma}_{\mu_0,\mu_0} \le \frac{R_1^2}{k}$. Combining the two bounds yields

$$\left\|\nabla\frac{\delta\mathcal{L}}{\delta\mu}(\mu,\theta)\right\| \le \left(\frac{r^{\circ 2}R_1^3}{2\lambda^2}(R_1+4k^{1/2}r^{\circ 1/2}R_2^2M_4^{1/2})\right)^{1/2}=:C_1.$$

Furthermore, for $\theta_1=(\boldsymbol{a}_1,\boldsymbol{w}_1)$, $\theta_2=(\boldsymbol{a}_2,\boldsymbol{w}_2)$ we have

$$\begin{aligned}\left\|\nabla_{\boldsymbol{a}}\frac{\delta\mathcal{L}}{\delta\mu}(\mu,\theta_1)-\nabla_{\boldsymbol{a}}\frac{\delta\mathcal{L}}{\delta\mu}(\mu,\theta_2)\right\| &= \left\|\mathbb{E}_{\boldsymbol{x}}\left[\zeta_{\mu^\circ,\mu}(\boldsymbol{x})^\top\boldsymbol{\Sigma}_{\mu^\circ,\mu}\boldsymbol{\Sigma}_{\mu,\mu}^{-1}(\sigma(\boldsymbol{w}_1^\top\boldsymbol{x})-\sigma(\boldsymbol{w}_2^\top\boldsymbol{x}))\right]\right\|\\ &\le R_2\cdot\mathbb{E}_{\boldsymbol{x}}\left[\|\zeta_{\mu^\circ,\mu}(\boldsymbol{x})^\top\boldsymbol{\Sigma}_{\mu^\circ,\mu}\boldsymbol{\Sigma}_{\mu,\mu}^{-1}\|\cdot\|\boldsymbol{w}_1^\top\boldsymbol{x}-\boldsymbol{w}_2^\top\boldsymbol{x}\|\right]\\ &\le R_2M_2^{1/2}\left(\frac{kr^{\circ 2}}{2\lambda}\right)^{1/2}\|\boldsymbol{w}_1-\boldsymbol{w}_2\|,\end{aligned}$$

and also

$$\begin{aligned}&\left\|\nabla_{\boldsymbol{w}}\frac{\delta\mathcal{L}}{\delta\mu}(\mu,\theta_1)-\nabla_{\boldsymbol{w}}\frac{\delta\mathcal{L}}{\delta\mu}(\mu,\theta_2)\right\|\\ &= \left\|\mathbb{E}_{\boldsymbol{x}}\left[\zeta_{\mu^\circ,\mu}(\boldsymbol{x})^\top\boldsymbol{\Sigma}_{\mu^\circ,\mu}\boldsymbol{\Sigma}_{\mu,\mu}^{-1}(\boldsymbol{a}_1\sigma'(\boldsymbol{w}_1^\top\boldsymbol{x})-\boldsymbol{a}_2\sigma'(\boldsymbol{w}_2^\top\boldsymbol{x}))\boldsymbol{x}\right]\right\|\\ &\le \left\|\mathbb{E}_{\boldsymbol{x}}\left[\zeta_{\mu^\circ,\mu}(\boldsymbol{x})^\top\boldsymbol{\Sigma}_{\mu^\circ,\mu}\boldsymbol{\Sigma}_{\mu,\mu}^{-1}(\boldsymbol{a}_1-\boldsymbol{a}_2)\sigma'(\boldsymbol{w}_1^\top\boldsymbol{x})\boldsymbol{x}\right]\right\|\end{aligned}$$

$$
\begin{aligned}
&\quad + \left\|\mathbb{E}_{\boldsymbol{x}}\left[\zeta_{\mu^\circ,\mu}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_{\mu^\circ,\mu}\boldsymbol{\Sigma}_{\mu,\mu}^{-1}\boldsymbol{a}_2(\sigma'(\boldsymbol{w}_1^\top\boldsymbol{x}) - \sigma'(\boldsymbol{w}_2^\top\boldsymbol{x}))\boldsymbol{x}\right]\right\| \\
&\le R_2 M_2^{1/2}\left(\frac{kr^{\circ 2}}{2\lambda}\right)^{1/2}\|\boldsymbol{a}_1 - \boldsymbol{a}_2\| \\
&\quad + R_3 \cdot \mathbb{E}_{\boldsymbol{x}}\left[\|\zeta_{\mu^\circ,\mu}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_{\mu^\circ,\mu}\boldsymbol{\Sigma}_{\mu,\mu}^{-1}\| \cdot \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|\|\boldsymbol{x}\|^2\right] \\
&\le R_2 M_2^{1/2}\left(\frac{kr^{\circ 2}}{2\lambda}\right)^{1/2}\|\boldsymbol{a}_1 - \boldsymbol{a}_2\| + R_3 M_4^{1/2}\left(\frac{kr^{\circ 2}}{2\lambda}\right)^{1/2}\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|.
\end{aligned}
$$

Combining the two yields that

$$
\left\|\nabla\frac{\delta\mathcal{L}}{\delta\mu}(\mu,\theta_1) - \nabla\frac{\delta\mathcal{L}}{\delta\mu}(\mu,\theta_2)\right\| \le \left(\frac{kr^{\circ 2}}{2\lambda}(2R_2^2 M_2 + R_3^2 M_4)\right)^{1/2}\|\theta_1 - \theta_2\| =: C_2\|\theta_1 - \theta_2\|.
$$

□

Lemma E.5. *For any $\mu \in \mathscr{P}_2^+(\Theta)$ such that $\lambda_{\min}(\boldsymbol{\Sigma}_{\mu,\mu}) \ge \lambda$ it holds that $\|\mathbf{H}_\mu(\theta,\theta')\| \le C_3$, $\mathbf{H}_\mu(\theta,\theta')$ is uniformly C_4-Lipschitz w.r.t. θ and θ', and $\mathbf{H}_\mu$ is C_5-Lipschitz w.r.t. μ in 1-Wasserstein distance, where $C_3, C_4 = O(\lambda^{-2})$ and $C_5 = O(d\lambda^{-3})$.*

Proof. To derive regularity estimates of $\mathbf{H}_\mu$, we start from the expansion in Lemma E.3 and perform explicit computations for only the first trace term $t(\mu,\theta,\theta') = \operatorname{tr}\left(\boldsymbol{\Sigma}_{\mu^\circ,\theta'}\boldsymbol{\Sigma}_{\mu,\mu}^{-1}\boldsymbol{\Sigma}_{\theta,\mu^\circ}\right)$. $\nabla_\theta\nabla_{\theta'}t(\mu,\theta,\theta')$ consists of block matrices

$$
\begin{aligned}
\nabla_{\boldsymbol{a}}\nabla_{\boldsymbol{a}'}t(\mu,\theta,\theta') &= \mathbb{E}_{\boldsymbol{x}}\left[\sigma(\boldsymbol{w}^\top\boldsymbol{x})h_{\mu^\circ}(\boldsymbol{x})^\top\right]\mathbb{E}_{\boldsymbol{x}}\left[\sigma(\boldsymbol{w}'^\top\boldsymbol{x})h_{\mu^\circ}(\boldsymbol{x})\right]\boldsymbol{\Sigma}_{\mu,\mu}^{-1}, \\
\nabla_{\boldsymbol{a}}\nabla_{\boldsymbol{w}'}t(\mu,\theta,\theta') &= \boldsymbol{\Sigma}_{\mu,\mu}^{-1}\boldsymbol{a}'\mathbb{E}_{\boldsymbol{x}}\left[\sigma(\boldsymbol{w}^\top\boldsymbol{x})h_{\mu^\circ}(\boldsymbol{x})^\top\right]\mathbb{E}_{\boldsymbol{x}}\left[\sigma'(\boldsymbol{w}'^\top\boldsymbol{x})h_{\mu^\circ}(\boldsymbol{x})\boldsymbol{x}^\top\right], \\
\nabla_{\boldsymbol{w}}\nabla_{\boldsymbol{a}'}t(\mu,\theta,\theta') &= \mathbb{E}_{\boldsymbol{x}}\left[\sigma'(\boldsymbol{w}^\top\boldsymbol{x})\boldsymbol{x}h_{\mu^\circ}(\boldsymbol{x})^\top\right]\mathbb{E}_{\boldsymbol{x}}\left[\sigma(\boldsymbol{w}'^\top\boldsymbol{x})h_{\mu^\circ}(\boldsymbol{x})\right]\boldsymbol{a}^\top\boldsymbol{\Sigma}_{\mu,\mu}^{-1}, \\
\nabla_{\boldsymbol{w}}\nabla_{\boldsymbol{w}'}t(\mu,\theta,\theta') &= \mathbb{E}_{\boldsymbol{x}}\left[\sigma'(\boldsymbol{w}^\top\boldsymbol{x})\boldsymbol{x}h_{\mu^\circ}(\boldsymbol{x})^\top\right]\mathbb{E}_{\boldsymbol{x}}\left[\sigma'(\boldsymbol{w}'^\top\boldsymbol{x})h_{\mu^\circ}(\boldsymbol{x})\boldsymbol{x}^\top\right]\boldsymbol{a}'^\top\boldsymbol{\Sigma}_{\mu,\mu}^{-1}\boldsymbol{a}.
\end{aligned}
$$

It follows from Lemma C.1 that $\|\nabla_\theta\nabla_{\theta'}t(\mu,\theta,\theta')\| \le (R_1^4 + 2R_1^2R_2k^{1/2}r^{\circ 1/2}M_2^{1/2} + R_2^2kr^\circ M_2)\lambda^{-1} = O(\lambda^{-1})$. Each term of $\mathbf{H}_\mu$ is likewise uniformly bounded so that $\mathbf{H}_\mu$ is a valid kernel.

The Lipschitz constant of $\nabla_\theta\nabla_{\theta'}t(\mu,\theta,\theta')$ w.r.t. θ can also be controlled by separately bounding

$$
\begin{aligned}
&\|\nabla_{\boldsymbol{a}}\nabla_{\boldsymbol{a}'}t(\mu,\theta_1,\theta') - \nabla_{\boldsymbol{a}}\nabla_{\boldsymbol{a}'}t(\mu,\theta_2,\theta')\| \\
&\quad\le \mathbb{E}_{\boldsymbol{x}}\left[|\sigma(\boldsymbol{w}_1^\top\boldsymbol{x}) - \sigma(\boldsymbol{w}_2^\top\boldsymbol{x})| \cdot \|h_{\mu^\circ}(\boldsymbol{x})\|\right]\mathbb{E}_{\boldsymbol{x}}\left[|\sigma(\boldsymbol{w}'^\top\boldsymbol{x})| \cdot \|h_{\mu^\circ}(\boldsymbol{x})\|\right]\|\boldsymbol{\Sigma}_{\mu,\mu}^{-1}\| \\
&\quad\le R_1R_2M_2^{1/2}kr^\circ\lambda^{-1}\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|, \\
&\|\nabla_{\boldsymbol{a}}\nabla_{\boldsymbol{w}'}t(\mu,\theta_1,\theta') - \nabla_{\boldsymbol{a}}\nabla_{\boldsymbol{w}'}t(\mu,\theta_2,\theta')\| \\
&\quad\le \|\boldsymbol{\Sigma}_{\mu,\mu}^{-1}\boldsymbol{a}'\| \cdot \mathbb{E}_{\boldsymbol{x}}\left[|\sigma(\boldsymbol{w}_1^\top\boldsymbol{x}) - \sigma(\boldsymbol{w}_2^\top\boldsymbol{x})| \cdot \|h_{\mu^\circ}(\boldsymbol{x})\|\right]\mathbb{E}_{\boldsymbol{x}}\left[|\sigma'(\boldsymbol{w}'^\top\boldsymbol{x})| \cdot \|h_{\mu^\circ}(\boldsymbol{x})\boldsymbol{x}^\top\|\right] \\
&\quad\le R_2^2M_2kr^\circ\lambda^{-1}\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|, \\
&\|\nabla_{\boldsymbol{w}}\nabla_{\boldsymbol{a}'}t(\mu,\theta_1,\theta') - \nabla_{\boldsymbol{w}}\nabla_{\boldsymbol{a}'}t(\mu,\theta_2,\theta')\| \\
&\quad\le \mathbb{E}_{\boldsymbol{x}}\left[|\sigma'(\boldsymbol{w}_1^\top\boldsymbol{x}) - \sigma'(\boldsymbol{w}_2^\top\boldsymbol{x})| \cdot \|\boldsymbol{x}h_{\mu^\circ}(\boldsymbol{x})^\top\|\right]\mathbb{E}_{\boldsymbol{x}}\left[|\sigma(\boldsymbol{w}'^\top\boldsymbol{x})| \cdot \|h_{\mu^\circ}(\boldsymbol{x})\|\right]\|\boldsymbol{a}_1^\top\boldsymbol{\Sigma}_{\mu,\mu}^{-1}\| \\
&\qquad + \mathbb{E}_{\boldsymbol{x}}\left[|\sigma'(\boldsymbol{w}_2^\top\boldsymbol{x})| \cdot \|\boldsymbol{x}h_{\mu^\circ}(\boldsymbol{x})^\top\|\right]\mathbb{E}_{\boldsymbol{x}}\left[|\sigma(\boldsymbol{w}'^\top\boldsymbol{x})| \cdot \|h_{\mu^\circ}(\boldsymbol{x})\|\right]\|(\boldsymbol{a}_1 - \boldsymbol{a}_2)^\top\boldsymbol{\Sigma}_{\mu,\mu}^{-1}\| \\
&\quad\le R_1R_3M_4^{1/2}kr^\circ\lambda^{-1}\|\boldsymbol{w}_1 - \boldsymbol{w}_2\| + R_1R_2M_2^{1/2}kr^\circ\lambda^{-1}\|\boldsymbol{a}_1 - \boldsymbol{a}_2\|, \\
&\|\nabla_{\boldsymbol{w}}\nabla_{\boldsymbol{w}'}t(\mu,\theta_1,\theta') - \nabla_{\boldsymbol{w}}\nabla_{\boldsymbol{w}'}t(\mu,\theta_2,\theta')\| \\
&\quad\le \mathbb{E}_{\boldsymbol{x}}\left[|\sigma'(\boldsymbol{w}_1^\top\boldsymbol{x}) - \sigma'(\boldsymbol{w}_2^\top\boldsymbol{x})| \cdot \|\boldsymbol{x}h_{\mu^\circ}(\boldsymbol{x})^\top\|\right]\mathbb{E}_{\boldsymbol{x}}\left[|\sigma'(\boldsymbol{w}'^\top\boldsymbol{x})| \cdot \|h_{\mu^\circ}(\boldsymbol{x})\boldsymbol{x}^\top\|\right]\|\boldsymbol{a}'^\top\boldsymbol{\Sigma}_{\mu,\mu}^{-1}\boldsymbol{a}_1\| \\
&\qquad + \mathbb{E}_{\boldsymbol{x}}\left[|\sigma'(\boldsymbol{w}_2^\top\boldsymbol{x})| \cdot \|\boldsymbol{x}h_{\mu^\circ}(\boldsymbol{x})^\top\|\right]\mathbb{E}_{\boldsymbol{x}}\left[|\sigma'(\boldsymbol{w}'^\top\boldsymbol{x})| \cdot \|h_{\mu^\circ}(\boldsymbol{x})\boldsymbol{x}^\top\|\right]\|\boldsymbol{a}'^\top\boldsymbol{\Sigma}_{\mu,\mu}^{-1}(\boldsymbol{a}_1 - \boldsymbol{a}_2)\| \\
&\quad\le R_2R_3M_2^{1/2}M_4^{1/2}kr^\circ\lambda^{-1}\|\boldsymbol{w}_1 - \boldsymbol{w}_2\| + R_2^2M_2kr^\circ\lambda^{-1}\|\boldsymbol{a}_1 - \boldsymbol{a}_2\|.
\end{aligned}
$$

Therefore, $\nabla_\theta\nabla_{\theta'}t(\mu,\theta,\theta')$ is uniformly $O(\lambda^{-1})$-Lipschitz w.r.t. both θ and θ' by symmetry. All the remaining terms can also be bounded with at most an $O(\lambda^{-2})$ Lipschitz constant; in particular,

the terms including three factors of $\mathbf{\Sigma}_{\mu,\mu}^{-1}$ can be controlled by removing a factor of $\lambda^{-1/2}$ twice and isolating $\mathbf{\Sigma}_{\mu,\mu}^{-1/2}\mathbf{\Sigma}_{\mu,\mu^\circ}$ and $\mathbf{\Sigma}_{\mu^\circ,\mu}\mathbf{\Sigma}_{\mu,\mu}^{-1/2}$ as in the proof of Lemma E.4.

Finally, the third-order functional derivative $\nabla_{\tilde{\theta}}\frac{\delta}{\delta\mu}\mathbf{H}_\mu(\theta,\theta')(\tilde{\theta})$ can be bounded in a similar manner with spectral norm at most $O(\lambda^{-3})$, yielding via Kantorovich-Rubinstein duality that

$$\begin{aligned}\|\mathbf{H}_{\mu_1}(\theta,\theta') - \mathbf{H}_{\mu_2}(\theta,\theta')\| &= \left\|\int \frac{\delta}{\delta\mu}\mathbf{H}_{(1-s)\mu_1+s\mu_2}(\theta,\theta')(\tilde{\theta})(\mu_2-\mu_1)(\mathrm{d}\tilde{\theta})\right\| \\ &\lesssim (k+d)\lambda^{-3}\cdot\mathcal{W}_1(\mu_1,\mu_2).\end{aligned}$$

The additional $k+d$ factor arises from bounding each entry of $\mathbf{H}_{\mu_1} - \mathbf{H}_{\mu_2}$ separately. We omit the details. □

Proposition E.6. *Let F be a functional depending on μ only through the MLP layer h_μ. Suppose MFD (3) at time t admits a distribution $\bar{\mu} \in \mathscr{P}_2(\Theta)$ with $\chi^2(\bar{\mu},\mu_t) \le \bar{\chi}^2$ such that $\frac{\mathrm{d}^2}{\mathrm{d}s^2}\big|_{s=0}F(\bar{\mu}_s) \le -\Lambda$. Then the smallest eigenvalue λ_0 of $\mathcal{H}_{\mu_t}$ satisfies $\lambda_0 \le -\bar{\chi}^{-2}\Lambda$.*

Proof. The second derivative along the linear homotopy $\bar{\mu}_s$ can be expanded as

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}s^2}\bigg|_{s=0}F(\bar{\mu}_s) &= \frac{\mathrm{d}}{\mathrm{d}s}\bigg|_{s=0}\int\frac{\delta F}{\delta\mu}(\bar{\mu}_s,\theta)(\bar{\mu}-\mu_t)(\mathrm{d}\theta) \\ &= \iint\frac{\delta^2 F}{\delta\mu^2}(\mu_t,\theta,\theta')(\bar{\mu}-\mu_t)(\mathrm{d}\theta)(\bar{\mu}-\mu_t)(\mathrm{d}\theta').\end{aligned}$$

Now similarly to the proof of Proposition 5.1, denoting $\theta=(\boldsymbol{a},\boldsymbol{w}), \theta'=(\boldsymbol{a}',\boldsymbol{w}')$ we can exploit the fact that $\frac{\delta^2 F}{\delta\mu^2}(\mu,\theta,\theta')$ is bilinear in $\boldsymbol{a},\boldsymbol{a}'$ to relate it to the kernel $\mathbf{H}_\mu$,

$$\begin{aligned}\frac{\delta^2 F}{\delta\mu^2}(\mu,\theta,\theta') &= \boldsymbol{a}^\top\left[\nabla_{\boldsymbol{a}}\nabla_{\boldsymbol{a}'}\frac{\delta^2 F}{\delta\mu^2}(\mu,\theta,\theta')\right]\boldsymbol{a}' = (\boldsymbol{a}\ \ 0_d)^\top\left[\nabla_\theta\nabla_{\theta'}\frac{\delta^2 F}{\delta\mu^2}(\mu,\theta,\theta')\right](\boldsymbol{a}'\ \ 0_d) \\ &= (\boldsymbol{a}\ \ 0_d)^\top\mathbf{H}_\mu(\theta,\theta')(\boldsymbol{a}'\ \ 0_d).\end{aligned}$$

Writing the eigenfunction decomposition of $\mathbf{H}_{\mu_t}$ as (omitting the dependency on t for brevity)

$$\mathbf{H}_{\mu_t}(\theta,\theta') = \sum_{j\in\mathbb{Z}}\lambda_j\psi_j(\theta)\psi_j(\theta')^\top, \quad \int\|\psi_j\|^2\,\mathrm{d}\mu_t = 1 \quad \forall j\in\mathbb{Z},$$

with the ordering $\lambda_1\ge\lambda_2\ge\cdots\ge 0$ and $\lambda_0\le\lambda_{-1}\le\cdots\le 0$, we may thus bound

$$\begin{aligned}-\Lambda \ge \frac{\mathrm{d}^2}{\mathrm{d}s^2}\bigg|_{s=0}F(\bar{\mu}_s) &= \iint(\boldsymbol{a}\ \ 0_d)^\top\mathbf{H}_\mu(\theta,\theta')(\boldsymbol{a}'\ \ 0_d)(\bar{\mu}-\mu_t)(\mathrm{d}\theta)(\bar{\mu}-\mu_t)(\mathrm{d}\theta') \\ &= \sum_{j\in\mathbb{Z}}\lambda_j\left(\int(\boldsymbol{a}\ \ 0_d)^\top\psi_j(\theta)(\bar{\mu}-\mu_t)(\mathrm{d}\theta)\right)^2 \\ &\ge -|\lambda_0|\cdot\sum_{j\in\mathbb{Z}}\left(\int(\boldsymbol{a}\ \ 0_d)^\top\psi_j(\theta)(\bar{\mu}-\mu_t)(\mathrm{d}\theta)\right)^2 \\ &= -|\lambda_0|\cdot\sum_{j\in\mathbb{Z}}\left(\int\left(\frac{\mathrm{d}\bar{\mu}}{\mathrm{d}\mu_t}-1\right)(\boldsymbol{a}\ \ 0_d)^\top\psi_j(\theta)\mu_t(\mathrm{d}\theta)\right)^2 \\ &= -|\lambda_0|\int\left(\frac{\mathrm{d}\bar{\mu}}{\mathrm{d}\mu_t}-1\right)^2\|\boldsymbol{a}\|^2\mu_t(\mathrm{d}\theta) \\ &\ge -\bar{\chi}^2|\lambda_0|,\end{aligned}$$

where we have made use of Parseval's identity. Hence the largest negative eigenvalue is bounded as $\lambda_0 \le -\bar{\chi}^{-2}\Lambda$. □

Theorem E.7. *Assume $F : \mathscr{P}_2(\Omega) \to \mathbb{R}$, $\Omega \subseteq \mathbb{R}^m$ satisfies $\|\nabla \frac{\delta F}{\delta \mu}\| \le C_1$, $\nabla \frac{\delta F}{\delta \mu}$ is C_2-Lipschitz, $\mathbf{H}_\mu$ is Hilbert-Schmidt, $\|\mathbf{H}_\mu\| \le C_3$, $\mathbf{H}_\mu(\theta, \theta')$ is C_4-Lipschitz w.r.t. θ, θ' and C_5-Lipschitz w.r.t. μ in $\mathcal{W}_1$. Further suppose that $\lambda_0 := \lambda_{\min}(\mathcal{H}_{\mu^\dagger}) < 0$ and the corresponding eigenfunction ψ_0 satisfies $|\int \psi_0^\top \nabla \frac{\delta \mathcal{L}}{\delta \mu}(\mu_t)\, \mathrm{d}\mu_t| \ge \alpha$ for some $\alpha > 0$. Then WGF initialized at $\mu_0 = \mu^\dagger$ decreases F by at least $F(\mu_\tau) \le F(\mu_0) - \Omega\left(\frac{|\lambda_0|\alpha}{\sqrt{m}\tau}\right)$ in time $\tau = O\left(\frac{1}{|\lambda_0|} \log \frac{|\lambda_0|}{\sqrt{m}\alpha}\right)$.*

Unlike before, F can be completely general and does not need to depend on μ through an MLP layer.

Proof. First note that the function $\theta' \mapsto \mathbf{H}_{\mu^\dagger}(\theta, \theta') \nabla \frac{\delta F}{\delta \mu}(\mu_t, \theta')$ is uniformly Lipschitz: for any θ_1', θ_2',

$$\begin{aligned}
&\left\| \mathbf{H}_{\mu^\dagger}(\theta, \theta_1') \nabla \frac{\delta F}{\delta \mu}(\mu_t, \theta_1') - \mathbf{H}_{\mu^\dagger}(\theta, \theta_2') \nabla \frac{\delta F}{\delta \mu}(\mu_t, \theta_2') \right\| \\
&\le \|\mathbf{H}_{\mu^\dagger}(\theta, \theta_1') - \mathbf{H}_{\mu^\dagger}(\theta, \theta_2')\| \cdot \left\| \nabla \frac{\delta F}{\delta \mu}(\mu_t, \theta_1') \right\| \\
&\qquad + \|\mathbf{H}_{\mu^\dagger}(\theta, \theta_2')\| \cdot \left\| \nabla \frac{\delta F}{\delta \mu}(\mu_t, \theta_1') - \nabla \frac{\delta F}{\delta \mu}(\mu_t, \theta_2') \right\| \\
&\le C_1 C_4 \|\theta_1' - \theta_2'\| + C_2 C_3 \|\theta_1' - \theta_2'\|.
\end{aligned}$$

We re-expand the evolution equation (Lemma 5.3) for the dynamics $(\mu_t)_{t \ge 0}$ around $\mu^\dagger$ as

$$\begin{aligned}
\partial_t \left[\nabla_\theta \frac{\delta F}{\delta \mu}(\mu_t, \theta) \right] &= -\int \mathbf{H}_{\mu_t}(\theta, \theta') \nabla_{\theta'} \frac{\delta F}{\delta \mu}(\mu_t, \theta') \mu_t(\mathrm{d}\theta') \\
&=: -\int \mathbf{H}_{\mu^\dagger}(\theta, \theta') \nabla_{\theta'} \frac{\delta F}{\delta \mu}(\mu_t, \theta') \mu^\dagger(\mathrm{d}\theta') + e(t, \theta),
\end{aligned}$$

where the difference or error function $e(t, \theta)$ can be bounded as

$$\begin{aligned}
\|e(t, \theta)\| &\le \left\| \int (\mathbf{H}_{\mu_t}(\theta, \theta') - \mathbf{H}_{\mu^\dagger}(\theta, \theta')) \nabla_{\theta'} \frac{\delta F}{\delta \mu}(\mu_t, \theta') \mu_t(\mathrm{d}\theta') \right\| \\
&\qquad + \left\| \int \mathbf{H}_{\mu^\dagger}(\theta, \theta') \nabla_{\theta'} \frac{\delta F}{\delta \mu}(\mu_t, \theta') (\mu_t - \mu^\dagger)(\mathrm{d}\theta') \right\| \\
&\le \left(C_1 C_5 + (C_1 C_4 + C_2 C_3) m^{1/2} \right) \mathcal{W}_1(\mu_t, \mu^\dagger). \\
&=: C_6\, \mathcal{W}_1(\mu_t, \mu^\dagger).
\end{aligned}$$

For the second term, we have used the Lipschitz constant derived above to bound each entry separately. Then the ψ_0-component $\alpha_0(t) := \int \psi_0^\top \nabla \frac{\delta F}{\delta \mu}(\mu_t)\, \mathrm{d}\mu^\dagger$ of the gradient evolves according to

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} \alpha_0(t) &= -\iint \psi_0(\theta)^\top \mathbf{H}_{\mu^\dagger}(\theta, \theta') \nabla_{\theta'} \frac{\delta F}{\delta \mu}(\mu_t, \theta') \mu^\dagger(\mathrm{d}\theta') \mu^\dagger(\mathrm{d}\theta) + \int \psi_0(\theta)^\top e(t, \theta) \mu^\dagger(\mathrm{d}\theta) \\
&= -\lambda_0 \int \psi_0(\theta)^\top \nabla_\theta \frac{\delta F}{\delta \mu}(\mu_t, \theta) \mu^\dagger(\mathrm{d}\theta) + \int \psi_0(\theta)^\top e(t, \theta) \mu^\dagger(\mathrm{d}\theta),
\end{aligned}$$

and hence

$$\left| \frac{\mathrm{d}}{\mathrm{d}t} \alpha_0(t) + \lambda_0 \alpha_0(t) \right| \le \left(\int \|\psi_0\|^2\, \mathrm{d}\mu^\dagger \right)^{1/2} \sup_{\theta \in \Theta} \|e(t, \theta)\| \le C_8\, \mathcal{W}_1(\mu_t, \mu^\dagger).$$

Without loss of generality, assume initially $\alpha_0(0)$ is positive so that $\alpha_0(0) \ge \alpha$. We consider a 1-Wasserstein ball centered at $\mu^\dagger$ with radius small enough so that the error term is negligible compared to the exponential growth,

$$\mathcal{B}_{\mathcal{W}}(\Delta) = \left\{ \mu \in \mathscr{P}_2(\Theta) : \mathcal{W}_1(\mu, \mu^\dagger) \le \Delta := \frac{|\lambda_0|\alpha}{2C_6} \right\}.$$

Then for a set time interval $\tau > 0$ to be determined, either of the following must happen:

(i) $(\mu_t)_{t\in[0,\tau]} \subset \mathcal{B}_{\mathcal{W}}(\Delta)$. In this case, $\alpha_0(t)$ grows exponentially during the entire interval $t \in [0,\tau]$ as

$$\frac{\mathrm{d}}{\mathrm{d}t}\alpha_0(t) \geq |\lambda_0|\alpha_0(t) - C_6\Delta = |\lambda_0|\left(\alpha_0(t) - \frac{\alpha}{2}\right) > 0,$$

showing that

$$\alpha_0(t) \geq e^{|\lambda_0|t}\left(\alpha_0(0) - \frac{\alpha}{2}\right) + \frac{\alpha}{2} \geq \frac{\alpha(e^{|\lambda_0|t}+1)}{2}.$$

Then the decrease of F after time τ can be bounded below by retrieving the ψ_0-component as

$$\begin{aligned}
F(\mu_0) - F(\mu_\tau) &= \int_0^\tau \int \left\|\nabla\frac{\delta F}{\delta\mu}(\mu_t,\theta)\right\|^2 \mu_t(\mathrm{d}\theta)\,\mathrm{d}t \\
&\geq \int_0^\tau \left(\int \left\|\nabla\frac{\delta F}{\delta\mu}(\mu_t,\theta)\right\|^2 \mu^\dagger(\mathrm{d}\theta)\,\mathrm{d}t - 2C_1C_2\,\mathcal{W}_1(\mu_t,\mu^\dagger)\right) \\
&\geq \int_0^\tau \left(\int \psi_0(\theta)^\top \nabla\frac{\delta F}{\delta\mu}(\mu_t,\theta)\mu^\dagger(\mathrm{d}\theta)\right)^2 \mathrm{d}t - 2C_1C_2\Delta\tau \\
&= \int_0^\tau \alpha_0(t)^2\,\mathrm{d}t - 2C_1C_2\Delta\tau \\
&\geq \frac{\alpha^2}{4}\left(\frac{1}{2|\lambda_0|}(e^{2|\lambda_0|\tau}-1) + \frac{2}{|\lambda_0|}(e^{|\lambda_0|\tau}-1) + \tau\right) - \frac{C_1C_2}{C_6}|\lambda_0|\alpha\tau.
\end{aligned}$$

(ii) $\mu_{\tau_e} \notin \mathcal{B}_{\mathcal{W}}(\Delta)$ for some $\tau_e \leq \tau$. If the mean-field flow has managed to escape the ball $\mathcal{B}_{\mathcal{W}}(\Delta)$ in time τ_e, the Benamou-Brenier formula (Proposition A.3) immediately guarantees that

$$\begin{aligned}
F(\mu_0) - F(\mu_\tau) &\geq F(\mu_0) - F(\mu_{\tau_e}) \\
&= \int_0^{\tau_e} \int \left\|\nabla\frac{\delta F}{\delta\mu}(\mu_t,\theta)\right\|^2 \mu_t(\mathrm{d}\theta)\,\mathrm{d}t \\
&\geq \frac{\mathcal{W}_2(\mu_{\tau_e},\mu^\dagger)}{\tau_e} > \frac{\Delta}{\tau}.
\end{aligned}$$

Thus we have proved that:

$$\begin{aligned}
&F(\mu_0) - F(\mu_\tau) \\
&\geq \left(\frac{\alpha^2}{4}\left(\frac{1}{2|\lambda_0|}(e^{2|\lambda_0|\tau}-1) + \frac{2}{|\lambda_0|}(e^{|\lambda_0|\tau}-1) + \tau\right) - \frac{C_1C_2}{C_6}|\lambda_0|\alpha\tau\right) \wedge \frac{|\lambda_0|\alpha}{2C_6\tau}. \qquad (10)
\end{aligned}$$

Due to the exponential terms, we see $\tau \approx \log\frac{1}{\alpha}$ is enough to ensure that the two terms become roughly equal so that the guarantees is close to optimal. For the remainder of the proof, we derive the exact formula. Choose

$$\tau = \frac{1}{|\lambda_0|}\log\frac{C_7}{\alpha}$$

for some $C_7 > \alpha$. The first term in the right-hand side of (10) can be bounded as

$$\begin{aligned}
&\frac{\alpha^2}{4}\left(\frac{1}{2|\lambda_0|}(e^{2|\lambda_0|\tau}-1) + \frac{2}{|\lambda_0|}(e^{|\lambda_0|\tau}-1) + \tau\right) - \frac{C_1C_2}{C_6}|\lambda_0|\alpha\tau \\
&= \frac{\alpha^2}{8|\lambda_0|}\left(\frac{C_7^2}{\alpha^2} - 1\right) + \frac{\alpha^2}{2|\lambda_0|}\left(\frac{C_7}{\alpha} - 1\right) + \frac{\alpha^2}{4|\lambda_0|}\log\frac{C_7}{\alpha} - \frac{C_1C_2}{C_6}\alpha\log\frac{C_7}{\alpha} \\
&\geq \left(\frac{C_7^2}{24|\lambda_0|} - \frac{5\alpha^2}{8|\lambda_0|}\right) + \left(\frac{C_7^2}{24|\lambda_0|} - \frac{C_1C_2}{C_6}\alpha\log\frac{C_7}{\alpha}\right) + \frac{C_7^2}{24|\lambda_0|} + \frac{C_7\alpha}{2|\lambda_0|} + \frac{\alpha^2}{4|\lambda_0|}\log\frac{C_7}{\alpha} \\
&\geq \left(\frac{C_7^2}{24|\lambda_0|} - \frac{5\alpha^2}{8|\lambda_0|}\right) + \left(\frac{C_7^2}{24|\lambda_0|} - \frac{C_1C_2C_7}{C_6e}\right) + \frac{C_7^2}{24|\lambda_0|},
\end{aligned}$$

where we have used the fact that the function $x \mapsto x\log\frac{c}{x}$ has maximum $\frac{c}{e}$. Then the first term of (10) will dominate the second as long as

$$\frac{C_7^2}{24|\lambda_0|} \geq \frac{5\alpha^2}{8|\lambda_0|} \vee \frac{C_1C_2C_7}{C_6e} \vee \frac{|\lambda_0|^2\alpha}{2C_6\log\frac{C_7}{\alpha}}.$$

Manipulating terms shows that

$$C_7 = \sqrt{15}\alpha \vee \frac{24C_1C_2|\lambda_0|}{C_6 e} \vee \left(\frac{24|\lambda_0|^3\alpha}{C_6 \log 15}\right)^{1/2}$$

is sufficient. For the purposes of the general statement, we focus on asymptotic behavior w.r.t. α and hide all regularity constants $C_1, \cdots, C_5$, yielding $C_6 = O(m^{1/2})$ and $C_7 = O(|\lambda_0| m^{-1/2})$. □

Proof of Theorem 5.5. Let us fix the lower bound $\lambda_{\min}(\mathbf{\Sigma}_{\mu_r,\mu_r}) \geq \lambda = \Theta(\frac{1}{k})$. (The bound only needs to hold either locally for the $\mathcal{W}_2$-ball of radius Δ in the proof of Theorem E.7, or along the dynamics μ_t until escape.) We first need a robust version of Theorem 3.3(ii) since μ_t cannot be exactly on a critical point. If $\frac{\mathrm{d}}{\mathrm{d}s}\big|_{s=0} \mathcal{L}(\bar{\mu}_s) > -\delta$ it must hold that $\|\mathbf{L}_\mu \mathbf{\Sigma}_{\mu^\circ,\mu} \mathbf{\Sigma}_{\mu,\mu}^{-1}\|_* < \frac{\delta}{2}$ by (4). Then from (5), again choosing $\mathbf{R} \in \mathcal{O}(k)$ such that $\mathbf{\Sigma}_{\mu^\circ,\mu}\mathbf{\Sigma}_{\mu,\mu}^{-1}\mathbf{R}$ is symmetric,

$$\begin{aligned}
&\frac{\mathrm{d}^2}{\mathrm{d}s^2}\Big|_{s=0} \mathcal{L}(\bar{\mu}_s) \\
&= -4\,\mathrm{tr}\left(\mathbf{L}_\mu^2 \mathbf{R}^\top \mathbf{\Sigma}_{\mu,\mu}^{-1}\mathbf{R}\right) + 2\,\mathrm{tr}\left(\mathbf{L}_\mu(2\mathbf{\Sigma}_{\mu^\circ,\mu}\mathbf{\Sigma}_{\mu,\mu}^{-1}\mathbf{R} + \mathbf{R}^\top \mathbf{\Sigma}_{\mu,\mu}^{-1}\mathbf{\Sigma}_{\mu,\mu^\circ} - 2\mathbf{I}_k)\mathbf{\Sigma}_{\mu^\circ,\mu}\mathbf{\Sigma}_{\mu,\mu}^{-1}\mathbf{R}\right) \\
&= -4\,\mathrm{tr}\left(\mathbf{L}_\mu^2 \mathbf{R}^\top \mathbf{\Sigma}_{\mu,\mu}^{-1}\mathbf{R}\right) + 2\,\mathrm{tr}\left((2\mathbf{\Sigma}_{\mu^\circ,\mu}\mathbf{\Sigma}_{\mu,\mu}^{-1}\mathbf{R} + \mathbf{R}^\top \mathbf{\Sigma}_{\mu,\mu}^{-1}\mathbf{\Sigma}_{\mu,\mu^\circ} - 2\mathbf{I}_k)^\top \mathbf{L}_\mu \mathbf{\Sigma}_{\mu^\circ,\mu}\mathbf{\Sigma}_{\mu,\mu}^{-1}\mathbf{R}\right) \\
&\leq -\frac{4}{kR_1^2}\mathcal{L}(\mu_t)^2 + 2\|\mathbf{L}_\mu \mathbf{\Sigma}_{\mu^\circ,\mu}\mathbf{\Sigma}_{\mu,\mu}^{-1}\|_* \|2\mathbf{\Sigma}_{\mu^\circ,\mu}\mathbf{\Sigma}_{\mu,\mu}^{-1}\mathbf{R} + \mathbf{R}^\top \mathbf{\Sigma}_{\mu,\mu}^{-1}\mathbf{\Sigma}_{\mu,\mu^\circ} - 2\mathbf{I}_k\| \\
&\leq -\frac{4}{kR_1^2}\mathcal{L}(\mu_t)^2 + \left(3\|\mathbf{\Sigma}_{\mu^\circ,\mu}\mathbf{\Sigma}_{\mu,\mu}^{-1/2}\| \cdot \|\mathbf{\Sigma}_{\mu,\mu}^{-1/2}\mathbf{R}\| + 2\right)\delta \\
&\leq -\frac{4}{kR_1^2}\mathcal{L}(\mu_t)^2 + (3r^{\circ 1/2}\lambda^{-1/2} + 2)\delta.
\end{aligned}$$

Hence if we take

$$\delta \leq \frac{2}{kR_1^2(3r^{\circ 1/2}\lambda^{-1/2} + 2)}\mathcal{L}(\mu_t)^2$$

then $\frac{\mathrm{d}^2}{\mathrm{d}s^2}\big|_{s=0} \mathcal{L}(\bar{\mu}_s) \leq -\frac{2}{kR_1^2}\mathcal{L}(\mu_t)^2$, and by Proposition E.6 it holds that

$$\lambda_0 = \lambda_{\min}(\mathcal{H}_{\mu_t}) \leq -\frac{2\gamma}{kR_1^2 R_4}\mathcal{L}(\mu_t)^2.$$

Then Theorem E.7 applies to $F = \mathcal{L}$ by virtue of Lemma E.3 and the regularity constants derived in Lemma E.4 and E.5. One can check that

$$C_6 = C_1C_5 + (C_1C_4 + C_2C_3)(k+d)^{1/2} = O\left(\frac{d}{k\lambda^4}\right)$$

and

$$C_7 = O\left(\alpha \vee \frac{\lambda^{5/2}\gamma}{k^{3/2}d}\mathcal{L}(\mu_t)^2 \vee \frac{\lambda^2\gamma^{3/2}\alpha^{1/2}}{kd^{1/2}}\mathcal{L}(\mu_t)^3\right) = O\left(\alpha + \frac{\gamma}{k^4 d}\right);$$

the third term is dominated by the geometric mean of the first two, and $\mathcal{L}(\mu_t) = O(1)$. Hence the time interval of interest is

$$\tau = O\left(\frac{k}{\gamma\,\mathcal{L}(\mu_t)^2}\left(\log\frac{\gamma}{k^4 d\alpha}\right) \vee 1\right),$$

and the guaranteed decrease of the objective is

$$\mathcal{L}(\mu_t) - \mathcal{L}(\mu_{t+\tau}) \geq \frac{|\lambda_0|\alpha}{2C_6\tau} \geq \Omega\left(\frac{\gamma^2\alpha\,\mathcal{L}(\mu_t)^4}{k^5 d}\left(\log\frac{\gamma}{k^4 d\alpha} \vee 1\right)^{-1}\right).$$

E.3 Escaping from Saddle Points Efficiently

Theorem 5.5 on its own cannot ensure convergence rates. The flow might be initialized at or pass near multiple saddle points with very small α values, taking longer to escape. This is an unavoidable problem of nonconvex gradient descent even in finite dimensions (Du et al., 2017). In contrast, it has

been shown that simply adding uniform noise allows GD to escape saddle points efficiently (Ge et al., 2015; Jin et al., 2017). Here, we suggest an adaptation to WGF.

The main problem is how to apply 'random' perturbations in $\mathscr{P}_2(\Omega)$. Motivated by the characterization of the tangent space (6), we propose a scheme which constructs perturbations in the velocity space using vector-valued Gaussian processes.

Definition E.8 (vector-valued Gaussian process)**.** The random function $\boldsymbol{\xi} : \Omega \to \mathbb{R}^m$ is said to follow a Gaussian process if any finite collection of variables $\boldsymbol{\xi}(\theta_1), \cdots, \boldsymbol{\xi}(\theta_j)$ are jointly normally distributed. The process is determined by the mean function $\boldsymbol{m} : \Omega \to \mathbb{R}^m$, $\boldsymbol{m}(\theta) = \mathbb{E}[\boldsymbol{\xi}(\theta)]$ and matrix-valued covariance function

$$\mathbf{K} : \Omega \times \Omega \to \mathbb{R}^{m\times m}, \quad \mathbf{K}(\theta, \theta') = \mathbb{E}[(\boldsymbol{\xi}(\theta) - \boldsymbol{m}(\theta))(\boldsymbol{\xi}(\theta') - \boldsymbol{m}(\theta'))^\top].$$

We denote this process as $\boldsymbol{\xi} \sim \mathrm{GP}(\boldsymbol{m}, \mathbf{K})$. See Álvarez et al. (2012) for further details.

1. Generate a random velocity field $\boldsymbol{\xi} : \Omega \to \mathbb{R}^m$ from a stationary Gaussian process $\mathrm{GP}(0, \mathbf{K})$ with bounded kernel $\mathbf{K} : \Omega \times \Omega \to \mathbb{R}^{m\times m}$.
2. Run the pushforward dynamics $\partial_t \mu_t = \nabla \cdot (\boldsymbol{\xi} \mu_t)$ from $\mu_0 = \mu^\dagger$ for fixed time Δt.

This can bypass the dimensional dependency in Ge et al. (2015) and ensure a nonzero ψ_0-component for $\nabla \frac{\delta F}{\delta \mu}(\mu_{\Delta t})$, which is approximately normally distributed with variance $O(\Delta t)$.

Lemma E.9. *For any $\mu \in \mathscr{P}_2(\Omega)$, square-integrable test function $\psi \in L^2(\Omega, \mu; \mathbb{R}^m)$ and covariance function $\mathbf{K} : \Omega \times \Omega \to \mathbb{R}^{m\times m}$ satisfying $\int \|\mathbf{K}(\theta, \theta)\| \mu(\mathrm{d}\theta) < \infty$ the inner product $\langle \psi, \boldsymbol{\xi} \rangle_{L^2(\Omega,\mu;\mathbb{R}^m)}$ for $\boldsymbol{\xi} \sim \mathrm{GP}(0, \mathbf{K})$ is normally distributed.*

Proof. Note that the inner product is defined almost surely since

$$\mathbb{E}_{\boldsymbol{\xi}} \left[\|\boldsymbol{\xi}\|^2_{L^2(\Omega,\mu;\mathbb{R}^m)} \right] = \int \mathbb{E}_{\boldsymbol{\xi}}[\|\boldsymbol{\xi}(\theta)\|^2] \mu(\mathrm{d}\theta) = \int \mathrm{tr}\, \mathbf{K}(\theta, \theta) \mu(\mathrm{d}\theta) < \infty.$$

We denote by $\mathscr{E}$ the closed linear span of the set of square-integrable random variables $\{\psi(\theta)^\top \boldsymbol{\xi}(\theta) : \theta \in \Omega\}$. For any $\mathbf{Z} \in \mathscr{E}^\perp$ it holds that $\mathbb{E}_{\boldsymbol{\xi}}[\mathbf{Z}\psi(\theta)^\top \boldsymbol{\xi}(\theta)] = 0$, so that by Fubini's theorem

$$\mathbb{E}_{\boldsymbol{\xi}} \left[\mathbf{Z} \langle \psi, \boldsymbol{\xi} \rangle_{L^2(\Omega,\mu;\mathbb{R}^m)} \right] = \mathbb{E}_{\boldsymbol{\xi}} \left[\int \mathbf{Z} \psi(\theta)^\top \boldsymbol{\xi}(\theta) \mu(\mathrm{d}\theta) \right] = 0.$$

Hence $\langle \psi, \boldsymbol{\xi} \rangle_{L^2(\Omega,\mu;\mathbb{R}^m)} \in (\mathscr{E}^\perp)^\perp = \mathscr{E}$, and so is normally distributed. □

For the proposed perturbation process, the change in the gradient field along the flow of $\boldsymbol{\xi}$ can be quantified as

$$\begin{aligned}
\nabla \frac{\delta F}{\delta \mu}(\mu_{\Delta t}, \theta) - \nabla \frac{\delta F}{\delta \mu}(\mu^\dagger, \theta) &= \int_0^{\Delta t} \partial_t \left[\nabla \frac{\delta F}{\delta \mu}(\mu_t, \theta) \right] \mathrm{d}t \\
&= -\int_0^{\Delta t} \int \nabla_\theta \nabla_{\theta'} \frac{\delta^2 F}{\delta \mu^2}(\mu_t, \theta, \theta') \boldsymbol{\xi}(\theta') \mu_t(\mathrm{d}\theta') \, \mathrm{d}t \\
&= -\int_0^{\Delta t} \mathcal{H}_{\mu_t}[\boldsymbol{\xi}] \, \mathrm{d}t.
\end{aligned}$$

The resulting ψ_0-component is

$$\begin{aligned}
\alpha(\boldsymbol{\xi}) &= \int \psi_0(\theta)^\top \nabla \frac{\delta F}{\delta \mu}(\mu_{\Delta t}, \theta) \mu^\dagger(\mathrm{d}\theta) \\
&= \int \psi_0(\theta)^\top \left(\nabla \frac{\delta F}{\delta \mu}(\mu^\dagger, \theta) - \int_0^{\Delta t} \mathcal{H}_{\mu^\dagger}[\boldsymbol{\xi}] \, \mathrm{d}t + \int_0^{\Delta t} (\mathcal{H}_{\mu^\dagger} - \mathcal{H}_{\mu_t})[\boldsymbol{\xi}] \, \mathrm{d}t \right) \mu^\dagger(\mathrm{d}\theta) \\
&= -\lambda_0 \Delta t \int \psi_0(\theta)^\top \boldsymbol{\xi}(\theta) \mu^\dagger(\mathrm{d}\theta) + \alpha + O(\Delta t^2),
\end{aligned}$$

and first term is normally distributed by Lemma E.9.

Algorithm 1 Mean-field dynamics with birth-death and perturbation

Require: i.i.d. samples $\theta_0^{(1)}, \cdots, \theta_0^{(N)} \sim \mu_0$
  while $\mathcal{L}(\widehat{\mu}_k) > \epsilon$ **do**
    Update all particles as $\theta_{k+1}^{(j)} = \theta_k^{(j)} - \eta \nabla \frac{\delta \mathcal{L}}{\delta \mu}(\widehat{\mu}_k, \theta_k^{(j)})$, $j \in [N]$
    if $\mathcal{L}(\widehat{\mu}_k) - \mathcal{L}(\widehat{\mu}_{k+1}) \le \delta_b$ **then**
      Randomly replace $\lfloor \gamma N \rfloor$ neurons with i.i.d. samples from π
    end if
    if $\mathcal{L}(\widehat{\mu}_k) - \mathcal{L}(\widehat{\mu}_{k+1}) \le \delta_p$ **and** $k - k_p > \tau$ **then**
      $k_p \leftarrow k$
      Generate a Gaussian process $\boldsymbol{\xi} \sim \mathrm{GP}(0, \mathbf{K})$
      Update all particles as $\theta_{k+1}^{(j)} = \theta_k^{(j)} - \eta_p \nabla \boldsymbol{\xi}(\theta_k^{(j)})$, $j \in [N]$
    end if
    $k \leftarrow k + 1$
  end while

Unfortunately this naive approach is not enough to ensure large α, at least in polynomial time, since the eigenfunction ψ_0 and base measure also change along the perturbation. Jin et al. (2017) bypass this issue in finite dimensions via a geometric argument; we conjecture that our method also guarantees polynomial escape time. If this is true, we may combine Proposition 5.1 with $\delta = O(k^{-1}\,\mathcal{L}(\mu_t)^2)$, yielding $O(\frac{k^6}{\gamma^3 t^3})$ convergence away from saddle points, and Theorem 5.5 to conclude that *perturbed* WGF enjoys polynomial convergence to global minima.

Dimensional dependency of Theorem 5.5. The rate is polynomial in the number of features k but only linear in d, mitigating the curse of dimensionality. Initially $\mathcal{L}$ decreases by $\widetilde{\Omega}(k^{-5}d^{-1})$ in time $\widetilde{O}(k)$ when $\mathcal{L} = \Theta(1)$. As training progresses, the rate worsens to $\widetilde{\Omega}(k^{-9}d^{-1})$ in time $\widetilde{O}(k^3)$ when $\mathcal{L} = \Theta(\frac{1}{k})$ due to the smaller curvature of $\mathcal{L}$, until we enter the accelerated phase and Theorem 5.2 takes over. Since $\mathcal{L}$ becomes ill-conditioned if $h_\mu(\boldsymbol{x})$ is nearly constrained on a subspace, we have assumed that $\lambda_{\min}(\boldsymbol{\Sigma}_{\mu,\mu})$ is locally bounded below (on the same order as the upper bound R_1^2/k) to obtain regularity estimates. We expect this to not be a problem in practice since $\mathbf{W}$ will not diverge without timescale separation. In our experiments, $\lambda_{\min}(\boldsymbol{\Sigma}_{\mu,\mu})$ never varied by over 25% during each run.

F Numerical Experiments

F.1 Implementation

We provide a simple summary of the proposed modified mean-field dynamics in Algorithm 1. For the birth-death process, a fraction γ of all neurons are randomly deleted and replaced with samples from π whenever $\mathcal{L}$ does not sufficiently decrease. Here $\theta_k^{(1)}, \cdots, \theta_k^{(N)}$ denote the values of the N particles at step k with empirical distribution $\widehat{\mu}_k = \frac{1}{N}\sum_{j=1}^N \delta_{\theta_k^{(j)}}$, ϵ is the convergence error and δ_b, δ_p are improvement thresholds for applying the birth-death and perturbation procedures, respectively. We also set learning rate η, perturbation step size η_p and a waiting time τ for escaping saddle points. More generally, δ_b, δ_p could be decreased and τ could be increased depending on the current objective value as suggested in Theorem 5.5. In addition, the density ratio $\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}$ could be estimated at certain steps to directly check for the birth-death condition; see Sugiyama et al. (2018) for an overview of applicable methods, especially in high dimensions.

F.2 Experimental Results

Complementing our theoretical analyses, we now explore some empirical aspects of in-context feature learning of a toy Transformer. We compare three models: the *attention* Transformer jointly optimizes the loss $\mathcal{L}_{\mathrm{TF}}(\mu, \mathbf{W})$, while the *static* and *modified* Transformers directly minimize $\mathcal{L}(\mu)$ without passing through the LSA layer. All models are pretrained using SGD on 10K prompts each containing 1K token pairs. For the MLP we set $d = 20$, $k = 5$ with 500 sigmoid neurons and $\mathcal{D}_{\mathscr{X}} \sim \mathcal{N}(0, \mathbf{I}_d)$.

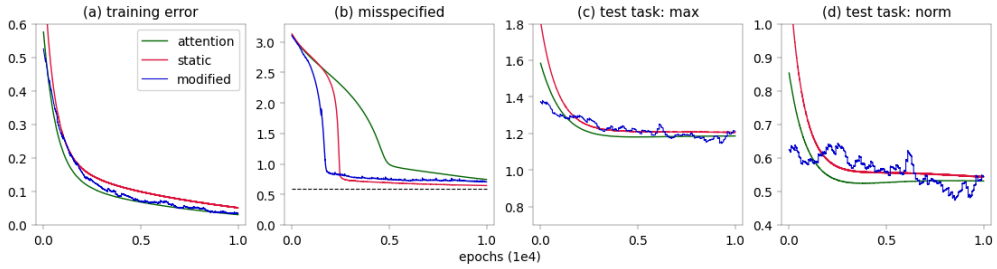

Figure 1: (a) Training error of the attention, static and modified Transformers. (b) Learning a misspecified task containing two extra features. (c) Test error for the nonlinear maximum task $\max_{1\le j\le k} h_{\mu^\circ}(\boldsymbol{x})$; (d) for the norm task $\|h_{\mu^\circ}(\boldsymbol{x})\|$.

The modified model additionally implements the birth-death and perturbation dynamics of Section 5 if $\mathcal{L}$ has not decreased by 1% every 100 epochs.

Figure 1(a) shows that attention and static Transformers exhibit similar dynamics and successfully converge to global optima, justifying the two-timescale approach. Next, Figure 1(b) plots the training curve for a misspecified model where the true features h_{μ° are 7-dimensional. While zero loss is not achievable due to the increased complexity, all models still find a well-behaved minimum, and the modified dynamics escapes a potential saddle point more quickly. Finally, we compute the test loss w.r.t. two nonlinear feature-based tasks $\max_{1\le j\le k} h_{\mu^\circ}(\boldsymbol{x})_j$ and $\|h_{\mu^\circ}(\boldsymbol{x})\|$ in Figures 1(c),(d). Accuracy sharply improves when the relevant features are learned, confirming that ICFL can generalize beyond linear regression even in one-layer Transformers and further demonstrating the importance of feature learning.