Mean-field analysis of polynomial-width two-layer neural network beyond finite time horizon

Margalit Glasgow MGLASGOW@MIT.EDU
Massachusetts Institute of Technology

Denny Wu DENNYWU@NYU.EDU
New York University, Flatiron Institute

Joan Bruna XYZ@SAMPLE.COM
New York University, Flatiron Institute

Editors: Nika Haghtalab and Ankur Moitra

Abstract

We study the approximation gap between the dynamics of a polynomial-width neural network and its infinite-width counterpart, both trained using projected gradient descent in the mean-field scaling regime. We demonstrate how to tightly bound this approximation gap through a differential equation governed by the mean-field dynamics. A key factor influencing the growth of this ODE is the *local Hessian* of each particle, defined as the derivative of the particle's velocity in the mean-field dynamics with respect to its position. We apply our results to the canonical feature learning problem of estimating a well-specified single-index model; we permit the information exponent to be arbitrarily large, leading to convergence times that grow polynomially in the ambient dimension d. We show that, due to a certain "self-concordance" property in these problems — where the local Hessian of a particle is bounded by a constant times the particle's velocity — polynomially many neurons are sufficient to closely approximate the mean-field dynamics throughout training.

1. Introduction

The Mean-field Regime. We consider the training of the following one-hidden-layer neural network with m neurons via gradient-based optimization:

$$f(x) = \frac{1}{m}\sum_{i=1}^{m}\sigma(\langle x, w_i\rangle), \quad w_1, w_2, ..., w_m \in \mathbb{S}^{d-1}, \tag{1}$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is the nonlinear activation function (e.g., ReLU), and $\{w_i\}_{i=1}^m$ are trainable parameters, constrained to the sphere. Due to the nonlinearity of the activation function, the optimization landscape is generally non-convex. In this context, two approaches have been developed to "convexify" the problem through overparameterization (i.e., increasing the network width m) and to establish global optimization guarantees: the *neural tangent kernel* (NTK) [JGH18, DZPS18, AZLS19, ZCZG20] and the *mean-field* analysis [NS17, CB18, MMN18, RVE18, SS20]. The NTK approach linearizes the training dynamics around initialization under appropriate scalings, ensuring that the trainable parameters remain close to their random initialization [COB19]. However, this condition prevents feature learning and often leads to suboptimal statistical rates, as it fails to capture the adaptivity of neural networks [GMMM19, CB20, YH20, BES+22].

The mean-field analysis, on the other hand, lifts (1) into the (infinite-dimensional) space of measures by considering the empirical distribution of neurons $\hat{\rho}^m = m^{-1}\sum_{i=1}^m \delta_{w_i}$. Under certain

regularity conditions, one can establish weak convergence of the empirical distribution to the limiting mean-field measure as the number of neurons tends to infinity: $\hat{\rho}^m \overset{m\to\infty}{\to} \rho^{\mathrm{MF}}$, and the trajectory of the limiting parameter distribution is characterized by a partial differential equation (PDE). This (McKean-Vlasov type) PDE description can capture the nonlinear evolution of the neural network beyond the kernel (lazy) regime.

Studying the mean-field dynamics has several advantages, particularly with regard to learning sparse or low-dimensional target functions such as multi-index models. First, in contrast to the NTK regime, the mean-field dynamics describes feature learning which often leads to improved statistical efficiency (see e.g., [Bac17, CB20, AAM22, MZD+23]). Further, overparameterized neural networks are useful for fitting functions that are not well-specified, for instance a multi-index function with an unknown link function. In such instances, prior correlation loss analyses [AAM23, LOSW24] that ignore the interaction between neurons cannot establish learnability[1]. Second, from a purely analytical perspective, the infinite-width limit allows us to exploit certain problem symmetries that simplify the mean-field PDE into low-dimensional descriptions as done in [AAM22, HC23, ASKL23, CG24, MU25].

Propagation of Chaos. Since training infinite-width networks is computationally infeasible, the practical significance of the above theoretical benefits hinges on having a quantitative connection between finite-width networks and their associated mean-field limit. This is precisely the goal we embark upon in this work. The dynamics of polynomial-width neural networks can be viewed as a finite (interacting) particle discretization of the limiting mean-field PDE. Therefore, one of the main challenges in transferring learning guarantees of the infinite-width limit to the finite-width system lies in the non-asymptotic control of particle discretization error — known as the *propagation of chaos* [Szn91, CD22].

In the context of neural network theory, existing propagation of chaos results typically fall short of delivering this non-asymptotic control. On one hand, *exponential-in-time Grönwall-type* estimates leverage the regularity of the dynamics to propagate the Monte-Carlo error at initialization (at scale $O(1/m)$) to obtain an estimate of the form $\sup_{t\in[0,T]}(f_{\rho_t^{\mathrm{MF}}}(x) - f_{\rho_t^m}(x))^2 \lesssim \exp(T)\cdot(m^{-1}\wedge\eta)$ where $\eta > 0$ is the learning rate [MMN18, MMM19, DBDFS20]. Hence, this type of discretization error analysis is only quantitative when the time horizon is short, such as $T = O_d(1)$ for learning low "leap" functions [AAM22, BMZ23, MU25] and $T = O_d(\log d)$ for learning certain quartic polynomials [MZD+23]. On the other hand, for the *mean-field Langevin dynamics* (MFLD) [HRSS19, NWS22, Chi22a], which introduces additive Gaussian noise to the gradient updates, exponential dependency on time can be removed under a uniform logarithmic Sobolev inequality (LSI), leading to *uniform-in-time propagation of chaos* [CLRW24, SWN23, KZC+24, Nit24]. However, the LSI assumption ultimately transfers the exponential dependency to the runtime [SWON23, WMHC24, MHWE24, TS24]. Finally, [CRBVE20, PN21, Chi22b] proved uniform-in-time fluctuations around the mean-field limit, but in the asymptotic width limit. To our knowledge, the only work that coupled a poly-width network with the infinite-width limit for $\mathrm{poly}(d)$ time is [RZG23], which considered a specific bottleneck architecture for learning a symmetric target function.

Consequently, despite the feature learning advantage, the function class that can be learned by two-layer neural networks trained via gradient descent in the mean-field regime with *polynomial compute* is largely unknown, except for target functions reachable within finite (or at most $\log d$)

1. In Section B, we give several concrete examples of this, along with simulations.

time horizon. It is likely that for many interesting problems, this $T = O_d(\log d)$ horizon is not sufficient for the mean-field dynamics to converge to a low-loss solution. For instance, when the target function is low-dimensional, prior works have shown that gradient-based feature learning often requires $T \gtrsim d^{\Theta(k^*)}$ runtime, where k^* is the *information/leap exponent* (IE) of the link function, which may be arbitrarily large [BAGJ21, AAM23, BBPV23]. The goal of this work is to identify sufficient and verifiable conditions under which the mean-field limit is well-approximated by $m = \text{poly}(d)$ neurons up to $T = \text{poly}(d)$ time horizon.

1.1. Our Contributions

In this work, we study a teacher-student setting where the target function is parameterized by finitely many "teacher" neurons. Let ρ_t^{MF} denote the distribution at time t of the infinite-width mean-field dynamics trained with projected (spherical) gradient flow on infinite data, and ρ_t^m the m-particle mean-field discretization of this dynamics, trained with n samples. We establish a set of conditions under which ρ_t^m is well approximated by ρ_t^{MF} up to the time required to learn the teacher model. The crux of these conditions is twofold:

1. The mean-field dynamics satisfy a certain *local strong convexity* (Assumption LSC), which states that when a neuron is close to a teacher neuron, the local landscape is strongly convex.
2. A certain average *stability* parameter J_{avg} (Assumption Stability) is at most $O(1/T)$, where T is the convergence time. Loosely speaking, J_{avg} is a measure of the average sensitivity of the neurons with respect to a small perturbation in any one neuron.

We show in Theorem 7 that if these conditions hold (along with several other regularity and technical conditions), then for $t \le T$, with high probability one has

$$W_1\left(\rho_t^{\text{MF}}, \rho_t^m\right) \lessapprox \frac{\text{poly}(d,t)}{\min\left(\sqrt{m}, \sqrt{n}\right)}.$$

This means that $\text{poly}(d,T)$ neurons suffice to approximate the mean-field limit up to the time of convergence. This result also gives a non-asymptotic rate of convergence of ρ_t^m to ρ_t^{MF} with time dependence that goes beyond the pessimistic Grönwall estimate.

In Theorem 9, we apply our result to a setting of learning a single-index model (SIM) with high information exponent $k^* \ge 4$, for which gradient flow converges in time $T = \Theta(d^{k^*/2})$. First, we prove that in this setting, the limiting mean-field network, trained on the population loss, can learn the target function at time T. Then we use Theorem 7 to deduce that with $m, n = d^{\Theta(k^*)}$, at time T the difference $W_1(\rho_t^{\text{MF}}, \rho_t^m)$ is small, and thus the finite-width model ρ_t^m also achieves small population loss.

Remark *To our knowledge, our work is the first to prove propagation of chaos (i.e., the above bound on $W_1\left(\rho_t^{\text{MF}}, \rho_t^m\right)$) with polynomially many neurons at timescales longer than $\log(d)$. We remark that we do not believe all the conditions we impose to be necessary – we discuss this in detail in Section B. Existing techniques (see [CD22] for review) primarily leverage either (a) convexity in the system, (b) Grönwall's method, or (c) a large diffusion term. Our techniques go beyond these approaches, and as such they could be useful to establish quantitative propagation of chaos in interacting particle systems with little or no noise.*

Outline. In Section 2, we provide preliminaries on the setting and explain the basic objects we will analyze. In Section 3, we state our main results, as outlined in the contributions. In Section 4, we give an overview of the proofs. We conclude in Section 5. In Appendix A, we provide additional discussion of related work. In Appendices B and C, we discuss the assumptions of our settings, comment on their necessity, and provide simulations. Full proofs are in the following appendices.

Notations. $\mathcal{P}(\Omega)$ denotes the space of probability distributions over Ω. $W_1(\rho, \rho')$ denotes the 1-Wasserstein distance between distributions ρ and ρ'. We will use lower-case letters (f, g, h) to denote functions defined on $\mathbb{S}^{d-1}$, Greek letters (Δ, ξ, etc) to denote vector-valued functions $\mathbb{S}^{d-1} \to \mathbb{R}^d$, and upper-case letters to denote matrix-valued functions $\mathbb{S}^{d-1} \to \mathbb{R}^{d\times d}$ or $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to \mathbb{R}^{d\times d}$. When $\hat{\mu}$ is an empirical measure of the form $\hat{\mu} = \frac{1}{m}\sum_i \delta_{w_i}$, we will use the shorthand $f(i) = f(w_i)$, and denote $\mathbb{E}_i f(i) := \frac{1}{m}\sum_i f(w_i)$. We write $P_w^\perp := (I - ww^\top)$. For $H \in L^2(\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}, \mu^2, \mathbb{R}^{d\times d})$, $D \in L^2(\mathbb{S}^{d-1}, \mu, \mathbb{R}^{d\times d})$ and $\Lambda, \in L^2(\mathbb{S}^{d-1}, \mu, \mathbb{R}^d)$, we use $H\Lambda(w) := \mathbb{E}_{w'\sim\mu} H(w, w')\Lambda(w')$, and $D \odot \Lambda(w) = D(w)\Lambda(w)$. For $f \in L^2(\mathbb{R}^d, \nu)$, we write $\|f\|_\nu^2 := \mathbb{E}_x |f(x)|^2$, and omit the subscript when the context is clear.

Throughout this paper, we use the asymptotic notation $O_C(X)$ to denote X times some constant that depends arbitrarily on C. Whenever a term of the form C (usually with some subscript) appears, this term is referring to a constant, meaning that its value does not depend on m, n, d (which we will take to infinity). We write "with high probability" when the probability approaches 1 as m or n goes to infinity. This probability is taken over the neural network initialization $\{w_i\}_{i\in[m]}$ and the random sample of n data points.

2. Setting and Preliminaries

2.1. Projected Gradient Dynamics on Neural Networks

Consider a neural network to be parameterized by some distribution $\rho \in \mathcal{P}(\mathbb{S}^{d-1})$, such that

$$f_\rho(x) := \mathbb{E}_{w\sim\rho}\sigma(w^\top x),$$

for a link function (activation) σ. We require σ to satisfy the regularity conditions in Regularity assumption.

A supervised regression problem is parameterized by an initial distribution for the network weights, ρ_0, and a distribution $\mathcal{D}$ over points $(x, y) \in \mathbb{R}^d \times \mathbb{R}$. Given $(\rho_0, \mathcal{D})$, we define $f^*(x) = \mathbb{E}_{\mathcal{D}}[y|x]$. We will train the neural network to minimize the squared loss

$$L_{\mathcal{D}}(\rho) := \mathbb{E}_{(x,y)\sim\mathcal{D}}(f_\rho(x) - y)^2 .$$

We study the projected gradient flow dynamics of ρ induced by moving each particle $w \sim \rho$ in the direction of the gradient of the loss $L_{\mathcal{D}}(\rho)$, and then projecting the particle back on the sphere:

$$\tfrac{d}{dt}w = \nu_{\mathcal{D}}(w, \rho) := -(I - ww^\top)\nabla_w f_{\mathcal{D}}(w) + (I - ww^\top)\nabla_w \mathbb{E}_{w'\sim\rho} \hbar_{\mathcal{D}}(w, w') \tag{2}$$

where

$$f_{\mathcal{D}}(w) := \mathbb{E}_{(x,y)\sim\mathcal{D}} y\sigma(w^\top x) \qquad \text{and} \qquad \hbar_{\mathcal{D}}(w, w') := \mathbb{E}_{(x,y)\sim\mathcal{D}} \sigma(w'^\top x)\sigma(w^\top x).$$

In the case where we train on infinite data, the relevant problem parameters are $(f^*, \rho_0, \mathcal{D}_x)$, where $\mathcal{D}_x$ is the x-marginal of $\mathcal{D}$. In such setting, and when $\mathcal{D}_x$ is clear from context, we will use $\nu(w, \rho)$

(without any distribution subscripted) to denote the case where $x \sim \mathcal{D}_x$, $y = f^*(x)$ deterministically. Whenever an expectation over x appears in this paper without explicit distribution, it should be interpreted as over $x \sim \mathcal{D}_x$. In this paper, we will primarily be interested in a teacher-student setting with a ground truth measure ρ^*, such that $f^*(x) = \mathbb{E}_{w^* \sim \rho^*} \sigma(x^\top w^*)$. Thus we will sometimes describe a problem by $(\rho^*, \rho_0, \mathcal{D}_x)$.

2.2. Coupling between Mean Field and Finite-Neuron Dynamics

We will study the evolution of two different learning dynamics in this paper.

Infinite-width, infinite-data *mean-field* dynamics. We denote the mean-field distribution at time t by $\rho_t^{\text{MF}} \in \mathcal{P}(\mathbb{S}^{d-1})$, where we initialize $\rho_0^{\text{MF}} = \rho_0$. Each particle $w \in \mathbb{S}^{d-1}$ in the mean-field dynamics evolves according to the infinite-data velocity $\nu(w, \rho_t^{\text{MF}}) \in T_w \mathbb{S}^{d-1}$. $\xi_t(w) \in \mathbb{S}^{d-1}$ denotes the *characteristic* of a particle initialized at w and evolved under the mean-field dynamics:

$$\tfrac{d}{dt}\xi_t(w) = \nu(\xi_t(w), \rho_t^{\text{MF}}) \qquad \xi_0(w_i) = w_i \, .$$

This dynamics can also be expressed though the *continuity equation:* $\frac{d}{dt}\rho_t^{\text{MF}} = \nabla \cdot (\nu(w, \rho_t^{\text{MF}})\rho_t^{\text{MF}})$.

Finite-width, finite-data dynamics. Let ρ_t^m denote the empirical measure defined by m *neurons* under the projected gradient flow induced by the *empirical loss* from n training samples. Let $\hat{\mathcal{D}}$ denote the empirical distribution of the n training samples. We initialize $\rho_0^m = \frac{1}{m}\sum_{i=1}^m \delta_{w_i}$, where $w_i \sim \rho_0$ i.i.d. for each $i \in [m]$. Each particle $w \in \mathbb{S}^{d-1}$ in the finite dynamics evolves according to the empirical velocity $\nu_{\hat{\mathcal{D}}}(w, \rho_t^m)$. This defines an ODE in $(\mathbb{S}^{d-1})^{\otimes m}$, whose characteristics are now denoted by $\hat{\xi}_t(w_i)$, and solve

$$\tfrac{d}{dt}\hat{\xi}_t(w_i) = \nu_{\hat{\mathcal{D}}}(\hat{\xi}_t(w_i), \rho_t^m) \qquad \hat{\xi}_0(w_i) = w_i \, , \; i \in [m] \, .$$

We will study the setting where the training data are drawn i.i.d. from a sub-Gaussian distribution with sub-Gaussian label noise (See Regularity Assumption **R2**).

Coupling the dynamics. Let $\bar{\rho}_t^m$ be the distribution initialized at ρ_0^m, but that evolves according to the dynamics $\nu(\cdot, \rho_t^{\text{MF}})$. That is, $\bar{\rho}_t^m = \frac{1}{m}\sum_{i=1}^m \delta_{\xi_t(w_i)}$. Note that $\bar{\rho}_t^m$ is equivalent in distribution to a random sample of m particles drawn iid from ρ_t^{MF}. Define the coupling error at neuron w_i as

$$\Delta_t(i) := \hat{\xi}_t(w_i) - \xi_t(w_i) \in \mathbb{R}^d \, , \; i \in [m] \, ,$$

such that $\Delta_0(i) = 0$ for all i. Now by definition, $W_1(\rho_t^m, \bar{\rho}_t^m) \le \mathbb{E}_i \|\Delta_t(i)\|$; thus it is easy to show that $\mathbb{E}_i\|\Delta_t(i)\|$ gives a good bound on the function-error distance between ρ_t^{MF} and ρ_t^m:

Lemma 1 *Suppose Regularity Assumption **R1** holds. With high probability over the draw ρ_0^m, we have*

$$\|f_{\rho_t^{\text{MF}}} - f_{\rho_t^m}\|^2 \le 2C_{reg}^2 \left(\mathbb{E}_i\|\Delta_t(i)\|\right)^2 + \frac{\log(m)}{m}.$$

Here, and throughout, we use the notation $\|f - g\|^2$ to denote $\mathbb{E}_{x \sim \mathcal{D}_x}[(f(x) - g(x))^2]$.

2.3. Description of the Dynamics of Δ

The main result of this section is Lemma 5, which gives a first-order approximation of the dynamics of $\Delta_t(i)$. The quantities $\{\Delta_t(i)\}_i$ evolve via their own particle interaction system, governed by two main terms: a self-interaction term, and an interaction term. The self-interaction term is described by what we call the *local Hessian,* the derivative of a particle's velocity with respect to that particle's position.

Definition 2 (Local Hessian) *The* local Hessian $D_t^\perp : \mathbb{S}^{d-1} \to \mathbb{R}^{d\times d}$ *of neuron* w *at time* t *is*

$$D_t^\perp(w) := \left(\nabla_{\xi_t(w)}\nu(\xi_t(w), \rho_t^{\mathrm{MF}})\right)(I - \xi_t(w)\xi_t(w)^\top).$$

We will also use the abbreviated notation $D_t^\perp(i) := D_t^\perp(w_i)$.

Remark 3 *We call this the local Hessian because it equals the* negative *Hessian of the landscape of the map* $\xi_t(w_i) \to U_t(\xi_t(w_i)) := U(\xi_t(w_i); \rho_t^{\mathrm{MF}})$*, where* $U = \frac{\delta L}{\delta \rho}$ *is the first-variation of the loss, so that* $V = \nabla U$*, and* $\xi_t(w_i)$ *is restricted to the manifold* $\mathbb{S}^{d-1}$*. Thus if the local landscape* $U_t(\xi_t(w_i))$ *is convex on* $\mathbb{S}^{d-1}$*, then* $D_t^\perp(i)$ *is negative semi-definite.*

The part of the dynamics driven by the other $\Delta_t(j)$ is described by what we term the *interaction Hessian,* the (rescaled) derivative of a particle's velocity with respect to the other particles' position.

Definition 4 (Interaction Hessian) *Define the* interaction Hessian $H_t^\perp : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to \mathbb{R}^{d\times d}$ *by*

$$H_t^\perp(w, w') := \left(I - \xi_t(w)\xi_t(w)^\top\right)\nabla_{\xi_t(w')}\nabla_{\xi_t(w)}\hbar(\xi_t(w), \xi_t(w'))\left(I - \xi_t(w')\xi_t(w')^\top\right),$$

We will also use the abbreviated notation $H_t^\perp(i,j) := H_t^\perp(w_i, w_j)$.

Fact 1 *For any* w, w'*,* $H_t^\perp(w, w')$ *is a positive semi-definite kernel.*

Proof By definition of $\hbar_{\mathcal{D}}$ in Equation 2.1, one can check that $H_t^\perp(w, w') = \mathbb{E}_x \phi_x(w)\phi_x(w')^\top$, where we define the feature map $\phi_x(w) := (I - \xi_t(w)\xi_t(w)^\top)\sigma'(\xi_t(w)^\top x)x$ ■

We make the following basic regularity assumptions on the activation function and the data.

Assumption Regularity (Regularity Assumptions)

- ***R1*** *For a constant* C_{reg}*, the activation* σ *satisfies that for* $j = 0, 1, 2, 3$ *and any subGaussian variable* X*, we have* $\mathbb{E}_X|\sigma^{(j)}(X)|^5 \le (C_{reg}/11)^5$*, where* $\sigma^{(j)}$ *denotes the* j*th derivative of* σ*.*
- ***R2*** *The distribution* $\mathcal{D}_x$ *on the covariates is subGaussian, and the noise has covariance at most* 1*, that is* $\mathbb{E}_{y\sim\mathcal{D}|x}(y - f^*(x))^2 \le 1$*.*

We introduce the control parameters

$$\epsilon_m := \frac{d^{3/2}\log(mT)}{\sqrt{m}}, \qquad \epsilon_n := \frac{\sqrt{d}\log^2(n)}{\sqrt{n}}.$$

We will show in Lemma 23 that with high probability, the error $\|\nu(\xi_t(w_i), \rho_t^{\mathrm{MF}}) - \nu(\xi_t(w_i), \bar{\rho}_t^m)\|$ due to sampling only m neurons is uniformly (over i and t) bounded by ϵ_m. Similarly, we will show in Lemma 27 that the error $\|\nu_{\hat{\mathcal{D}}}(\hat{\xi}_t(w_i), \rho_t^m) - \nu(\hat{\xi}_t(w_i), \rho_t^m)\|$ due to using the empirical data distribution $\mathcal{D}$ is uniformly bounded by ϵ_n.

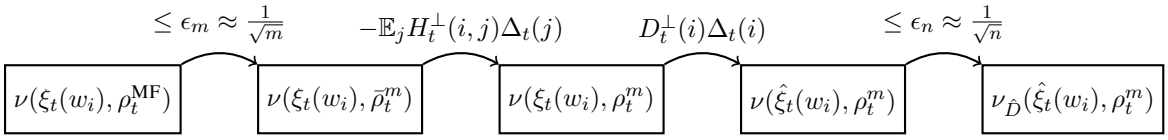


Figure 1: Decomposing $\frac{d}{dt}\Delta_t(i) = \nu(\xi_t(w_i), \rho_t^{\mathrm{MF}}) - \nu_{\hat{D}}(\hat{\xi}_t(w_i), \rho_t^m)$. The approximate differences between the terms in the rectangles are given above the arrows.

Lemma 5 (Parameter-Space Error Dynamics) *Suppose Regularity Assumption holds. With high probability, for all $t \leq T$ and $i \in [m]$,*

$$\frac{d}{dt}\Delta_t(i) = D_t^{\perp}(i)\Delta_t(i) - \mathbb{E}_{j\sim[m]} H_t^{\perp}(i,j)\Delta_t(j) + \boldsymbol{\epsilon}_{t,i},$$

where $\|\boldsymbol{\epsilon}_{t,i}\| \leq 2\epsilon_m + \epsilon_n + 2C_{reg}\left(\|\Delta_t(i)\|^2 + \mathbb{E}_j\|\Delta_t(j)\|^2\right)$.

We prove Lemma 5 by decomposing $\frac{d}{dt}\Delta_t(i) = \nu(\xi_t(w_i), \rho_t^{\mathrm{MF}}) - \nu(\hat{\xi}_t(w_i), \rho_t^m)$ into four differences (see Figure 1), and separating the first order terms (in Δ_t) from higher order terms in these differences.

An integral form for $\Delta_t(i)$. Duhamel's principle gives us a way to solve the ODE in Lemma 5 using the solution to a simpler dynamics which only involves the local Hessian.

Definition 6 (Local Stability Matrix) *Define $J_{t,s}^{\perp}(w)$ to be the matrix that solves*

$$\tfrac{d}{dt}J_{t,s}^{\perp}(w) = D_t^{\perp}(w)J_{t,s}^{\perp}(w); \qquad J_{s,s}^{\perp}(w) = (I - \xi_s(w)\xi_s(w)^{\top}).$$

We call this the local stability matrix, because $J_{t,s}^{\perp}(w) = \mathbf{J}_{\xi_{t,s}}(\xi_s(w))$, where $\xi_{t,s}(u)$ denotes the position of a neuron at time t which evolves in the mean field dynamics starting at position u at time s, and $\mathbf{J}$ denotes the Jacobian. We use the shorthand $J_{t,s}(i) := J_{t,s}(w_i)$.

On the same assumptions as Lemma 5, Duhamel's principle yields

$$\Delta_t(i) = \int_0^t J_{t,s}^{\perp}(i)\left(-\mathbb{E}_j H_s^{\perp}(i,j)\Delta_s(j) + \boldsymbol{\epsilon_{s,i}}\right) ds. \tag{3}$$

3. Main Result: Propagation of Chaos

3.1. Intuition and Key Challenges

To bound $W_1(\rho_t^m, \rho_t^{\mathrm{MF}})$, it suffices to analyze the dynamics of Δ_t given by the ODE in Lemma 5:

$$\tfrac{d}{dt}\Delta_t(i) = D_t^{\perp}(i)\Delta_t(i) - \mathbb{E}_{j\sim[m]} H_t^{\perp}(i,j)\Delta_t(j) + \boldsymbol{\epsilon}_{t,i} \qquad \|\boldsymbol{\epsilon}_{t,i}\| \leq \epsilon. \tag{4}$$

One might hope to leverage the linearity of (4) to solve this ODE in closed form, but unfortunately, the time-dependent coefficient matrix, $\mathrm{diag}(D_t^{\perp}) - H_t^{\perp}$, does not commute at different times t.

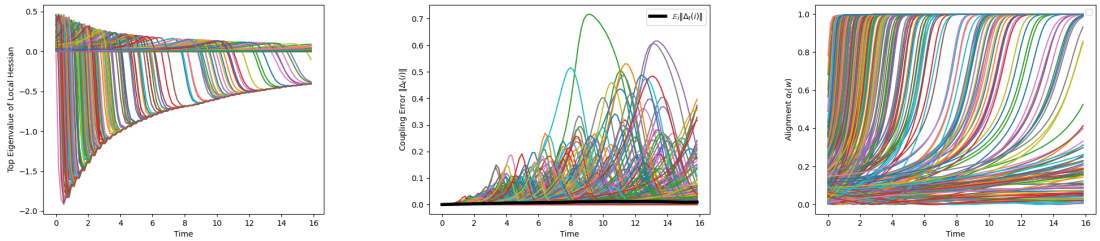


Figure 2: Non-Uniform Dynamics in SIM with IE 4 ($f^*(x) = \mathrm{He}_4(x^\top w^*)$ for $x \in \mathbb{R}^{32}$. We plot $D_t^\perp(i), \|\Delta_t(i)\|, \alpha_t(w_i) = |w^* \xi_t(w_i)|$ for each neuron. Left: Top eigenvalue of the local hessians $D_t^\perp(i)$. Center: $\|\Delta_t(i)\|$, Right: Alignment $\alpha_t(w_i)$ with the teacher neuron. A key challenge in the IE > 2 setting is the variance in Lipschitzness among the different neurons, and in $\|\Delta_t(i)\|$.

Going Beyond Grönwall. The conventional approach (see e.g., [MMN18, MMM19]), uses the maximum Lipschitzness of $\nu(w, \rho)$ – in our spherical case, this translates to a bound on $\sup_{i,j,t} \|D_t^\perp\|, \|H_t^\perp(i,j)\|$ – to bound the RHS of (4) as

$$\frac{d}{dt}\|\Delta_t(i)\| \le 2\,\mathrm{Lip}_{\max} \sup_{j\in[m]} \|\Delta_t(j)\| + \epsilon. \tag{5}$$

In standard settings, this maximum Lipschitzness is a constant, so this method can achieve no better than the bound $W_1(\rho_t^m, \rho_t^{\mathrm{MF}}) \le \exp(\Theta(t))\epsilon$. The work of [MZD+23] goes further to bound (5) using a tight time-dependent Lipschitz constant, yielding propagation of chaos for $\log(d)$ time. However, for problems with polynomial-in-d time to converge, such as learning a SIM with a high information exponent, the approach in (5) is overly pessimistic, because both the local Lipschitzness at neuron i, and the $\|\Delta_t(j)\|$ are extremely non-uniform in i and j (See Figure 2).

Equation (3) gives us an alternative way to approach (4) which can leverage the non-uniform Lipschitzness. Ignoring for a moment the interaction terms in Equation (3), we have $\|\Delta_t(i)\| \approx \int_0^t J_{t,s}^\perp(i)\boldsymbol{\epsilon_{s,i}} ds$, where we recall that the perturbation matrix $J_{t,s}^\perp(i)$ measures of the stability of $\xi_t(w)$ with respect to perturbations at time s. Naively, $J_{t,s}^\perp(w_i)$ appears to grow at an exponential rate whenever the local landscape of the linearized loss around $\xi_t(w_i)$ (see Remark 3) is non-convex.

A key observation of our work is that when w_i escapes certain higher-order saddles, $\|J_{t,s}^\perp(i)\|$ will be bounded polynomially in $t - s$. We achieve this by showing a certain *self-concordance*-like property which upper bounds $D_t^\perp(i)$ using the velocity (which is small near the saddle). Thus one part of our assumptions will be a worst-case polynomial bound on $\|J_{t,s}^\perp(w)\|$ (see Stability Assumption).

The Interaction Term: A Blessing and a Curse. At first glance, the presence of the PSD interaction term $H_t^\perp$ in (4) seems like it can only help us bound $\mathbb{E}_i\|\Delta_t(i)\|$. Indeed, if we ignore the local $D_t^\perp$ terms in the ODE, we would have that $\frac{d}{dt}\Delta_t = -H_t^\perp \Delta_t$, and thus we could show that $\mathbb{E}_i\|\Delta_t(i)\|^2$, an upper bound on the Wasserstein-2 distance $W_2(\rho_t^m, \rho_t^{\mathrm{MF}})$, is non-increasing.

However, the interaction of $H_t^\perp$ and $D_t^\perp$ can lead to precarious situations if the neurons move at non-uniform rates. To see this possibility, suppose for some neuron w_i, $\Delta_t(i)$ first grows by a polynomial factor due to $D_t^\perp(i)$, and then propagates that error, via the interaction term, to a different neuron w_j. Later on, when neuron w_j escapes the saddle, it will grow $\Delta_t(j)$ by a polynomial factor. The process can then continue by "passing off" the error between neurons such that it grows in an exponential fashion, without any neuron doing more than "polynomial growth" of the error itself.

To rule out such a scenario, we will impose an assumption that leverages the intuition that in many teacher-student settings with uniform initialization, the neurons are dispersed before converging to the teacher neurons. Thus on average, the interaction term – whose scale is dictated by inner product $w_i^\top w_j$ – is small, and cannot propagate too much error to these neurons. Specifically, the interaction term drives changes in the error according to the interaction Hessian, $H_t^\perp$: an error of $\Delta_t(j)$ at neuron w_j causes a force of $-H_t^\perp(i,j)\Delta_t(j)$ on the error of neuron w_i. Following Equation (3), this force propagates into an error of scale $R_{t,s}(i,j)\Delta_s(j)$ on neuron w_i at time t, where $R_{t,s}(i,j) := J_{t,s}^\perp(i)H_s^\perp(i,j)$.

The second part of Assumption Stability states that the *average* of $R_{t,s}(i,j)$, over all neurons i far from $\mathrm{supp}(\rho^*)$, is small.

Behavior Near the Teacher Neurons. While the second part of Assumption Stability is quite powerful, we cannot hope that it holds for neurons near the teacher neurons. Indeed, when i and j are both near some $w^* \in \mathrm{supp}(\rho^*)$, then $\|R_{t,t}(i,j)\| = \|H_t^\perp(i,j)\| = \Omega(1)$. Thus for neurons near $\mathrm{supp}(\rho^*)$, we will need to leverage the fact that $H_t^\perp$ is PSD. A key contribution of our work is constructing a novel potential function which can leverage this term. We discuss this at length in Section 4.

3.2. Theorem Statement

We will now present an informal version of our assumptions and propagation of chaos result. Due to the technicality of some of the assumptions, we defer some full statements to Section B. Define

$$B_\tau := \{w \in \mathbb{S}^{d-1} : \exists w^* \in \mathrm{supp}(\rho^*) : \|w^* - w\| \le \tau\}.$$

The following key assumption gives average and worst-case bounds on some of the stability parameters of the MF dynamics.

Assumption Stability (Worst-Case and Average Stability) *Suppose that we have*

$$J_{max} := \sup_{s\le t\le T, w\in\mathbb{S}^{d-1}} \left(\|J_{t,s}^\perp(w)\|, \mathbb{E}_{w\sim\rho_0}\|J_{t,s}^\perp(w)\|^2\right) \le \mathrm{poly}(d,T).$$

Further suppose that for all $\tau > 0$, and given a target horizon $T > 0$,

$$J_{avg}(\tau) := \sup_{s\le t\le T, w',v\in\mathbb{S}^{d-1}} \mathbb{E}_{w\sim\rho_0}\|J_{t,s}^\perp(w)H_s^\perp(w,w')v\|\mathbf{1}(\xi_t(w)\notin B_\tau) \le \frac{\mathrm{poly}(1/\tau)}{T}.$$

Next, we will state our local strong convexity assumption. We remark that such an assumption can only hold when ρ^* is atomic (see Remark 8, and additional comments in Section B.4).

Assumption LSC (abbv) (Local Strong Convexity (Abbreviated; see Assumption LSC)) *We have* (C_{LSC}, τ) locally strongly convex *up to time T, meaning that for any $t \le T$, for any w with $\xi_t(w) \in B_\tau$, we have*

$$D_t^\perp(w) \preceq -C_{\mathrm{LSC}}P_{\xi_t(w)}^\perp\|f_{\rho_t^{\mathrm{MF}}} - f^*\|.$$

Both Stability and LSC (abbv) assumptions are verifiable via solving the deterministic mean-field dynamics ρ_t^{MF}. For technical reasons, our result requires two additional conditions. First, our theorem depends on the rank of the interaction Hessian as $\rho_t^{\mathrm{MF}} \to \rho^*$ being a constant independent

of the ambient dimension d. This rank can be bounded by the following parameter, which will appear in our main theorem:

$$C_{\rho^*} := \min\left(|\mathrm{supp}(\rho^*)|, \dim(\mathrm{supp}(\rho^*))^{2\,\mathrm{degree}(\sigma)+1}\right). \tag{6}$$

Here $\mathrm{degree}(\sigma)$ is the degree of the polynomial σ (or ∞ if σ is not a polynomial). We do not expect such an assumption to be critical; see Section B.7.

Second, we require a technical symmetry condition stated in Assumption Symmetry (in Section B). Loosely, this requires that the atomic set $\mathrm{supp}(\rho^*)$ is *transitive* with respect to the group of rotational symmetries that describe the problem. We remark that such an assumption still covers many non-trivial problems, for instance, learning two teacher neurons in non-orthogonal positions, many neurons in orthogonal positions, or a ring of evenly spaced neurons in a circle. See Section B.6 for further discussion.

We are now ready to state the main theorem.

Theorem 7 (Propagation of Chaos) *Assume that Regularity, Stability, LSC and Symmetry hold up to time T (if relevant). Let C be a constant depending on $C_{\mathrm{LSC}}, \tau, C_{\rho^*}$ and $\delta_T := \|f_{\rho_T^{\mathrm{MF}}} - f^*\|$. Suppose n, m are large enough such that $J_{max}^4 T^3(\epsilon_n + \epsilon_m) \le 1/C$. Then with high probability over the draw ρ_0^m, for all $t \le T$,*

$$\|f_{\rho_t^{\mathrm{MF}}} - f_{\rho_t^m}\|^2 \le 2(W_1(\rho_t^m, \rho_t^{\mathrm{MF}}))^2 + \frac{2C_{reg}\log(m)}{m} \le (CJ_{max}t(\epsilon_m + \epsilon_n))^2.$$

where $\epsilon_m = \frac{\log(mT)\max(d^{1/2}J_{max}, d^{3/2})}{\sqrt{m}}$ *and* $\epsilon_n = \frac{\sqrt{d}\log^2(n)}{\sqrt{n}}$.

Theorem 7 follows directly from Lemma 1 and Corollary 40 in Section F. In Theorem 9, we will apply this theorem to the example of learning a single-index function with high information exponent which takes $T = \mathrm{poly}(d)$ time to learn.

Remark 8 (Local Strong Convexity) *Our local strong convexity is similar to assumptions appearing in prior mean-field analyses [Chi22b, Assumption A5][CRBVE20, Lemma D.9]. In comparison to these works, our assumption is* stronger *in that we require it for all t, not just as $t \to \infty$; this is necessary for our non-asymptotic analysis. However, our assumption is also* weaker *in that we allow the strong convexity parameter to depend on the loss, similarly to the notion of one-point strong convexity (see e.g., [SYS21]). Attaining the stronger non-loss-dependent strong convexity requires a strongly convex regularization term.*

In problems where the mean-field dynamics converge to ρ^*, our local strong convexity condition enforces that when a neuron w_t is close a teacher neuron $w^* \in \mathrm{supp}(\rho^*)$, it will be attracted to w^* and thus any small perturbations are dampened. Local strong convexity can only *hold when ρ^* is atomic. Similar properties have been shown for various sparse optimization problem over measures [FDGW21, PKP23].*

3.3. Application to Single-index Model with High Information Exponent

We now study the setting of learning a well-specified single-index function $f^*(x) = \sigma(x^\top w^*)$, where $w^* \in \mathbb{S}^{d-1}$, and $\sigma(z) = \sum_{k=k^*}^K c_k \mathrm{He}_k(z)$, where (a) $k^* \ge 4$, and $\frac{1}{C_{\mathrm{SIM}}} \le c_{k^*} \le C_{\mathrm{SIM}} \max_k c_k$, (b) for all k, $c_k \ge 0$, and (c) σ is an even function[2]. We restrict to the case when $k^* > 2$ because the

2. If σ is not even, the loss may not go to zero, since $1/2$ of the neurons may be stuck on the side of the equator with $w^\top w^* < 0$.

analysis for $k^* = 2$ has notable differences – namely, the escape times of the neurons is no longer non-uniform as in Figure 2(right); see Remark 17 for further comments. We assume the initial distribution ρ_0 of the neurons is uniform on $\mathbb{S}^{d-1}$, and the data is drawn i.i.d from the distribution $\mathcal{D}$, which has Gaussian covariates, and subGaussian label noise: that is,

$$x \sim \mathcal{N}(0, I_d), \qquad y = f^*(x) + \zeta(x); \qquad \mathbb{E}[\zeta(x)] = 0, \quad \mathbb{E}[\zeta(x)^2] \le 1.$$

Theorem 9 (PoC in Single-Index Model) *Fix any $\delta > 0$, and suppose d is large enough in terms of δ, C_{SIM} and K. Let $T(\delta) := \arg\min\{t : \|f_{\rho_t^{\mathrm{MF}}} - f^*\|^2 \le \delta^2\}$. Then $T(\delta) = O_{K,C_{SIM}}(\sqrt{d}^{k^*-2}\delta^{-(k^*-1)})$. If $n \ge d^{11k^*}$ and $m \ge d^{13k^*}$, then with high probability, for all $t \le T(\delta)$,*

$$\|f_{\rho_t^{\mathrm{MF}}} - f_{\rho_t^m}\|^2 \le \frac{O_{K,\delta}(d^{3k^*})}{\min(\sqrt{m}, \sqrt{n})} \le 3\delta^2 .$$

Remark *The above theorem provides, to the best of our knowledge, the first polynomial-width learning guarantee for one-hidden-layer neural network in the mean-field regime that holds for polynomial-in-d time horizon. When $\mathrm{degree}(\sigma) \gg k^*$, our result demonstrates the statistical advantage of the mean-field parameterization over the lazy/NTK alternative; specifically, under the NTK parameterization, when the width m is sufficiently large, the sample complexity of gradient descent training on the empirical risk must scales as $n \gtrsim d^{\Theta(\mathrm{degree}(\sigma))}$ [GMMM21], whereas the mean-field scaling only requires $n \gtrsim d^{\Theta(k^*)}$ samples.*

4. Overview of Proof Ideas

4.1. Potential-Based Analysis to Prove Theorem 7

We introduce a potential function of Δ_t which dominates $W_1(\rho_t^m, \rho_t^{\mathrm{MF}})$. Building upon the observations from Section 3.1, we design this potential function to have the following three properties:

- **P1** When many neurons are near the teacher neurons, the dynamics due to the interaction hessian $H_t^{\perp}$ should cause the potential to decrease.
- **P2** When a neuron w_i is in a locally convex region ($D_t^{\perp}(i) \preceq 0$), the dynamics due to the local Hessian at w_i should decrease the potential.
- **P3** The change in potential due to a perturbation of Δ should be bounded proportionally to the *average* change over the Δ_i.

A natural choice of potential function would be $\mathbb{E}_i\|\Delta_t(i)\|^2$ (which upper bounds $W_2(\rho_t^m, \rho_t^{\mathrm{MF}})$) because when $\rho_t^{\mathrm{MF}} \approx \rho^*$, $D_t(i)$ are negative definite so

$$\frac{d}{dt}\mathbb{E}_i\|\Delta_t(i)\|^2 \approx -\Delta_t^{\top} H_t^{\perp}\Delta_t - 2\mathbb{E}_i\Delta_t(i)^{\top} D_t(i)\Delta_t(i) \le 0.$$

However, such a function does not satisfy **P3** whenever there is a lot of variance among the $\|\Delta_t(i)\|$.

To achieve **P3**, intuitively, the potential should behave more like $W_1(\rho_t^m, \rho_t^{\mathrm{MF}})$ than $W_2(\rho_t^m, \rho_t^{\mathrm{MF}})$, making $\mathbb{E}_i\|\Delta_t(i)\|$ another natural choice. Unfortunately, this alone does not work as potential function, because even when all neurons have converged to the support of ρ^*, it may *increase* under the

dynamics from the interaction Hessian[3]. As an example, consider the case where $\rho^* = \delta_{w^*}$, and thus near convergence, $H_t^\perp \approx \mathbf{1}\mathbf{1}^\top \otimes P_{w^*}^\perp$, where $\mathbf{1} \in \{\mathbb{S}^{d-1} \to \mathbb{R}\}$ sends all inputs to 1; then if Δ_t is very "imbalanced" (in the sense that $H_t^\perp \Delta_t = \mathbb{E}_i \Delta_t(i)$ is large), we may have $\frac{d}{dt}\mathbb{E}_i\|\Delta_t(i)\| > 0$. For instance suppose $\Delta_t(i) = u$ for a p fraction of the neurons, and $\Delta_t(i) = 0$ for the remaining neurons. Then $\frac{d}{dt}\mathbb{E}_i\|\Delta_t(i)\| = -p + (1-p) > 0$ for $p < 0.5$. To counteract the increase in $\mathbb{E}_i\|\Delta_t(i)\|$, we need to include in the potential function a term which decreases whenever Δ_t is very imbalanced, yet it retains a flavor of an ℓ_1 norm. In order to tame the interactions, such a term should naturally take into account the eigendecomposition of $H_t^\perp$. To construct such a potential function, we will instead consider the eigendecomposition of the map $H_\infty^\perp$ (defined explicitly in Defintion 10), which closely approximates $H_t^\perp$ on neurons in B_τ and avoids tracking the temporal evolution of the eigendecomposition. This ultimately lets us leverage the PSD structure of $H_t^\perp$.

Definition 10 *Define*

$$H_\infty^\perp(w, w') = P^\perp_{\xi^\infty(w)} \nabla_{\xi^\infty(w')} \nabla_{\xi^\infty(w)} K(\xi^\infty(w), \xi^\infty(w')) P^\perp_{\xi^\infty(w')},$$

where $\xi^\infty(w) := \operatorname{argmin}_{w^* \in \operatorname{supp}(\rho^*)} \|\xi_T(w) - w^*\|$ *and we break ties in the argmin arbitrarily.*

Let $\mathcal{Z} := L^2(\mathbb{S}^{d-1}, \rho_0; \mathbb{R}^d)$ be the Hilbert space with the dot product $\langle f, g\rangle_{\mathcal{Z}} := \mathbb{E}_{w\sim\rho_0} f(w)^\top g(w)$. Define the action of $H : (\mathbb{S}^{d-1})^{\otimes 2} \to \mathbb{R}^{d\times d}$ on $\mathcal{Z}$ as $v \mapsto \overline{H}v(w) := \mathbb{E}_{w'\sim\rho_0} H(w, w')v(w')$. In Section F.2.2, we verify that $\overline{H_\infty^\perp}$ is well defined, self-adjoint, and due to the atomic nature of ρ^*, the span of $\overline{H_\infty^\perp}$ is has some finite dimension J. Therefore, $\overline{H_\infty^\perp}$ admits a spectral decomposition in $\mathcal{Z}$ in terms of an orthonormal basis $\{\varphi_j\}_{j\le J}$:

$$\overline{H_\infty^\perp} = \sum_{j\le J} \lambda_j \varphi_j \otimes \varphi_j \ , \ \lambda_j \in \mathbb{R} \ , \ \varphi_j \in \mathcal{Z} \ , \tag{7}$$

such that $\|\overline{H_\infty^\perp}\|_* := \sum_j |\lambda_j| < \infty$. Note that one can have multiplicities in this spectral decomposition. For that purpose, denote by $\Lambda = \{\lambda_j; j \le J\}$ the support of the spectrum. For each $\lambda \in \Lambda$, we denote by V_λ the subspace spanned by $\{\varphi_j; \lambda_j = \lambda\}$, and let P_λ be the orthogonal projector onto that space.

Definition 11 (Balanced Spectral Decomposition of $H_\infty^\perp$ (BSD)) *We say that the spectral decomposition (7) is* C_b-balanced *if, for all $\lambda \in \Lambda$, there exists an orthonormal basis $\mathcal{B}_\lambda$ of V_λ, and some $\eta_\lambda > 0$ such that for all $w \in \mathbb{S}^{d-1}$, $\sum_{v\in\mathcal{B}_\lambda} v(w)v(w)^\top \preceq \eta_\lambda^2 I_d$, and $\sum_{\lambda\in\Lambda} \eta_\lambda^2 \le C_b$. We denote by $\mathcal{Q} := \{(\mathcal{B}_\lambda, \eta_\lambda)\}_{\lambda\in\Lambda}$ the resulting set of eigenfunctions and constants.*

Now, for any $v \in \mathcal{Z}$ and $\Delta \in (\mathbb{R}^d)^{\otimes m}$, we define $\phi_v(\Delta) := |\mathbb{E}_i v(w_i)^\top \Delta(i)|$, and

$$\Psi_{\mathcal{Q}}(\Delta) := \sum_{\lambda\in\Lambda} \eta_\lambda \left(\sum_{v\in\mathcal{B}_\lambda} \phi_v(\Delta)^2\right)^{1/2},$$

Finally, our potential function is

$$\Phi_{\mathcal{Q}}(\Delta) := \Omega(\Delta) + \Psi(\Delta),$$

with $\Omega(\Delta) = \mathbb{E}_i\|\Delta(i)\|$. When the context is clear, we will write $\Phi_{\mathcal{Q}}(t) = \Phi_{\mathcal{Q}}(\Delta_t)$.

When the context is clear, we will write $\Phi_{\mathcal{Q}}(t) = \Phi_{\mathcal{Q}}(\Delta_t)$.

3. Using $W_1(\rho_t^m, \rho_t^{\mathrm{MF}})$ alone (instead of $\mathbb{E}_i\|\Delta_t(i)\|$) fails for the same reason.

Lemma 12 (Balanced Spectral Decomposition) *Suppose Assumption Symmetry holds. Then there exists an spectral distribution $\mathcal{Q}$ which is $C_{\rho^*} = \min\left(|\mathrm{supp}(\rho^*)|, \dim(\mathrm{supp}(\rho^*))^{\mathrm{degree}(\sigma)}\right)$-balanced.*

The next three lemmas show that the potential function $\Phi_{\mathcal{Q}}$ has the desired properties **P1**-**P3**.

Lemma 13 (Descent with Respect to Interaction Term) *Let $\Phi_{\mathcal{Q}}(t)$ be as defined above, where $\mathcal{Q}$ is a C_b-balanced spectral decomposition of $H_\infty^\perp$. Then for any $\tau > 0$ for which the concentration event of Lemma 25 holds for $S = B_\tau$, we have*

$$\langle \nabla \Phi_{\mathcal{Q}}(t), -H_t^\perp \Delta_t \rangle \le (1 + C_b)\mathbb{E}_i \|\mathbb{E}_j H_t^\perp(i,j)\Delta_t(j)\| \mathbf{1}(\xi_t(w_i) \notin B_\tau) + \mathcal{E}_{13},$$

where $\mathcal{E}_{13} = O_{C_{reg}, C_b}(\mathbb{E}_i \|\Delta_t(i)\| \mathbf{1}(\xi_t(w_i) \notin B_\tau) + (\tau + C_b \epsilon_m^{25})\Omega(t))$.

Lemma 14 (Descent with Respect to Local Term) *Suppose Assumption LSC holds with (C_{LSC}, τ). Let $\mathcal{Q}$ be a C_b-balanced spectral distribution. Then with $C_{14} = O_{C_{reg}, C_b}(1)$, we have*

$$\langle \nabla \Phi_{\mathcal{Q}}(t), D_t^\perp \odot \Delta_t \rangle \le -\left(\frac{c\sqrt{L_{\mathcal{D}}(\rho_t^{\mathrm{MF}})}}{2} - C_{14}\tau\right)\Phi_{\mathcal{Q}}(t) + C_{14}\mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i) \notin B_\tau) + C_b \mathbb{E}_i \|\Delta_t(i)\|^2.$$

Lemma 15 (L1 Perturbation Lemma) *Let $\mathcal{Q}$ be a C_b-balanced spectral distribution. Let $G : [m] \to \mathbb{R}^d$. Then $|\langle \nabla \Phi_{\mathcal{Q}}(t), G \rangle| \le (1 + C_b)\mathbb{E}_i \|G(i)\|$.*

Combining the three key properties of the potential function, along with Assumption Stability allows us to bound the dynamics of the potential function in the following way (formalized in Theorem 39):

$$\frac{d}{dt}\Phi_{\mathcal{Q}}(t) \le -\frac{C_{\mathrm{LSC}}\sqrt{L(\rho_t^{\mathrm{MF}})}}{C}\Phi_{\mathcal{Q}}(t) + C J_{\mathrm{avg}} \int_{s=0}^{t} \Phi_{\mathcal{Q}}(s) ds + C J_{\max}(\epsilon_m + \epsilon_n), \tag{8}$$

where $C = O_{C_{\rho^*}, C_{\mathrm{reg}}}(1)$. Theorem 7 follows by analyzing this differential equation. We leverage Assumption Stability to prove (8), by bounding the term $\mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i) \notin B_\tau)$ which arises from Lemmas 13 and 14.

4.2. Self-Concordance Argument to Bound $J_{\max}$

To avoid exponential growth in $J_{t,s}^\perp$, we make the following observation.

Observation 1 *When the velocity $\nu(w, \rho_t^{\mathrm{MF}})$ of a particle w is small, so is $\|D_t^\perp\|$.*

To make this observation more concrete, consider the simplified case of learning a single-index function $f^*(x) = \sigma(x^\top w^*)$ with Gaussian data, where $\sigma(z) = \mathrm{He}_k(z)$ for $k > 2$. We expect a similar property may hold in other low-dimensional feature-learning problems, where the local non-convexity arises only in a low-dimensional subspace. For a neuron w_t, when $\alpha_t := w_t^\top w^*$ is small (and assume for simplicity that α_t is positive), we have that

$$\nu(\alpha_t) := \frac{d}{dt}\alpha_t \approx \alpha_t^{k-1}\ , \text{ thus } \ \frac{d}{d\alpha}\nu(\alpha_t) \approx (k-1)\alpha_t^{k-2} \approx \frac{k-1}{\alpha_t}\nu(\alpha_t)\ .$$

By showing that $\|D_t^\perp\|$ is dominated by $\frac{d}{d\alpha}\nu(\alpha_t)$, we get the desired "self-concordance" property:

$$\|D_t^\perp\| = \left\|\nabla_{w_t}\nu(w_t, \rho_t^{\mathrm{MF}})\right\| \lessapprox \frac{(k-1)}{\alpha_t}\nu(\alpha_t).$$

Recalling the differential equation of $J_{t,s}^\perp$, we have just shown that $\frac{d}{dt}\|J_{t,s}^\perp\| \leq \frac{(k-1)}{\alpha_t}\nu(\alpha_t)\|J_{t,s}^\perp\|$. Note that trivially, α_t satisfies the differential equation $\frac{d}{dt}\alpha_t = \frac{1}{\alpha_t}V(\alpha_t)\alpha_t$. As a result, one can easily deduce that $\|J_{t,s}^\perp\| \leq \left(\frac{|\alpha_t|}{|\alpha_s|}\right)^{k-1}$; see Lemma 50.

4.3. Averaging Argument to Bound J_{avg}

Recall that in order to use our approach to achieve a propagation of chaos for polynomially sized networks, for any $w', v \in \mathbb{S}^{d-1}$ and τ, we must have

$$\sup_{s,t\leq T, w', v\in\mathbb{S}^{d-1}} \mathbb{E}_{w\sim\rho_0}\|J_{t,s}^\perp(w)H_s^\perp(w,w')v\|\mathbf{1}(\xi_t(w) \notin B_\tau) \leq O_\tau\left(\frac{1}{T}\right),$$

where T is the desired training time. We briefly give some intuition for why this holds in the single-index model $f^*(x) = \mathrm{He}_k(x^\top w^*)$, which requires $T = \Theta(d^{(k-2)/2})$. To tightly bound $J_{\text{avg}}(\tau)$, we leverage the fact that neurons far from $\pm w^*$ are dispersed. By averaging over the "level set" of neurons with $\alpha_s(w) = \alpha$ (where $\alpha_t(w) := |{w^*}^\top \xi_t(w)|$) we have

$$\sup_{w', v\in\mathbb{S}^{d-1}} \mathbb{E}_{w:|\alpha_s(w)|=\alpha}\|H_s^\perp(w,w')v\| \leq \max\left(d^{-1/2}, \alpha\right)^{k-1}.$$

Plugging this in for $t \leq T$, along with the bound $\|J_{t,s}^\perp\| \leq \left(\frac{|\alpha_t|}{|\alpha_s|}\right)^{k-1}$ from above, yields

$$\begin{aligned} J_{\text{avg}}(\tau) &\leq \mathbb{E}_w\left(\frac{|\alpha_t(w)|}{|\alpha_s(w)|}\right)^{k-1}\max\left(\sqrt{d}^{-1}, \alpha_s(w)\right)^{k-1}\mathbf{1}(|\alpha_t(w)| \leq 1-\tau) \\ &\lessapprox \mathbb{E}_w|\alpha_t(w)|^{k-1}\mathbf{1}(|\alpha_t(w)| \leq 1-\tau), \end{aligned}$$

Bounding this final term results from the observation the particles escape the saddle at roughly uniform time in the interval $[0, T]$ (see Figure 2(right) and Proposition 47).

5. Conclusion

We studied propagation of chaos in the context of gradient-based training of shallow neural networks. By leveraging several key geometric assumptions of the optimization landscape, we established non-asymptotic guarantees of finite-width dynamics with polynomial dependency in all relevant parameters. At the heart of our technical contributions is a tailored potential function that balances the intricate interactions that arise between particle fluctuations around their idealized mean-field evolution. In essence, our assumptions exploit a form of self-concordance in the instantaneous potentials, as well as symmetries in the minimizing mean-field measure. While these assumptions rule out generic interaction particle systems, they crucially capture several problems of interest, such as planted models including single-index target functions. An enticing future direction is remove the local strong convexity assumptions to extend to the case when ρ^* is a manifold; among other settings, this captures the learning a misspecified SIM. Another interesting direction is to go beyond the Monte Carlo scale of fluctuations, which has been established asymptotically under certain conditions [CRBVE20, PN21].

Acknowledgments

The authors would like to thank Yuqing Wang, Jamie Simon, Taiji Suzuki, Jason Lee, and Andrea Montanari and anonymous referees for useful discussions during the completion of this work. JB acknowledges funding support from NSF DMS-MoDL 2134216 and NSF CAREER CIF 1845360. This material is based upon MG's work supported by the NSF under award 2402314. This work was done in part while MG and DW were visiting the Simons Institute for the Theory of Computing.

References

[AAM22] Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks. In *Conference on Learning Theory*, pages 4782–4887. PMLR, 2022.

[AAM23] Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2552–2623. PMLR, 2023.

[ADK+24] Luca Arnaboldi, Yatin Dandi, Florent Krzakala, Luca Pesce, and Ludovic Stephan. Repetita iuvant: Data repetition allows sgd to learn high-dimensional multi-index functions. *arXiv preprint arXiv:2405.15459*, 2024.

[AGP24] Gérard Ben Arous, Cédric Gerbelot, and Vanessa Piccolo. High-dimensional optimization for multi-spiked tensor pca. *arXiv preprint arXiv:2408.06401*, 2024.

[ASKL23] Luca Arnaboldi, Ludovic Stephan, Florent Krzakala, and Bruno Loureiro. From high-dimensional & mean-field dynamics to dimensionless odes: A unifying approach to sgd in two-layers networks. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1199–1227. PMLR, 2023.

[AZLS19] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International conference on machine learning*, pages 242–252. PMLR, 2019.

[Bac17] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

[BAGJ21] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *J. Mach. Learn. Res.*, 22:106–1, 2021.

[BBPV23] Alberto Bietti, Joan Bruna, and Loucas Pillaud-Vivien. On learning gaussian multi-index models with gradient flow. *arXiv preprint arXiv:2310.19793*, 2023.

[BBSS22] Alberto Bietti, Joan Bruna, Clayton Sanford, and Min Jae Song. Learning single-index models with shallow neural networks. In *Advances in Neural Information Processing Systems*, 2022.

[BES+22] Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. *Advances in Neural Information Processing Systems*, 35:37932–37946, 2022.

[BMZ23] Raphaël Berthier, Andrea Montanari, and Kangjie Zhou. Learning time-scales in two-layers neural networks. *arXiv preprint arXiv:2303.00055*, 2023.

[CB18] Lenaic Chizat and Francis Bach. On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport. 2018.

[CB20] Lénaïc Chizat and Francis Bach. Implicit Bias of Gradient Descent for Wide Two-layer Neural Networks Trained with the Logistic Loss. 2020.

[CD22] Louis-Pierre Chaintron and Antoine Diez. Propagation of chaos: a review of models, methods and applications. i. models and methods. *arXiv preprint arXiv:2203.00446*, 2022.

[CG24] Ziang Chen and Rong Ge. Mean-field analysis for learning subspace-sparse polynomials with gaussian input. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[Chi22a] Lénaïc Chizat. Mean-field langevin dynamics: Exponential convergence and annealing. *Transactions on Machine Learning Research*, 2022.

[Chi22b] Lenaic Chizat. Sparse optimization on measures with over-parameterized gradient descent. *Mathematical Programming*, 194(1-2):487–532, 2022.

[Chi22c] Lenaic Chizat. Sparse optimization on measures with over-parameterized gradient descent. *Mathematical Programming*, 2022.

[Cla12] Pete L Clark. The instructor's guide to real induction. *arXiv preprint arXiv:1208.0973*, 2012.

[CLRW24] Fan Chen, Yiqing Lin, Zhenjie Ren, and Songbo Wang. Uniform-in-time propagation of chaos for kinetic mean field langevin dynamics. *Electronic Journal of Probability*, 29:1–43, 2024.

[COB19] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in neural information processing systems*, 32, 2019.

[CRBVE20] Zhengdao Chen, Grant Rotskoff, Joan Bruna, and Eric Vanden-Eijnden. A dynamical central limit theorem for shallow neural networks. *Advances in Neural Information Processing Systems*, 33:22217–22230, 2020.

[CVEB22] Zhengdao Chen, Eric Vanden-Eijnden, and Joan Bruna. On feature learning in neural networks with global convergence guarantees. *arXiv preprint arXiv:2204.10782*, 2022.

[CWPPS23] Elizabeth Collins-Woodfin, Courtney Paquette, Elliot Paquette, and Inbar Seroussi. Hitting the high-dimensional notes: An ode for sgd learning dynamics on glms and multi-index models. *arXiv preprint arXiv:2308.08977*, 2023.

[DBDFS20] Valentin De Bortoli, Alain Durmus, Xavier Fontaine, and Umut Simsekli. Quantitative propagation of chaos for sgd in wide neural networks. *Advances in Neural Information Processing Systems*, 33:278–288, 2020.

[DJL+17] Simon S Du, Chi Jin, Jason D Lee, Michael I Jordan, Aarti Singh, and Barnabas Poczos. Gradient descent can take exponential time to escape saddle points. *Advances in neural information processing systems*, 30, 2017.

[DKL+23] Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. Learning two-layer neural networks, one (giant) step at a time. *arXiv preprint arXiv:2305.18270*, 2023.

[DLS22] Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural Networks can Learn Representations with Gradient Descent. 2022.

[DNGL23] Alex Damian, Eshaan Nichani, Rong Ge, and Jason D Lee. Smoothing the landscape boosts the signal for sgd: Optimal sample complexity for learning single index models. *Advances in Neural Information Processing Systems*, 36, 2023.

[DTA+24] Yatin Dandi, Emanuele Troiani, Luca Arnaboldi, Luca Pesce, Lenka Zdeborová, and Florent Krzakala. The benefits of reusing batches for gradient descent in two-layer networks: Breaking the curse of information and leap exponents. *arXiv preprint arXiv:2402.03220*, 2024.

[DZPS18] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2018.

[FDGW21] Axel Flinth, Frédéric De Gournay, and Pierre Weiss. On the linear convergence rates of exchange and continuous methods for total variation minimization. *Mathematical Programming*, 190(1):221–257, 2021.

[Gla23] Margalit Glasgow. Sgd finds then tunes features in two-layer neural networks with near-optimal sample complexity: A case study in the xor problem. *arXiv preprint arXiv:2309.15111*, 2023.

[GMMM19] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Limitations of lazy training of two-layers neural network. *Advances in Neural Information Processing Systems*, 32, 2019.

[GMMM21] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 2021.

[HC23] Karl Hajjar and Lénaïc Chizat. On the symmetries in the dynamics of wide two-layer neural networks. *Electronic Research Archive*, 31(4), 2023.

[HRSS19] Kaitong Hu, Zhenjie Ren, David Siska, and Lukasz Szpruch. Mean-field langevin dynamics and energy landscape of neural networks. *arXiv preprint arXiv:1905.07769*, 2019.

[JGH18] Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. 2018.

[JMM19] Adel Javanmard, Marco Mondelli, and Andrea Montanari. Analysis of a two-layer neural network via displacement convexity. *arXiv preprint arXiv:1901.01375*, 2019.

[JMS24] Nirmit Joshi, Theodor Misiakiewicz, and Nathan Srebro. On the complexity of learning sparse functions with statistical and gradient queries. *arXiv preprint arXiv:2407.05622*, 2024.

[KZC+24] Yunbum Kook, Matthew S Zhang, Sinho Chewi, Murat A Erdogdu, and Mufan (Bill) Li. Sampling from the mean-field stationary distribution. *arXiv preprint arXiv:2402.07355*, 2024.

[LMZ20] Yuanzhi Li, Tengyu Ma, and Hongyang R Zhang. Learning over-parametrized two-layer neural networks beyond NTK. 2020.

[LOSW24] Jason D Lee, Kazusato Oko, Taiji Suzuki, and Denny Wu. Neural network learns low-dimensional polynomials with sgd near the information-theoretic limit. *Advances in Neural Information Processing Systems*, 37:58716–58756, 2024.

[MHPG+23] Alireza Mousavi-Hosseini, Sejun Park, Manuela Girotti, Ioannis Mitliagkas, and Murat A Erdogdu. Neural networks efficiently learn low-dimensional representations with SGD. In *The Eleventh International Conference on Learning Representations*, 2023.

[MHWE24] Alireza Mousavi-Hosseini, Denny Wu, and Murat A Erdogdu. Learning multi-index models with neural networks via mean-field langevin dynamics. *arXiv preprint arXiv:2408.07254*, 2024.

[MMM19] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on learning theory*, pages 2388–2464. PMLR, 2019.

[MMN18] 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*, 115(33):E7665–E7671, 2018.

[MU25] Andrea Montanari and Pierfrancesco Urbani. Dynamical decoupling of generalization and overfitting in large two-layer networks. *arXiv preprint arXiv:2502.21269*, 2025.

[MZD+23] Arvind Mahankali, Haochen Zhang, Kefan Dong, Margalit Glasgow, and Tengyu Ma. Beyond ntk with vanilla gradient descent: A mean-field analysis of neural networks with polynomial width, samples, and time. *Advances in Neural Information Processing Systems*, 36:57367–57480, 2023.

[Nit24] Atsushi Nitanda. Improved particle approximation error for mean field neural networks. *arXiv preprint arXiv:2405.15767*, 2024.

[NS17] Atsushi Nitanda and Taiji Suzuki. Stochastic particle gradient descent for infinite ensembles. *arXiv preprint arXiv:1712.05438*, 2017.

[NWS22] Atsushi Nitanda, Denny Wu, and Taiji Suzuki. Convex analysis of the mean field langevin dynamics. In *International Conference on Artificial Intelligence and Statistics*, pages 9741–9757. PMLR, 2022.

[OSSW24] Kazusato Oko, Yujin Song, Taiji Suzuki, and Denny Wu. Learning sum of diverse features: computational hardness and efficient gradient-based training for ridge combinations. *arXiv preprint arXiv:2406.11828*, 2024.

[PKP23] Clarice Poon, Nicolas Keriven, and Gabriel Peyré. The geometry of off-the-grid compressed sensing. *Foundations of Computational Mathematics*, 23(1):241–327, 2023.

[PN21] Huy Tuan Pham and Phan-Minh Nguyen. Limiting fluctuation and trajectorial stability of multilayer neural networks with mean field training. *Advances in Neural Information Processing Systems*, 34:4843–4855, 2021.

[RJBVE19] Grant M Rotskoff, Samy Jelassi, Joan Bruna, and Eric Vanden-Eijnden. Global convergence of neuron birth-death dynamics. In *Proceedings of International Conference on Machine Learning 36*, pages 9689–9698, 2019.

[RL24] Yunwei Ren and Jason D Lee. Learning orthogonal multi-index models: A fine-grained information exponent analysis. *arXiv preprint arXiv:2410.09678*, 2024.

[RVE18] Grant M Rotskoff and Eric Vanden-Eijnden. Neural networks as Interacting Particle Systems: Asymptotic convexity of the Loss Landscape and Universal Scaling of the Approximation Error. *arXiv preprint arXiv:1805.00915*, 2018.

[RZG23] Yunwei Ren, Mo Zhou, and Rong Ge. Depth separation with multilayer mean-field networks. In *The Eleventh International Conference on Learning Representations*, 2023.

[SBH24] Berfin Simsek, Amire Bendjeddou, and Daniel Hsu. Learning gaussian multi-index models with gradient flow: Time complexity and directional convergence. *arXiv preprint arXiv:2411.08798*, 2024.

[SNW22] Taiji Suzuki, Atsushi Nitanda, and Denny Wu. Uniform-in-time propagation of chaos for the mean-field gradient langevin dynamics. In *The Eleventh International Conference on Learning Representations*, 2022.

[SS20] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020.

[SWN23] Taiji Suzuki, Denny Wu, and Atsushi Nitanda. Convergence of mean-field langevin dynamics: Time and space discretization, stochastic gradient, and variance reduction. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[SWON23] Taiji Suzuki, Denny Wu, Kazusato Oko, and Atsushi Nitanda. Feature learning via mean-field langevin dynamics: classifying sparse parities and beyond. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[SYS21] Itay M Safran, Gilad Yehudai, and Ohad Shamir. The Effects of Mild Over-parameterization on the Optimization Landscape of Shallow ReLU Neural Networks. 2021.

[Szn91] Alain-Sol Sznitman. Topics in propagation of chaos. *Lecture notes in mathematics*, pages 165–251, 1991.

[Tel23] Matus Telgarsky. Feature selection and low test error in shallow low-rotation relu networks. In *The Eleventh International Conference on Learning Representations*, 2023.

[TS24] Shokichi Takakura and Taiji Suzuki. Mean-field analysis on two-layer neural networks from a kernel perspective. *arXiv preprint arXiv:2403.14917*, 2024.

[WLLM19] Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. *Advances in Neural Information Processing Systems*, 32, 2019.

[WMHC24] Guillaume Wang, Alireza Mousavi-Hosseini, and Lénaïc Chizat. Mean-field langevin dynamics for signed measures via a bilevel approach. *Advances in Neural Information Processing Systems*, 37:35165–35224, 2024.

[YH20] Greg Yang and Edward J Hu. Feature learning in infinite-width neural networks. *arXiv preprint arXiv:2011.14522*, 2020.

[ZCZG20] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine learning*, 109:467–492, 2020.

Contents

Appendix A. Additional Related Works

Mean-field analysis of shallow neural networks. The mean-field analysis views the training of two-layer neural network (1) as an interacting particle system, and studies the evolution of the distribution of particles via the mean-field PDE [NS17, CB18, MMN18, RVE18, SS20]. While most optimization guarantees for mean-field neural networks are *qualitative* in nature, quantitative convergence rate can also be established under additional structural assumptions on the learning problem [JMM19, Chi22c, CRBVE20, CVEB22] or modification of the training dynamics [RJBVE19, WLLM19, NWS22, Chi22a].

Recent works have studied the statistical efficiency of mean-field neural networks in learning low-dimensional target functions including multi-index models and k-parity. These existing analyses can be divided into two approaches: (i) simplify the mean-field PDE using the symmetry and low-dimensional structure, and study the *dimensional-free* dynamics [HC23, ASKL23] at short timescale $T = \tilde{O}_d(1)$ [AAM22, MZD+23, JMS24]; (ii) directly characterize the converged solution using global optimality conditions [WLLM19, Tel23, SWON23, MHWE24]. While the latter approach establishes a much larger learnable function class (e.g., see [Bac17]), the computational complexity is exponential in the (intrinsic) dimensionality of the problem.

Gradient-based learning of single/multi-index models. Outside of the mean-field regime, feature learning in neural networks has also been studied in a "narrow-width" setting, where neurons evolve (almost) independently and align with the low-dimensional target function during gradient-based training. Prior analyses in this regime mostly considered target functions that depends on $k = O_d(1)$ directions of the input, such as single-index models [BAGJ21, BES+22, BBSS22, MHPG+23, DNGL23, DTA+24, LOSW24, ADK+24] and multi-index models [DLS22, AAM22, AAM23, DKL+23, BBPV23, CWPPS23, Gla23, AGP24]. For the "rank-extensive" setting $k \gg 1$, recent works have investigated the additive setting where the target function is a sum of k orthogonal single-index models [LMZ20, OSSW24, RL24, SBH24].

Appendix B. Full Statement of Assumptions and Discussion

In this section, we explore whether propagation of chaos may hold more generally than beyond our setting and conditions. We provide several remarks on the necessity of our assumptions, both in the context of our proof approach, and based on empirical simulations, which are given in full in Section C.

B.1. Omitted Assumptions

We begin by stating the full versions of the assumptions which were omitted in Section 3. We then briefly discuss the definition of propagation of chaos and several related phenomena in Section B.2. Finally, in Sections B.3-B.7, we provide remarks on the assumptions.

Let $V = \mathrm{span}(\mathrm{supp}(\rho^*))$ and let U be the space orthogonal to V in $\mathbb{R}^d$. Let

$$C_{\rho^*} := \min\left(|\mathrm{supp}(\rho^*)|, \dim(V)^{2\,\mathrm{degree}(\sigma)+1}\right).$$

Assumption LSC (Local Strong Convexity (Full Version of Assumption LSC (abbv))) *The problem is* (C_{LSC}, τ) locally strongly convex *up to time* T *if for any* $t \le T$ *and any* w *with* $\xi_t(w) \in B_\tau$,

we have

$$D_t^{\perp}(w) \preceq -C_{\mathrm{LSC}} P^{\perp}_{\xi_t(w)} \| f_{\rho_t^{\mathrm{MF}}} - f^* \| \, .$$

Further, the strong convexity is structured, *if there exist values* $c_t^1, c_t^2 \geq C_{\mathrm{LSC}}$ *such that for any* w *with* $\xi_t(w) \in B_\tau$*, we have*

$$\| c_t^1 V V^\top P^{\perp}_{\xi^\infty(w)} V V^\top + c_t^2 U U^\top - D_t^{\perp}(w) \| \leq \left(\frac{C_{\mathrm{LSC}} \| f_{\rho_t^{\mathrm{MF}}} - f^* \|}{2\sqrt{C_{\rho^*}}} + C_{reg} \tau \right) .$$

Assumption Symmetry (Symmetries of ρ^*) *The* automorphism group $\mathcal{G}$ *of a problem* $(\rho^*, \rho_0, \mathcal{D}_x)$ *is the group of rotations* g *on* $\mathbb{S}^{d-1}$ *where for any* $A \subset \mathbb{S}^{d-1}$*:*

$$\mathbb{P}_{\rho^*}[A] = \mathbb{P}_{\rho^*}[g(A)] \qquad \mathbb{P}_{\mathcal{D}}[A] = \mathbb{P}_{\mathcal{D}_x}[g(A)] \qquad \mathbb{P}_{\rho_0}[A] = \mathbb{P}_{\rho_0}[g(A)]$$

We assume:

- ***I1*** $\mathrm{supp}(\rho^*)$ *is* transitive *under* $\mathcal{G}$*, that is, for any* $w^*, w^{*\prime} \in \mathrm{supp}(\rho^*)$*, there exists* $g \in \mathcal{G}$ *such that* $g(w^*) = w^{*\prime}$*. Further,* $\mathbb{P}_{w \sim \rho_0}[\{\|w - w^*\| = \|w - w^{*\prime}\| \exists w^*, w^{*\prime} \in \mathrm{supp}(\rho^*)\}] = 0$*.*
- ***I2*** *Let* $V = \mathrm{span}(\mathrm{supp}(\rho^*))$ *and let* U *be the space orthogonal to* V *in* $\mathbb{R}^d$*. Then the distribution* $\mathcal{D}_x$ *on covariates* x *factorizes over* U *and* V*, that is* $\mathcal{D}_x = \mathcal{D}_U \otimes \mathcal{D}_V$*, where* $\mathcal{D}_U$ *is a distribution on* V *and* $\mathcal{D}_U$ *is a distribution on* U*. Further,* $\mathbb{E}_{x_U \sim \mathcal{D}_U} x = 0$*, and* $\mathbb{E}_{x_U \sim \mathcal{D}_U} x x^\top = U U^\top$*.*

B.2. Measuring Propagation of Chaos

A standard definition[4] of propagation of chaos (PoC) (see [Szn91, Prop. 2.2]) is that for all t, we have the convergence in law of the random distribution ρ_t^m to the constant distribution ρ_t^{MF}:

$$\lim_{m \to \infty} \rho_t^m \to \rho_t^{\mathrm{MF}}. \tag{9}$$

Equivalently, for any two continuous test functions ψ_1, ψ_2, we have that

$$\lim_{m \to \infty} \mathbb{E}_{w_1, w_2 \sim \rho_t^m} \psi_1(w_1) \psi_2(w_2) = \left(\mathbb{E}_{w \sim \rho_t^{\mathrm{MF}}} \psi_1(w) \right) \left(\mathbb{E}_{w \sim \rho_t^{\mathrm{MF}}} \psi_2(w) \right) \, . \tag{10}$$

Of primary interest in our paper is a weaker PoC phenomenon, which we will henceforth refer to as *PoC in function error*: almost surely with respect to the draw of $\hat{\rho}_0^m$,

$$\lim_{m \to \infty} \| f_{\rho_t^m} - f_{\rho_t^{\mathrm{MF}}} \|^2 \to 0.$$

It is easy to check that PoC in function error is implied by (10) by using test functions of the form $\psi_x(w) := \sigma(w^\top x)$. PoC in function error implies convergence of the risk of ρ_t^m to the risk of ρ_t^{MF}, and thus is the most practically relevant (see e.g. [SNW22]). On the other hand, our proof considers a much stronger PoC phenomenon. Our potential-function based proof yields almost surely over the initialization

$$\lim_{m \to \infty} \Omega(\Delta_t) := \mathbb{E}_i \| \Delta_t(i) \|_2 = 0.$$

4. Stronger notions of uniform convergence over t are also available; see [CD22, Section 3.4].

Here the m is implicit in Δ_t. This is a much stronger notion than (9) (it implies $\lim_{m\to\infty} W_1(\rho_t^m, \rho_t^{\mathrm{MF}}) = 0$), and we will refer to it as *PoC via fixed parameter-coupling*.

Remarkably in our neural network setting (though not necessarily for general interacting particle systems), up to the parameter $J_{\max}$ and a time horizon t, PoC in function error for all $s \le t$ implies PoC via fixed parameter-coupling. Indeed, by (3), we have that

$$\begin{aligned}\|\Delta_t(i)\| &\le \int_0^t \|J_{t,s}^{\perp}(i)\| \left(\|\mathbb{E}_j H_s^{\perp}(i,j)\Delta_s(j)\| + \|\boldsymbol{\epsilon_{s,i}}\|\right) ds \\ &= O\left(J_{\max}\int_0^t \left(\sqrt{\Delta_s^\top H_s^{\perp}\Delta_s} + \mathbb{E}_j\|\Delta_s(j)\|^2 + \|\Delta_s(i)\|^2 + \epsilon_m + \epsilon_n\right) ds\right) \\ &= O\left(J_{\max}\int_0^t \left(\|f_{\rho_s^m} - f_{\rho_s^{\mathrm{MF}}}\| + \mathbb{E}_j\|\Delta_s(j)\|^2 + \|\Delta_s(i)\|^2 + \epsilon_m + \epsilon_n\right) ds\right). \quad (11)\end{aligned}$$

Here the last equation follows from the following lemma proved using a Taylor expansion of $f_{\rho_t^m}$.

Lemma 16 *Suppose Regularity Assumption **R1** holds. For any t, we have*

$$\Delta_t^\top H_t^{\perp}\Delta_t \le 2\|f_{\rho_t^{\mathrm{MF}}} - f_{\rho_t^{\mathrm{MF}}}\|^2 + O\left(m^{-1}\right) + C_{reg}^2\left(\mathbb{E}_i\|\Delta_t(i)\|^2\right)^2.$$

Solving Eq. (11) yields that for m such that $\|f_{\rho_s^m} - f_{\rho_s^{\mathrm{MF}}}\| + \epsilon_m + \epsilon_n \ll \frac{1}{(tJ_{\max})^2}$ for all $s \le t$,

$$\mathbb{E}_i\|\Delta_t(i)\| = O\left(J_{\max} t \max_{s\le t}\left(\|f_{\rho_s^m} - f_{\rho_s^{\mathrm{MF}}}\| + \epsilon_m + \epsilon_n\right)\right). \quad (12)$$

Indeed, one can show this by inductively bounding the second order terms from time 0 to t.

All of the above PoC phenomena can be quantified non-asymptotically, and the main question of this paper is whether for certain problems the above quantities (or their differences) decay at a rate $\frac{\mathrm{poly}(t,d)}{\mathrm{poly}(m)}$, uniformly over all $d > 0$ and all $t \in [0, T(d)]$. Here $T(d)$ is a desired stopping time, e.g., when some fixed population loss ϵ is achieved.

B.3. Spherical Constraint and Second Layer Weights

When the weights of the neural network are not constrained to the sphere, propagation of chaos may fail even in simple well-specified settings: to see this, consider the case of learning a SIM with information exponent $k > 2$ using a neural network with homogeneous activation function. With polynomial width, we expect the standard $T \approx d^{(k-2)/2}$ convergence time. Whereas at the infinite-width limit, we may learn the target function by amplifying neurons that already attain large alignment at initialization due to homogeneity. In particular, we can achieve $T = d^{(k-2)/k}$ convergence time by leveraging neurons with initial alignment greater than $d^{-1/k}$ — to see this, observe there is roughly $\exp(-d^{(k-2)/k})$ fraction of neurons in the network with such initial overlap, and thus we need to grow these neurons to a scale of $\exp(d^{(k-2)/k})$, which takes $d^{(k-2)/k}$ time. A similar phenomenon occurs if we train the second-layer weights in the network and allow them to be unbounded. Note, however, that there is nothing precluding our results from holding if the (fixed) second layer is initialized differently.

B.4. Local Strong Convexity (Assumption LSC)

We focus here on the main part captured in Assumption LSC (abbv). See also the previous Remark 8. The additional *structured* condition in Assumption LSC is discussed with the symmetry conditions.

Local strong convexity plays a key part in how we bound the potential $\Phi_{\mathcal{Q}}(t)$ via the differential equation in (8). Indeed, plugging in the bound $J_{\text{avg}} \leq 1/T$ yields:

$$\frac{d}{dt}\Phi_{\mathcal{Q}}(t) \leq -\frac{C_{\text{LSC}}\sqrt{L(\rho_t^{\text{MF}})}}{C}\Phi_{\mathcal{Q}}(t) + \frac{C}{T}\int_{s=0}^{t}\Phi_{\mathcal{Q}}(s)ds + CJ_{\max}(\epsilon_m + \epsilon_n).$$

If C_{LSC} goes to 0, then the best bound on this differential equation becomes

$$\Phi_{\mathcal{Q}}(t) \lessapprox CJ_{\max}(\epsilon_m + \epsilon_n)\exp\left(t\sqrt{\frac{C}{T}}\right),$$

which would require that m be super-polynomially large in T in order to bound $\Phi_{\mathcal{Q}}(T)$.

As discussed in Remark 8, local strong convexity can only hold in problems where ρ^* is *atomic*. Thus it cannot capture example when ρ^* is distributed on a manifold, or for "misspecified" problems where the target link function differs from the network activation, e.g., $f^*(x) = \phi(x^\top w^*)$ for $\phi \neq \sigma$. These examples are particularly interesting because training with the correlation loss is insufficient. In our 1- or 2-index non-atomic experiments, however, we still observed propagation of chaos for the values of m we simulated (see e.g., the Misspecified and Circle problems depicted in Figures 4, 7).

In non-atomic examples, it is unreasonable to hope that $\mathbb{E}_i\|\Delta_t(i)\|$ will remain bounded for all t; thus addressing this case would require proving either a bound on the Wasserstein-1 distance, $W_1(\rho_t^m, \rho_t^{\text{MF}})$, or a bound on the function error $\|f_{\rho_t^m} - f_{\rho_t^{\text{MF}}}\|^2 \approx \Delta_t^\top H_t^\perp \Delta_t$.

B.5. Stability Conditions (Assumption Stability)

To achieve propagation of chaos with polynomially many neurons, we believe it is necessary in standard settings that $\sup_{s,t\leq T}\|J_{t,s}(w)\|$ is polynomially bounded with high probability over w. This is only a slightly weaker condition than the current $J_{\max}$ assumption. Getting around such an assumption would require strong directional control over the $\boldsymbol{\epsilon}_{t,i}$, which we do not expect to be possible.

The necessity of the strong assumption on J_{avg}, however, is mysterious to us. The neuron-to-neuron error-propagation described in section 3.1 seems hard to prevent without a similar assumption. Even if we leverage the fact that the interaction term is PSD (and thus creates a repulsion between the neurons), there could be oscillatory exponential growth of the $\Delta_t(i)$'s. Nevertheless, in our simulations, we were not able to find an example where violating the J_{avg} assumption precluded propagation of chaos; see for example the Staircase problem depicted in Figure 9.

Remark 17 (Order-1 Saddles /Information Exponent 2) *The reader may notice that both the assumption on J_{avg} and J_{max} may fail in simple cases where the information exponent is 2 (and thus, our application to SIMs in Theorem 9 restricts to $k^* > 2$). The J_{avg} assumption fails in this case because the neurons all escape the saddle at roughly the same time, and thus there exists some time t (roughly this escape time) where the expression in Assumption Stability is of order 1. We believe overcoming this obstacle should be possible by working with a (s,t)-dependent version of J_{avg}.*

If we allow T to grow polynomially large in d, the J_{max} assumption fails in this case because we expect to have a small fraction of neurons initialized exponentially close to the saddle (e.g., in SIM $|w^\top w^*| \le \exp(-T)$). For such neurons w, $\|J_{t,0}(w)\|$ will grow exponentially in t for t as large at T. In the SIM case, overcoming this challenge may be possible by defining J_{max} to exclude the worst $\mathrm{poly}(d,T)/\sqrt{m}$ fraction of neurons. However, more generally, if a constant fraction of the neurons get sucked in exponentially close to a saddle (similar to [DJL+17, Figure 1b]), we cannot expect to have propagation of chaos with polynomially many neurons. We speculate it is possible that adding a very small diffusion term to the dynamics could enable propagation of chaos.

B.6. Symmetry Conditions (Assumption Symmetry)

The *structured* condition in Assumption LSC. The structured condition Assumption LSC is used in proving Lemma 14, which shows that the potential decreases due to the local strong convexity near the teacher neurons. We believe its necessity is an artifact of how we designed the potential function.

In general, the structured condition holds for all 2-index problems with Gaussian data, because at $\xi^\infty(w) \in \mathrm{span}(V)$, (a) the projection of $D_t^\perp(w)$ onto the U-space will be a multiple of $UU^\top$, and (b) the projection of $D_t^\perp(w)$ onto the V-space will be one-dimensional, and thus a multiple of $VV^\top P^\perp_{\xi^\infty(w)} VV^\top$. Using a continuity argument between $\xi^\infty(w)$ and $\xi_t(w)$ yields the condition for all $\xi_t(w) \in B_\tau$. Beyond 2-index problems, we are not aware of exactly when this condition holds, though we expect it does not hold for many non-symmetric problems.

Symmetry Assumption (Assumption Symmetry). The transitivity condition (**I1** in Assumption Symmetry) plays an important role in our proof. Namely, Lemma 31 (see also Definition 28) uses transitivity to guarantee that the eigenfunctions of the interaction matrix *at convergence time,* are also eigenfunctions of the interaction submatrix of neurons that have converged at time t (for any t!). This "restricted isometry" property allows us to define our potential function *independently* of the time t. We expect that without restricted isometry, one would have to design a potential function which depends on the time t.

The transitive condition in **I1** holds for various non-trivial teacher-student problems, for example: learning k orthogonal teacher neurons for any k, learning any two non-orthogonal teacher neurons, or learning a ring of equally spaced teacher neurons on a circle. In the latter two examples, training with the correlation loss may fail (for instance, for two non-orthogonal neurons with a small angle between them, correlation loss may converge to the linear combination of teacher neurons); to our knowledge, gradient training of many of these "simple" symmetric examples is still not understood. Note also that the second part of **I1** holds when ρ_0 has bounded Radon-Nikodym derivative with respect to the Lebesgue measure on the sphere, or when $\rho_t^{\mathrm{MF}} \to \rho^*$. Both imply 0 mass on the boundary points.

In all the non-symmetric examples we simulated, the lack of symmetry does not pose an obstacle to propagation of chaos; see for example the XOR_4, Staircase, Misspecified, problems in Figures 4, 5, 9.

B.7. Dependence on C_{ρ^*}

The value C_{ρ^*} is bounded whenever ρ^* is atomic with constant number of neurons, or when ρ^* is in a finite-dimensional subspace, and σ is polynomial. This includes all polynomial multi-index functions.

The value C_{ρ^*}, defined in (6), functions as an upper bound on the rank of the interaction-kernel $k(w, w') = \mathbb{E}_x \sigma'(x^\top w)\sigma'(x^\top w')$ over points in $w, w' \in \mathrm{supp}(\rho^*)$. Having a constant upper bound on the rank of this kernel is useful in constructing a balanced spectral decomposition of $H_\infty^\perp$, which (loosely) ensures that near convergence time, small L_1-bounded changes in Δ_t cannot propagate (via the force of the interaction kernel) to large changes, measured in L_1. While it may be possible, we have not been able to find any simple ways to prevent L_1-growth in Δ_t near convergence time without this constant-rank assumption. In Section C, we simulate several examples in which C_{ρ^*} grows polynomially in d. The presence of various PoC phenomena did not seem to be correlated with the size of C_{ρ^*} — observe that the Misspecified example (Figure 4), which has C_{ρ^*} that grows polynomially in d, demonstrated PoC for relatively small widths.

Appendix C. Experiments

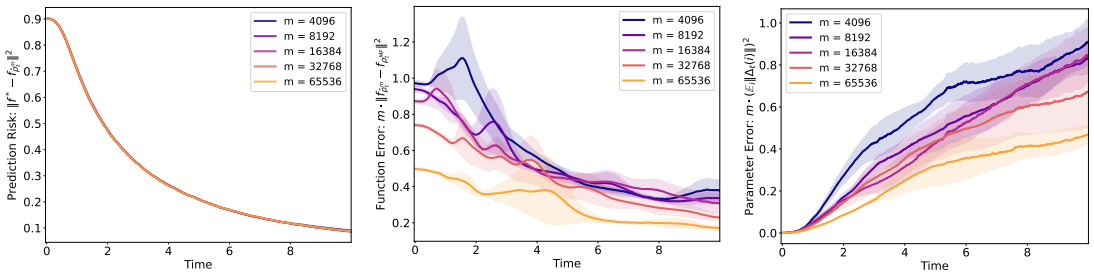
We conduct simulations both to validate our theory in settings which we expect satisfy our assumptions, and to examine what happens when these assumptions fail to hold. Table 1 in Appendix H describes all of the settings we simulated, and documents which assumptions we believe they satisfy. We remark that we did not preferentially chose these examples because we expected (or observed) propagation of chaos: we in fact ran these simulations with the goal of finding a multi-index function in which propagation of chaos fails, and have so far been unsuccessful.

C.1. Experiment Setting

Since we could not simulate an infinite-width network, we measured certain proxies for the distance between ρ_t^m and ρ_t^{MF} by comparing a neural network of width m to a neural network of width M, for $M \gg m$. The full experimental design is described in Section H.1. In brief, we initialize the smaller width-m network to be a subset of the neurons in the larger width-M network; in this way, we can track for all $i \in [m]$ the coupling differences $\hat{\Delta}_t(i)$ (a proxy for $\Delta_t(i)$[5]) throughout training. In our plots, we estimate the following quantities from data throughout the training dynamics: (a) the prediction risk or generalization error, (b) the function error $m \cdot \|f_{\rho_t^m} - f_{\rho_t^{\mathrm{MF}}}\|^2$, and (c) the fixed parameter-coupling error $m \cdot \mathbb{E}_i\|\hat{\Delta}_t(i)\|$. In our full plots in the appendix, we also include several histograms of $\|\hat{\Delta}_t(i)\|$. We plot the above values for a range of widths m, and examine the decay rate in m.

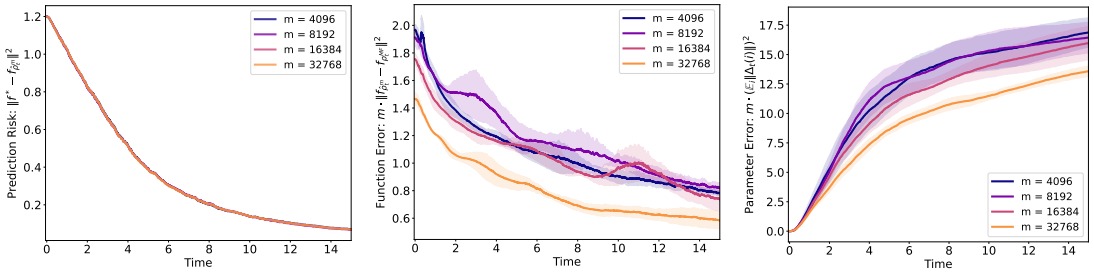
In Figures 3,4,5 we consider (i) the well-specified Gaussian single-index setting with He_4 activation function, which satisfies all our assumptions in Section B, (ii) a misspecified single-index setting where we do not expect ρ^* to be atomic (see e.g., [MZD+23]), and (iii) the multi-index setting of 4-parity function similar to that studied in [Gla23, Tel23, SWON23].

5. Triangle inequality yields that $\|\hat{\Delta}_t(i)\| \le 2\|\Delta_t(i)\|$, though the converse is not necessarily true.



(a) prediction risk $\|f_{\rho_t^m} - f^*\|^2$. (b) scaled function error $m\|f_{\rho_t^m} - f_{\rho_t^{\mathrm{MF}}}\|^2$. (c) scaled parameter coupling error $m(\mathbb{E}_i\|\hat{\Delta}_t(i)\|)^2$.

Figure 3: Well-specified single-index (He_4) target function $f^*(x) = \mathrm{He}_4(x^\top w^*), x \sim \mathcal{N}(0, I_d)$, and $\sigma = \mathrm{He}_4$. We set $d = 32$ and learning rate $\eta = 0.01$.



(a) prediction risk $\|f_{\rho_t^m} - f^*\|^2$. (b) scaled function error $m\|f_{\rho_t^m} - f_{\rho_t^{\mathrm{MF}}}\|^2$. (c) scaled parameter coupling error $m(\mathbb{E}_i\|\hat{\Delta}_t(i)\|)^2$.

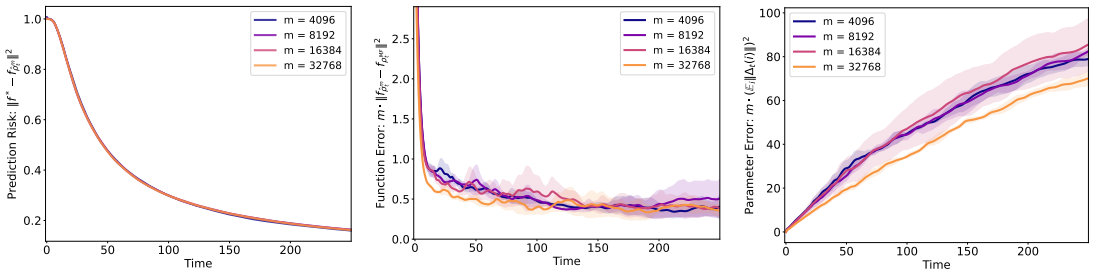
Figure 4: Misspecified single-index (Misspecified) target function $f^*(x) = 0.8\mathrm{He}_4(x^\top w^*) + 0.6\mathrm{He}_6(x^\top w^*), x \sim \mathcal{N}(0, I_d)$, and $\sigma = \mathrm{He}_4 + \mathrm{He}_6$. We set $d = 32$ and learning rate $\eta = 0.01$.

C.2. Takeaways from Simulations

We describe some of our takeaways from the experiments below. More figures can be found in Appendix H.

PoC in function error. In all examples we simulated, we observed that for m large enough, the function error $\|f_{\rho_t^m} - f_{\rho_t^{\mathrm{MF}}}\|^2$ decayed at least linearly with the width m – this is evident in Figures 3,4,5(b). Surprisingly, in all examples the function error decayed nearly monotonically in time.

PoC via fixed parameter-coupling. Similarly to the function error, we observed that the parameter-coupling error $\left(\mathbb{E}_i\|\hat{\Delta}_t(i)\|\right)^2$ decayed at at $1/m$ rate – this is evident in Figure 3,4,5(c). However, unlike the function error, the growth of this error appeared to be linear in t, which is consistent with the upper bound on parameter-coupling error in terms function error given in (12). We note that our experiments show that (12) is not quite tight, as the parameter-coupling error seems to grow slower than (a scaling of) the integral of the function error over time. More extensive experiments could provide more insight on this.



(a) prediction risk $\|f_{\rho_t^m} - f^*\|^2$.

(b) scaled function error $m\|f_{\rho_t^m} - f_{\rho_t^{\mathrm{MF}}}\|^2$.

(c) scaled parameter coupling error $m(\mathbb{E}_i\|\hat{\Delta}_t(i)\|)^2$.

Figure 5: 4-parity (XOR_4) target function $f^*(x) = \prod_{j\le 4}[x]_j, [x]_i \sim \mathrm{Unif}\{1,-1\}$, and $\sigma =$ SoftPlus with temperature 16. We set $d = 32$ and learning rate $\eta = 0.05$.

Remark 18 *We note that our experiments are technically insufficient to guarantee that the PoC rate is polynomial in d, T, because we did not conduct extensive comparisons of the decay rate across growing values of d, T. However, in all of the experiments we plotted, we observed linear decay in function error $\|f_{\rho_t^m} - f_{\rho_t^{\mathrm{MF}}}\|^2 \lesssim m^{-1}$ starting even at the smallest value of $m = 2^{12}$, which suggests that if the parameter-coupling PoC rate is some $g(d,t)/m$, then for all of experiments, $g(d,t) \le t$ for the value of d and range of t we simulated. Thus, we conjecture that in all these problems there is PoC in function error and in fixed parameter-coupling error at a rate at most $\mathrm{poly}(d,t)/m$.*

Appendix D. Proofs of Lemmas from Basic Setup

D.1. Notations

Throughout this section, we will use the following notation, which builds upon the notation in our setup from the main body. [6]

$$f(w) := \mathbb{E}_{x\sim\mathcal{D}_x} f^*(x)\sigma(w^\top x)$$
$$f'(w) := (I - ww^\top)\nabla_w f(w)$$

and

$$\hbar(w,w') := \mathbb{E}_{x\sim\mathcal{D}_x}\sigma(w'^\top x)\sigma(w^\top x)$$
$$\hbar'(w,w') := (I - ww^\top)\nabla_w \hbar(w,w').$$

In additional the interaction Hessian $H_t^\perp$ introduced in the introduction, we also define a versions without the orthogonal projection, that is:

$$H_t(w,w') := \hbar'(\xi_t(w),\xi_t(w'))$$
$$H_t^\perp(w,w') = H_t(w,w')(I - \xi_t(w')\xi_t(w'))$$

6. To emphasize the relationship with f and $\hbar$, we deviate from our standard notation convention here in using the lower-case letters f' and $\hbar'$ to denote vector-valued functions.

We also define the *empirical local Hessian* $\bar{D}_t$ (closely related to $D_t^\perp$), where the expectation is taken over $\bar{\rho}_t^m$ instead of ρ_t^{MF}:

$$\bar{D}_t(w) := \nabla_{\xi_t(w)}\nu(\xi_t(w), \bar{\rho}_t^m) = \nabla_{\xi_t(w)}f'(\xi_t(w)) - \mathbb{E}_{w'\sim\bar{\rho}_t^m}\nabla_{\xi_t(w)}\hbar'(\xi_t(w), w').$$
$$D_t^\perp(w) = \nabla_{\xi_t(w)}\nu(\xi_t(w), \rho_t^{\mathrm{MF}}) = \nabla_{\xi_t(w)}f'(\xi_t(w)) - \mathbb{E}_{w'\sim\rho_t^{\mathrm{MF}}}\nabla_{\xi_t(w)}\hbar'(\xi_t(w), w').$$

D.2. Proof of Lemma 5

We being with a basic lemma which uses the regularity of σ to bound the smoothness of various problem parameters.

Lemma 19 *Assume Assumption R1 holds. There exists a constant $C_{reg} = O_{C_{reg}}(1)$ such that the following holds for any w and w' with norm at most 1.*

- ***S1*** $\|\nabla_w\hbar'(w,w')\| \le C_{reg}$ *and* $\|\nabla_w f'(w)\| \le C_{reg}$
- ***S2*** $\left\|\nabla^2_{w'}\hbar'(w,w')\right\| \le C_{reg}$
- ***S3*** $\left\|\nabla^2_{w}\hbar'(w,w')\right\| \le C_{reg}$
- ***S4*** $\|\nabla_{w'}\nabla_w\hbar'(w,w')\| \le C_{reg}$
- ***S5*** $\left\|\nabla^2_w f'(w)\right\|_{op} \le C_{reg}$
- ***S6*** *For any distribution $\rho \in \Delta(\mathbb{S}^{d-1})$, we have* $\left\|\nabla^2_w\nu(w,\rho)\right\|_{op} \le C_{reg}$

Proof [Proof of Lemma 19] These are straightforward to check from the definitions. First note that the operator norm of the first and second derivatives of $I - ww^\top$ is at most 2. Thus for any vector-valued function $\xi(w)$, by chain rule, we have

$$\left\|\nabla_w(I-ww^\top)\xi(w)\right\| \le \|\nabla_w\xi(w)\| + 2\|\xi(w)\|$$
$$\left\|\nabla^2_w(I-ww^\top)\xi(w)\right\| \le 3\left\|\nabla^2_w\xi(w)\right\| + 8\|\nabla_w\xi(w)\|.$$

So to prove the lemma, it suffices to bound (over all $w, w' \in \mathbb{S}^{d-1}$):

$$\|\nabla_w f(w)\|, \left\|\nabla^2_w f(w)\right\|, \left\|\nabla^3_w f(w)\right\|,$$

and

$$\left\|\nabla_w\hbar(w,w')\right\|, \left\|\nabla^2_w\hbar(w,w')\right\|, \left\|\nabla^3_w\hbar(w,w')\right\|, \left\|\nabla_w\nabla_{w'}\nabla_w\hbar(w,w')\right\|, \left\|\nabla^2_w\nabla_{w'}\nabla_w\hbar(w,w')\right\|$$

As an example, for S2, we have

$$\begin{aligned}\left\|\nabla^2_{w'}\hbar'(w,w')\right\|_{op} &\le \sup_{v_2,v_2,v_3\in\mathbb{S}^{d-1}} \mathbb{E}_x\sigma(w^\top x)\sigma'''(w'^\top x)v_1^\top(I-ww^\top)x(v_2^\top x)(v_3^\top x)\\ &\le \sup_{z,z'\in B_2^d}\left(\mathbb{E}_x|\sigma(z^\top x)|^5\right)^{1/5}\left(\mathbb{E}_x|\sigma'''(z'^\top x)|^5\right)^{1/5}\sup_{v\in\mathbb{S}^{d-1}}\left(\mathbb{E}_x|(v^\top x)|^5\right)^{3/5}\\ &\le C_{\mathrm{reg}}/11,\end{aligned}$$

where here the second inequality holds by Holder's inequality, and the final inequality by Assumption **R1**. For **S3**, the argument is the same as the previous one, except we use the product rule to account for the derivatives of $(I - ww^\top)$, which have operator norm at most 1.

For the rest of the terms involving derivatives — up to third order — of K, the argument is near identical, following from Holder's inequality and Assumption **R1**. Thus each of these terms about bounded by $C_{\text{reg}}/11$.

For the terms involving F, as an example, lets expand the the third order term. We have

$$\begin{aligned}\left\|\nabla_w^3 f(w)\right\| &\leq \sup_{v_1,v_2,v_3\in\mathbb{S}^{d-1}} \mathbb{E}_x|\sigma^{(3)}(w^\top x)(v_1^\top x)(v_2^\top x)(v_3^\top x)f^*(x)| \\ &\leq \sup_{z,z'\in B_2^d}\left(\mathbb{E}_x|\sigma^{(3)}(z^\top x)|^5\right)^{1/5}\sup_{v\in\mathbb{S}^{d-1}}\left(\mathbb{E}_x|(v^\top x)|^5\right)^{3/5}\left(\mathbb{E}_x(f^*(x))^5\right)^{1/5} \\ &\leq C_{\text{reg}}/11.\end{aligned}$$

It follows that all the terms in the lemma are bounded by $11\left(C_{\text{reg}}/11\right) = C_{\text{reg}}$. ■

We also prove Lemma 1 and Lemma 16 here, which we restate for the reader's convenience.

Lemma 20 *Suppose Regularity Assumption **R1** holds. With high probability over the draw ρ_0^m, we have*

$$\|f_{\rho_t^{\text{MF}}} - f_{\rho_t^m}\|^2 \leq 2C_{reg}^2\left(\mathbb{E}_i\|\Delta_t(i)\|\right)^2 + \frac{\log(m)}{m}.$$

Lemma 21 *Suppose Regularity Assumption **R1** holds. For any t, we have*

$$\Delta_t^\top H_t^\perp \Delta_t \leq 2\|f_{\rho_t^{\text{MF}}} - f_{\rho_t^{\text{MF}}}\|^2 + O\left(m^{-1}\right) + C_{reg}^2\left(\mathbb{E}_i\|\Delta_t(i)\|^2\right)^2.$$

Proof [Proof of Lemma 1 and Lemma 16]

First we decompose

$$\mathbb{E}_x(f_{\rho_t^{\text{MF}}}(x) - f_{\rho_t^m}(x))^2 \leq 2\mathbb{E}_x(f_{\rho_t^{\text{MF}}}(x) - f_{\bar{\rho}_t^m}(x))^2 + 2\mathbb{E}_x(f_{\bar{\rho}_t^m}(x) - f_{\rho_t^m}(x))^2.$$

Now we can expand

$$\begin{aligned}&\mathbb{E}_x(f_{\bar{\rho}_t^m}(x) - f_{\rho_t^m}(x))^2 \\ &\quad= \mathbb{E}_x\left(\mathbb{E}_i\sigma(\xi_t(w_i)^\top x) - \sigma((\xi_t(w_i) - \Delta_t(i)^\top x)\right)^2 \\ &\quad= \mathbb{E}_x\left(\mathbb{E}_i\sigma'(\xi_t(w_i)^\top x)x^\top\Delta_t(i) + \int_{s=0}^1\int_{s'=0}^1(\sigma'((\xi_t(w_i) + s'\Delta_t(i)^\top x)(x^\top\Delta_t(i))^2 drds\right)^2.\end{aligned}$$

Letting $\zeta(i,x) := \int_{s=0}^1\int_{s'=0}^s(\sigma'((\xi_t(w_i) + s'\Delta_t(i)^\top x)drds$, we have

$$\mathbb{E}_x(f_{\bar{\rho}_t^m}(x) - f_{\rho_t^m}(x))^2 \leq 2\mathbb{E}_x\left(\mathbb{E}_i\sigma'(\xi_t(w_i)^\top x)x^\top\Delta_t(i)\right)^2 + 2\mathbb{E}_x\left(\mathbb{E}_i(x^\top\Delta_t(i))^2\zeta(i,x)\right)^2 \quad (13)$$

and likewise,

$$\mathbb{E}_x\left(\mathbb{E}_i\sigma'(\xi_t(w_i)^\top x)x^\top\Delta_t(i)\right)^2 \le 2\mathbb{E}_x(f_{\bar\rho_t^m}(x)-f_{\rho_t^m}(x))^2 + 2\mathbb{E}_x\left(\mathbb{E}_i(x^\top\Delta_t(i))^2\zeta(i,x)\right)^2 \quad (14)$$

Let us bound the second term. We have

$$\begin{aligned}\mathbb{E}_x\left(\mathbb{E}_i(x^\top\Delta_t(i))^2\zeta(i,x)\right)^2 &= \mathbb{E}_i\mathbb{E}_j\mathbb{E}_x(x^\top\Delta_t(i))^2\zeta(i,x)(x^\top\Delta_t(j))^2\zeta(j,x)\\ &\le \mathbb{E}_i\mathbb{E}_j\left(\mathbb{E}_x((x^\top\Delta_t(i))^2)^4\right)^{1/4}\left(\mathbb{E}_x(\zeta(i,x))^4\right)^{1/4}\left(\mathbb{E}_x((x^\top\Delta_t(j))^2)^4\right)^{1/4}\left(\mathbb{E}_x(\zeta(j,x))^4\right)^1\\ &= \mathbb{E}_i\left(\mathbb{E}_x((x^\top\Delta_t(i))^2)^4\right)^{1/2}\left(\mathbb{E}_x(\zeta(i,x))^4\right)^{1/2}\\ &= \sqrt{105}\left(\mathbb{E}_i\|\Delta_t(i)\|^2\right)^2\left(\mathbb{E}_x(\zeta(i,x))^4\right)^{1/2}\end{aligned}$$

Now since for any $s'\in[0,1]$, we have that $\|\xi_t(w_i)+s'\Delta_t(i)\|\le 1$ (as it interpolates between two points on the sphere), we have by Assumption Regularity that

$$\mathbb{E}_x(\zeta(i,x))^4 \le (C_{\text{reg}}/11)^4.$$

$$\zeta_s(i,x) = \int_{r=0}^{s}\sigma''((\xi_t(w_i)+r\Delta_t(i)^\top x)dr,$$

and thus since $\|\xi_t(w_i)+r\Delta_t(i)\|\le 1$ (as it interpolates between two points on the sphere), we have by Assumption Regularity that

$$\mathbb{E}_x(\zeta_s(i,x))^4 \le (C_{\text{reg}}/11)^4,$$

and thus

$$\mathbb{E}_x\left(\mathbb{E}_i(x^\top\Delta_t(i))^2\zeta(i,x)\right)^2 \le \left(\mathbb{E}_i\|\Delta_t(i)\|^2\right)^2 C_{\text{reg}}^2/11.$$

Returning to Equations (13) and (14), and observing that $\mathbb{E}_x\left(\mathbb{E}_i\sigma'(\xi_t(w_i)^\top x)x^\top\Delta_t(i)\right)^2 - \Delta_t^\top H_t^\perp\Delta_t \le C_{\text{reg}}\left(\mathbb{E}_i\|\Delta_t(i)\|^2\right)^2$ (to account for the projections orthogonal to $\xi_t(w_i)$ in $H_t^\perp$; we omit the details), we have that

$$\begin{aligned}\mathbb{E}_x(f_{\bar\rho_t^m}(x)-f_{\rho_t^m}(x))^2 &\le 2\Delta_t^\top H_t^\perp\Delta_t + (2C_{\text{reg}}+C_{\text{reg}}^2/11)\left(\mathbb{E}_i\|\Delta_t(i)\|^2\right)^2\\ &\le (4C_{\text{reg}}+C_{\text{reg}}^2/11)\left(\mathbb{E}_i\|\Delta_t(i)\|\right)^2,\end{aligned}$$

and

$$\Delta_t^\top H_t^\perp\Delta_t \le 2\mathbb{E}_x(f_{\bar\rho_t^m}(x)-f_{\rho_t^m}(x))^2 + (C_{\text{reg}}+C_{\text{reg}}^2/11)\left(\mathbb{E}_i\|\Delta_t(i)\|^2\right)^2. \quad (15)$$

It follows that

$$\mathbb{E}_x(f_{\bar\rho_t^m}(x)-f_{\rho_t^m}(x))^2 \le C_{\text{reg}}^2\left(\mathbb{E}_i\|\Delta_t(i)\|\right)^2.$$

We will use Chebychev's inequality to bound the first term $\mathbb{E}_x(f_{\rho_t^{\text{MF}}}(x) - f_{\rho_t^m}(x))^2$. We have

$$\begin{aligned}
\mathbb{E}_{\rho_0^m \sim \rho_0^{\otimes m}} \left(\mathbb{E}_x(f_{\rho_t^{\text{MF}}}(x) - f_{\bar{\rho}_t^m}(x))^2\right)^2 &\leq \mathbb{E}_{\rho_0^m \sim \rho_0^{\otimes m}} \mathbb{E}_x(f_{\rho_t^{\text{MF}}}(x) - f_{\bar{\rho}_t^m}(x))^4 \\
&= \mathbb{E}_x \mathbb{E}_{\rho_0^m \sim \rho_0^{\otimes m}} (f_{\rho_t^{\text{MF}}}(x) - f_{\bar{\rho}_t^m}(x))^4 \\
&\leq \mathbb{E}_x \frac{1}{m^3} \mathbb{E}_{w \sim \rho_t^{\text{MF}}} (\sigma(w^\top x))^4 + \frac{O(m^2)}{m^4} \left(\mathbb{E}_{w \sim \rho_t^{\text{MF}}} (\sigma(w^\top x))^2\right)^2 \\
&\leq O\left(\frac{C_{\text{reg}}^4}{m^2}\right),
\end{aligned}$$

where in the final inequality we used Assumption **R1**. By Chebychev, we have

$$\mathbb{P}_{\rho_0^m \sim \rho_0^{\otimes m}} \left[\mathbb{E}_x(f_{\rho_t^{\text{MF}}}(x) - f_{\bar{\rho}_t^m}(x))^2 \geq \frac{\log(m)}{2m}\right] \leq o(1).$$

We thus conclude that with high probability,

$$\mathbb{E}_x(f_{\rho_t^{\text{MF}}}(x) - f_{\rho_t^m}(x))^2 \leq 2C_{\text{reg}}^2 (\mathbb{E}_i \|\Delta_t(i)\|)^2 + \frac{\log(m)}{m},$$

which yields Lemma 1.

For Lemma 16, we have by (15) that

$$\begin{aligned}
\Delta_t^\top H_t^\perp \Delta_t &\leq 2\mathbb{E}_x(f_{\bar{\rho}_t^m}(x) - f_{\rho_t^m}(x))^2 + C_{\text{reg}}^2 \left(\mathbb{E}_i \|\Delta_t(i)\|^2\right)^2 \\
&\leq 2\mathbb{E}_x(f_{\rho_t^{\text{MF}}}(x) - f_{\rho_t^m}(x))^2 + O\left(\frac{1}{m}\right) + C_{\text{reg}}^2 \left(\mathbb{E}_i \|\Delta_t(i)\|^2\right)^2.
\end{aligned}$$

■

Finally, we prove Lemma 5, which we restate here.

Lemma 22 (Parameter-Space Error Dynamics) *Suppose Regularity Assumption holds. With high probability, for all $t \leq T$ and $i \in [m]$,*

$$\frac{d}{dt}\Delta_t(i) = D_t^\perp(i)\Delta_t(i) - \mathbb{E}_{j \sim [m]} H_t^\perp(i,j)\Delta_t(j) + \boldsymbol{\epsilon}_{t,i},$$

where $\|\boldsymbol{\epsilon}_{t,i}\| \leq 2\epsilon_m + \epsilon_n + 2C_{reg}\left(\|\Delta_t(i)\|^2 + \mathbb{E}_j\|\Delta_t(j)\|^2\right)$.

Proof [Proof of Lemma 5] We first decompose $\frac{d}{dt}\Delta_t(i)$ into four terms:

$$\begin{aligned}
\frac{d}{dt}\Delta_t(i) &= \nu(\xi_t(w_i), \rho_t^{\text{MF}}) - \nu(\hat{\xi}_t(w_i), \rho_t^m) \\
&= \left(\nu(\xi_t(w_i), \rho_t^{\text{MF}}) - \nu(\xi_t(w_i), \bar{\rho}_t^m)\right) + \left(\nu(\xi_t(w_i), \bar{\rho}_t^m) - \nu(\xi_t(w_i), \rho_t^m)\right) \\
&\quad + \left(\nu(\xi_t(w_i), \rho_t^m) - \nu(\hat{\xi}_t(w_i), \rho_t^m)\right) + \left(\nu(\xi_t(w_i), \rho_t^m) - \nu_{\hat{\mathcal{D}}}(\hat{\xi}_t(w_i), \rho_t^m)\right).
\end{aligned}$$

By Lemma 23 and Lemma 27, we can bound the first and fourth terms respectively with high probability:

$$\begin{aligned}\|\nu(\xi_t(w_i),\rho_t^{\mathrm{MF}})-\nu(\xi_t(w_i),\bar{\rho}_t^m)\|_2 &\le \epsilon_m.\\ \nu(\xi_t(w_i),\rho_t^m)-\nu_{\hat{\mathcal{D}}}(\hat{\xi}_t(w_i),\rho_t^m) &\le \epsilon_n.\end{aligned} \tag{16}$$

For the second term, we have

$$\begin{aligned}\nu(\xi_t(w_i),\bar{\rho}_t^m)-\nu(\xi_t(w_i),\rho_t^m) &= -f'(\xi_t(w_i))+\mathbb{E}_{w'\sim\bar{\rho}_t^m}\hbar'(\xi_t(w_i),w')\\ &\quad + f'(\xi_t(w_i))-\mathbb{E}_{w'\sim\rho_t^m}\hbar'(\xi_t(w_i),w')\\ &= -\mathbb{E}_j\left(\hbar'(\xi_t(w_i),\xi_t(w_j))-\hbar'(\xi_t(w_i),\xi_t(w_j)+\Delta_t(j))\right)\\ &= \mathbb{E}_{j\sim[m]}\left(H_t(i,j)\Delta_t(j)+\mathbf{v}_j\right),\end{aligned}$$

where $\|\mathbf{v}_j\|\le C_{\mathrm{reg}}\|\Delta_t(j)\|^2$. Indeed we can plug Lemma 19 **S2** into the Lagrange error bound

$$\|\hbar'(w,w')-\hbar'(w,w'+\Delta)-\nabla_{w'}\hbar'(w,w')\Delta\|\le\|\Delta\|^2\sup_{w':\|w'\|\le 1}\left\|\nabla^2_{w'}\hbar'(w,w')\right\|.$$

Now note that for any j, since both $\xi_t(w_j)$ and $w_t^{(j)}$ are on $\mathbb{S}^{d-1}$, we have that

$$|\langle\xi_t(w_j)\Delta_t(j)\rangle|=\frac{1}{2}\|\Delta_t(j)\|^2, \tag{17}$$

and so by **S1**,

$$H_t(i,j)\Delta_t(j)=H_t^{\perp}(i,j)\Delta_t(j)+\mathbf{v}'_j$$

where $\|\mathbf{v}'_j\|_2\le\frac{1}{2}C_{\mathrm{reg}}\|\Delta_t(j)\|^2$ Summarizing, we have that

$$\nu(\xi_t(w_i),\bar{\rho}_t^m)-\nu(\xi_t(w_i),\rho_t^m)=\mathbb{E}_{j\sim[m]}\left(H_t^{\perp}(i,j)\Delta_t(j)+\frac{3}{2}\mathbf{v}_j\right). \tag{18}$$

Finally for the third term, we have

$$\nu(\xi_t(w_i),\rho_t^m)-\nu(\hat{\xi}_t(w_i),\rho_t^m)=-\nabla_w\nu(w,\rho_t^m)|_{w=\xi_t(w_i)}\Delta_t(i)+\mathbf{v},$$

where by **S6**,

$$\|\mathbf{v}\|\le\|\Delta_t(i)\|^2\left\|\nabla^2_w\nu(w,\rho_t^m)\right\|_{op}\le C_{\mathrm{reg}}\|\Delta_t(i)\|^2$$

Recall that we have defined

$$\bar{D}_t(w):=\nabla_{\xi_t(w)}\nu(\xi_t(w),\bar{\rho}_t^m)=\nabla_{\xi_t(w)}f'(\xi_t(w))-\mathbb{E}_{w'\sim\bar{\rho}_t^m}\nabla_{\xi_t(w)}\hbar'(\xi_t(w),w').$$

Now

$$\begin{aligned}\nabla_{\xi_t(w_i)}\nu(\xi_t(w_i),\rho_t^m) &= \nabla_{\xi_t(w_i)}f'(\xi_t(w_i))-\mathbb{E}_j\nabla_{\xi_t(w_i)}\hbar'(\xi_t(w_i),\hat{\xi}_t(w_j))\\ &= \nabla_{\xi_t(w_i)}f'(\xi_t(w_i))-\mathbb{E}_j\nabla_{\xi_t(w_i)}(\hbar'(\xi_t(w_i),\xi_t(w_j))+\mathbf{M}_j)\\ &= \bar{D}_t(i)-\mathbb{E}_j\mathbf{M}_j.\end{aligned}$$

where by S4,

$$\|\mathbf{M}_j\|_{op} \leq \|\Delta_t(j)\| \sup_{w,w'} \left\|\nabla_w \nabla_{w'} k'(w,w')\right\|_{op} \leq C_{\text{reg}} \|\Delta_t(j)\|.$$

Thus, additionally using the fact that we have conditioned on the fact that $\|D_t(i) - \bar{D}_t(i)\| \leq \epsilon_m$ — and thus $\|D_t^\perp(i) - \bar{D}_t^\perp(i)\| \leq \epsilon_m$ — and again using (17) and S1 to swap $D_t(i)\Delta_t(i)$ for $D_t^\perp(i)\Delta_t(i)$ with an error term of magnitude $0.5C_{\text{reg}}\|\Delta_t(i)\|^2$, we have that

$$\nu(\xi_t(w_i), \rho_t^m) - \nu(\hat{\xi}_t(w_i), \rho_t^m) = D_t^\perp(i)\Delta_t(i) + \mathbf{v}_3, \tag{19}$$

where $\|\mathbf{v}_3\| \leq C_{\text{reg}}(1.5\|\Delta_t(i)\|^2 + \|\Delta_t(i)\|\mathbb{E}_j\|\Delta_t(j)\|) + \epsilon_m\|\Delta_t(i)\|$.

Putting together Equations (16), (18), and (19), we have

$$\frac{d}{dt}\Delta_t(i) = D_t^\perp(i)\Delta_t(i) - \mathbb{E}_{j\sim[m], j\neq i} H_t^\perp(i,j)\Delta_t(j) + \boldsymbol{\epsilon},$$

where

$$\begin{aligned}\|\boldsymbol{\epsilon}\| &\leq \epsilon_n + \epsilon_m(1 + \|\Delta_t(i)\|) + C_{\text{reg}}\left(1.5\|\Delta_t(i)\|^2 + \|\Delta_t(i)\|\mathbb{E}_j\|\Delta_t(j)\| + 1.5\mathbb{E}_j\|\Delta_j\|^2\right) \\ &\leq \epsilon_n + \epsilon_m(1 + \|\Delta_t(i)\|) + 2C_{\text{reg}}\left(\|\Delta_t(i)\|^2 + \mathbb{E}_j\|\Delta_j\|^2\right).\end{aligned}$$

■

Appendix E. Proof of Concentration Lemmas

Lemma 23 (Uniformly Bounded Sampling Error) *With probability $1 - o(1)$ over the initialization, for all $t \leq T$ and $i \in [m]$, the following holds with $\epsilon_m = \frac{d^{3/2}\log(Tm)}{\sqrt{m}}$.*

$$\begin{aligned}\|\nu(\xi_t(w_i), \rho_t^{\text{MF}}) - \nu(\xi_t(w_i), \bar{\rho}_t^m)\| &\leq \epsilon_m. \\ \|D_t(i) - \bar{D}_t(i)\| &\leq \epsilon_m.\end{aligned}$$

Proof [Proof of Lemma 23] Fix $t \leq T$ and $w \in \mathbb{S}^{d-1}$ By Equation (2), we have that

$$\nu(w, \rho_t^{\text{MF}}) - \nu(w, \bar{\rho}_t^m) := (I - ww^\top)\left(\mathbb{E}_{w'\sim\rho_t^{\text{MF}}}\nabla_w K(w,w') - \mathbb{E}_{w'\sim\bar{\rho}_t^m}\nabla_w K(w,w')\right)$$

Now

$$\mathbb{E}_{w_0\sim\rho_t^0}\mathbb{E}_{w'\sim\rho_t^{\text{MF}}}\nabla_w K(w,w') = \mathbb{E}_{w'\sim\rho_t^{\text{MF}}}\mathbb{E}_{w'\sim\rho_t^{\text{MF}}}\nabla_w K(w,w'),$$

and by Assumption R1, for any $w', w \in \mathbb{S}^{d-1}$, $\|\nabla_w K(w,w')\|_\infty \leq C_{\text{reg}}$. So by Hoeffding's inequality, taking a union bound over all d coordinates in the random vector, we have

$$\mathbb{P}\left[\|\nu(w, \rho_t^{\text{MF}}) - \nu(w, \bar{\rho}_t^m)\| \geq \frac{\epsilon_m}{2}\right] \leq 2d\exp\left(-\frac{\Omega(m\epsilon_m^2)}{4dC_{\text{reg}}^2}\right)$$

Now we need to take a union bound over all $w \in \mathbb{S}^{d-1}$, and $t \le T$. Create an net over $\mathbb{S}^{d-1}$ of maximum distance $\frac{\epsilon_m}{4C_{\text{reg}}}$ between any point and the net: this has size $O\left(\left(\frac{4C_{\text{reg}}}{\epsilon_m}\right)^d\right)$. Similarly make a net over $[0,T]$ of spacing $\frac{\epsilon_m}{4C_{\text{reg}}}$; this has size $\frac{4C_{\text{reg}}T}{\epsilon_m}$. By a union bound, with probability at least

$$1 - 2d\exp\left(-\frac{\Omega(m\epsilon_m^2)}{4dC_{\text{reg}}^2}\right) O\left(\left(\frac{4C_{\text{reg}}}{\epsilon_m}\right)^d\right)\frac{4C_{\text{reg}}T}{\epsilon_m},$$

for any w in the net and any t in the net, we have

$$\|\nu(w,\rho_t^{\text{MF}}) - \nu(w,\bar{\rho}_t^m)\| \le \frac{\epsilon_m}{3C_{\text{reg}}}.$$

For any $w,u \in \mathbb{S}^{d-1}$, for any ρ, by Lemma 19, we have

$$\nu(w,\rho) - \nu(u,\rho) \le C_{\text{reg}}\|w-u\|.$$

Similarly, by Lemma 19, for any $s,t \le T$, and any w_0, we have

$$\|\xi_t(w_0) - \xi_s(w_0)\| \le C_{\text{reg}}|t-s|.$$

Thus, for any $w \in \mathbb{S}^{d-1}$ and $t \le T$, there exist u and s in the respective nets of distance at most $\frac{\epsilon_m}{3C_{\text{reg}}}$. By a standard triangle inequality argument, we attain that with the probability in Equation E, for all $w \in \mathbb{S}^{d-1}$ and $t \le T$, we have

$$\|\nu(w,\rho_t^{\text{MF}}) - \nu(w,\bar{\rho}_t^m)\| \le \epsilon_m.$$

One can check that since $\epsilon_m \ge \frac{d\log(mT)}{\sqrt{m}}$, this probability is $1-o(1)$.

The argument for proving concentration for $\bar{D}_t(w)$ uniformly over w and t is identical. The only change is that since $\bar{D}_t(w)$ is a $d\times d$ matrix, we need to take a union bound over d^2 indices in this matrix, so we require that $\epsilon_m \ge \frac{d^{3/2}\log(mT)}{\sqrt{m}}$. ■

Lemma 24 (Concentration of $J_{t,s}$) *With high probability over the random choice of $\bar{\rho}_0^m$, for all $s \le t \le T$, all vectors $v \in \mathbb{S}^{d-1}$, and all $j \in [m]$, we have*

$$\left|\mathbb{E}_i\|J_{t,s}(i)H_s^{\perp}(i,j)v\|\mathbf{1}(\xi_t(w_i)\in S) - \mathbb{E}_{w\sim\rho_0}\|J_{t,s}(w)H_s^{\perp}(w,\bar{w}_0(j))v\|\mathbf{1}(\xi_t(w)\in S)\right| \le \epsilon_m,$$

for $\epsilon_m = \frac{\sqrt{d}J_{max}\log(mT)}{\sqrt{m}}$.

Proof [Proof of Lemma 24] Fix $w', v \in \mathbb{S}^{d-1}$ and $s \le t \le T$. Let

$$X(w) := \|J_{t,s}(w)H_s^{\perp}(w,w')v\|\mathbf{1}(\xi_t(w)\in S).$$

To prove the desired bound for j we must bound $\left|\mathbb{E}_{w\sim\rho_0^m}X(w) - \mathbb{E}_{w\sim\rho_0}X(w)\right|$ with high probability for $w' = \bar{w}_0(j)$.

By Lemma 19, we have $|X(w)| \le C_{\text{reg}} J_{\max}$. By Hoeffding's inequality, we have

$$\mathbb{P}\left[\left|\mathbb{E}_{w\sim\rho_0^m} X(w) - \mathbb{E}_{w\sim\rho_0} X(w)\right| \ge \frac{\epsilon_m}{2}\right] \le 2\exp\left(\frac{\Omega(m\epsilon_m^2)}{4C_{\text{reg}}^2 J_{\max}^2}\right).$$

Now we need to build an ϵ-net of scale $\frac{\epsilon_m}{6C_{\text{reg}}}$ over $s,t \in [0,T]$, $w' \in \mathbb{S}^{d-1}$, and $v \in \mathbb{S}^{d-1}$. The product of the size of these nets is

$$\left(\frac{6TC_{\text{reg}}}{\epsilon_m}\right)^2 O\left(\left(\frac{6C_{\text{reg}}}{\epsilon_m}\right)^{2d}\right)$$

Checking Lipschitzness of the various quantities as per the proof of Lemma 23, and then using a union bound gives the desired result with high probability whenever $\epsilon_m \ge \frac{\sqrt{d} J_{\max}\log(mT)}{\sqrt{m}}$. ■

Lemma 25 *Fix a set $S \subseteq \mathbb{S}^{d-1}$, any function $v : \mathbb{S}^{d-1} \to B_2^d$. With probability $1 - o(1/d)$ over the random choice of ρ_0^m, for any $w \in \mathbb{S}^{d-1}$, with $\epsilon_m^{25} = \frac{d\log(md)}{\sqrt{m}}$ we have*

$$\|\mathbb{E}_{w'\sim\rho_0} H_\infty^\perp(w,w')v(w')\mathbf{1}(\xi_t(w') \in S) - \mathbb{E}_{w'\sim\rho_0^m} H_\infty^\perp(w,w')v(w')\mathbf{1}(\xi_t(w') \in S)\| \le \|v\|_\infty \epsilon_m^{25}$$

$$\left|\mathbb{P}_{w'\sim\rho_0}[\xi_t(w') \in S] - \mathbb{P}_{w'\sim\rho_0^m}[\xi_t(w') \in S]\right| \le \epsilon_m^{25}.$$

Proof The second statement is immediate by a Chernoff bound. For the first statement, the proof is similar to the other concentration lemmas. Fix w. Let

$$X(w') := H_\infty^\perp(w,w')v(w')\mathbf{1}(\xi_t(w') \in S)$$

Since $\|H_\infty^\perp(w,w')\| \preceq C_{\text{reg}} I$ for all w, w', we have the following bound:

By Hoeffding's inequality (unioning over all coordinates of $X(w')$), we have

$$\mathbb{P}\left[\left\|\mathbb{E}_{w\sim\rho_0^m} X(w') - \mathbb{E}_{w\sim\rho_0} X(w')\right\| \ge \frac{\epsilon_m^{25}}{2}\right] \le 2\exp\left(\frac{\Omega(m{\epsilon_m^{25}}^2)}{4C_{\text{reg}}^2 d}\right).$$

We need to build an ϵ-net of scale $\frac{\epsilon_m^{25}}{4C_{\text{reg}}}$ over $w \in \mathbb{S}^{d-1}$ since by Lemma 19, $X(w)$ is C_{reg}-Lipschitz in w. This net has size $\left(\left(\frac{O(C_{\text{reg}})}{\epsilon_m}\right)^d\right)$. Thus with $\epsilon_m^{25} = \frac{d\log(m)}{\sqrt{m}}$, we have that with high probability, for all $w \in \mathbb{S}^{d-1}$, the desired quantity is uniformly bounded. ■

Lemma 26 (Averaging Lemma) *Suppose $\mathcal{Q}$ is C_b-balanced, and the high probability event in Lemma 24 holds for $S = B_\tau$. If Assumption Stability holds, then for any $s \le t$,*

$$\mathbb{E}_i\|J_{t,s}(i)m_s(i)\|\mathbf{1}(\xi_t(w_i) \notin B_\tau) \le (1+C_b)\left(\epsilon_m^{24} + J_{avg}(\tau)\right)\Phi_{\mathcal{Q}}(s).$$

In particular,

$$\mathbb{E}_i\|m_t(i)\|\mathbf{1}(\xi_t(w_i) \notin B_\tau) \le (1+C_b)\left(\epsilon_m^{24} + J_{avg}(\tau)\right)\Phi_{\mathcal{Q}}(t).$$

Proof Recall that

$$m_t(i) = \mathbb{E}_j H_t^{\perp}(i,j)\Delta_t(j).$$

Thus

$$\|J_{t,s}(i)m_s(i)\| \le \mathbb{E}_j \|J_{t,s}(i)H_s^{\perp}(i,j)\Delta_s(j)\|.$$

Now for any vector $v \in \mathbb{R}^d$, by Lemma 24 and Assumption Stability, we have that

$$\mathbb{E}_i \|J_{t,s}(i)H_s^{\perp}(i,j)v\| \le \epsilon_m^{24}\|v\| + J_{\text{avg}}(\tau)\|v\|,$$

and so

$$\mathbb{E}_i \|J_{t,s}(i)m_s(i)\|\mathbf{1}(\xi_t(w_i) \notin B_\tau) \le \left(\epsilon_m^{24} + J_{\text{avg}}(\tau)\right)\mathbb{E}_i\|\Delta_s(i)\| \le \left(\epsilon_m^{24} + J_{\text{avg}}(\tau)\right)\Phi_{\mathcal{Q}}(s).$$

The second line of the lemma holds by plugging in $s = t$. This concludes the lemma. ■

Lemma 27 *Suppose the empirical data distribution $\hat{D} = \sum_{i=1}^n \delta_{(x_i,y_i)}$ satisfies Assumption **R2**. Then with high probability over the draw of $\hat{D}$, we have uniformly over all $w \in \mathbb{S}^{d-1}$, and all $\rho \in \Delta(\mathbb{S}^{d-1})$, we have*

$$\|\nu_{\hat{\mathcal{D}}}(w,\rho) - \nu(w,\rho)\| \le \epsilon_n,$$

for $\epsilon_n = \frac{\sqrt{d}\log^2(n)}{\sqrt{n}}$.

Proof The velocity is linear in ρ, so it suffices to prove that (additionally) uniformly over w', we have

$$\|\nu_{\hat{\mathcal{D}}}(w,\delta_{w'}) - \nu(w,\delta_{w'})\| \le \epsilon_n.$$

We expand

$$\nu_{\hat{\mathcal{D}}}(w,\delta_{w'}) = (I - ww^\top)\mathbb{E}_{x\sim\hat{\mathcal{D}}}(y - \sigma(w'^\top x))\sigma'(w^\top x)x;$$

it suffices to prove that with high probability, uniformly over $w' \in \mathbb{S}^{d-1}$, and $v \in \mathbb{S}^{d-1}$, we have

$$\left|\mathbb{E}_{x\sim\hat{\mathcal{D}}}\sigma(w'^\top x)\sigma'(w^\top x)x^\top v - \mathbb{E}_{x\sim\mathcal{D}}\sigma(w'^\top x)\sigma'(w^\top x)x^\top v\right| \le \epsilon_n$$
$$\left|\mathbb{E}_{x\sim\hat{\mathcal{D}}} y\sigma'(w^\top x)x^\top v - \mathbb{E}_{x\sim\mathcal{D}} y\sigma'(w^\top x)x^\top v\right| \le \epsilon_n$$

For a fixed w, w', v, since by Assumption **R1**, all the terms in side the expectations are C_{reg}-subgaussian, this holds with probability $\exp(-n\epsilon_n^2/2C_{\text{reg}}^2)$. We now take three epsilon-nets over $\mathbb{S}^{d-1}$ (for w, w' and v respectively) at the scale $\frac{\epsilon_n}{6C_{\text{reg}}}$. Note that Lemma 19 implies these quantities are C_{reg}-Lipschitz with regard to w, w' or v. Since the epsilon nets have size $\left(O\left(\frac{C_{\text{reg}}}{\epsilon_n}\right)\right)^d$, with $\epsilon_n = \frac{\sqrt{d}\log^2(n)}{\sqrt{n}}$, we see that

$$\exp(-n\epsilon_n^2/2C_{\text{reg}}^2)\left(O\left(\frac{C_{\text{reg}}}{\epsilon_n}\right)\right)^{3d} = o(1).$$

■

Appendix F. Proof of Results Relating to Potential Function Analysis

F.1. Notations

For $g, h : \mathcal{X} \to \mathbb{R}^d$, and a set $S \subseteq \mathbb{S}^{d-1}$ we will denote the dot product and conditional dot products

$$\langle g, h \rangle = \mathbb{E}_{w \sim \rho_0} g(w)^\top h(w).$$
$$\langle g, h \rangle_S = \mathbb{E}_{w \sim \rho_0} g(w)^\top h(w) \mathbf{1}(w \in S).$$

For a kernel $H : (\mathbb{S}^{d-1})^2 \to \mathbb{R}^{d \times d}$, and two sets $S, T \subseteq \mathbb{S}^{d-1}$, for $g, h : \mathbb{S}^{d-1} \to \mathbb{R}^d$, we use the notation

$$\langle g, h \rangle_H^{S,T} := \mathbb{E}_{w,w' \sim \rho_0} g(w)^\top H(w, w') h(w') \mathbf{1}(w \in S, w' \in T).$$

If $S = T$ or $S = T = \mathbb{S}^{d-1}$, we will abbreviate and use the notation $\langle g, h \rangle_H^S$ or $\langle g, h \rangle_H$ respectively. If the functions g, h are only defined on $[m]$ (or respectively on $\mathrm{supp}(\rho_0^m)$), then in all the inner products / quadratic forms above, the default distribution should be taken to be Uniform$([m])$ (resp. ρ_0^m) instead of ρ_0.

We will use $\nabla \Phi_{\mathcal{Q}}(t)$ (resp. $\nabla \Omega(t)$, $\nabla \phi_v(t)$, $\nabla \Psi_{\mathcal{Q}}(t)$.) to denote the map on $[m]$ (resp. $\mathrm{supp}(\rho_0^m)$) which takes i (or w_i) to $m \nabla_{\Delta_t(i)} \Phi(t)$. We have rescaled these derivative so that this term is on order 1, so we can take inner products in our notation more easily.

For a set $B \subseteq \mathbb{S}^{d-1}$, we will use the shorthand $B^t := \xi_t^{-1}(B)$ to denote the set of all $w \in \mathbb{S}^{d-1}$ with $\xi_t(w) \in B$, and $\bar{B}$ to denote the complement of B in $\mathbb{S}^{d-1}$.

F.2. Proof of Lemmas on the Properties of the Potential

F.2.1. Restricted Isometry and Related Group Theoretic Definitions and Lemmas

Definition 28 *We say a problem (H, μ) has* consistent restricted isometry (CRI) *with a set S if for any eigenfunction v of (H, μ), (that is, where $\langle u, v \rangle_H = \lambda_v \langle u, v \rangle$ for all $u : \mathbb{S}^{d-1} \to \mathbb{R}^d$), we have that for all $w \in \mathbb{S}^{d-1}$, we have*

$$\mathbb{E}_{w' \sim \mu} H(w, w') v(w') \mathbf{1}(w' \in S) = \lambda_v v(w) \mathbb{P}_{w' \sim \rho_0}[w' \in S].$$

In other words, for any $u : \mathbb{S}^{d-1} \to \mathbb{R}^d$,

$$\langle u, v \rangle_H^S = \lambda_v \langle u, v \rangle^S \mathbb{P}_{\rho_0}[S],$$

Definition 29 *The* automorphism group $\mathcal{G}$ *of a problem $(\rho^*, \rho_0, \mathcal{D}_x)$ is the set group of rotations g on $\mathbb{S}^{d-1}$ where for any $A \subset \mathbb{S}^{d-1}$:*

$$\mathbb{P}_{\rho^*}[A] = \mathbb{P}_{\rho^*}[g(A)]$$
$$\mathbb{P}_{\mathcal{D}}[A] = \mathbb{P}_{\mathcal{D}_x}[g(A)]$$
$$\mathbb{P}_{\rho_0}[A] = \mathbb{P}_{\rho_0}[g(A)]$$

We say that a problem $(\rho^*, \mathcal{D}_x, \rho_0)$ is transitive *if for any $w^*, w^{*\prime} \in \mathrm{supp}(\rho^*)$, there exists some g in the automorphism group $\mathcal{G}$ such that $g(w^*) = w^{*\prime}$.*

Lemma 30 *Suppose I1 holds. For any time t, for all $g \in \mathcal{G}$ in the automorphism group of $(\rho^*, \rho_0, \mathcal{D}_x)$, we have*

A1 *If $\xi_t(w) \in A$, then $\xi_t(g(w)) \in g(A)$*

A2 *Almost surely over $w \sim \rho_0$, $\xi^\infty(w) = \operatorname{argmin}_{w^* \in \operatorname{supp}(\rho^*)} \|w - w^*\|$. So a.s., for all $A \subset \mathbb{S}^{d-1}, g \in \mathcal{G}$, if $\xi^\infty(w) \in A$, then $\xi^\infty(g(w)) \in g(A)$. Further, $\xi^\infty_\# \rho_0 = \rho^*$.*

A3 $g(B_\tau) = B_\tau$.

Proof We will prove the first item by induction. It suffices to prove the following claim, because if the velocity is symmetric, then ρ_t^{MF} will be symmetric.

Claim 1 *Conditional on A1 holding up to time t, we have*

$$\frac{d}{dt}\xi_t(w) = \nu(w, \rho_t^{\mathrm{MF}}) = g^{-1}(\nu(g(w), \rho_t^{\mathrm{MF}}))$$

Proof

$$\begin{aligned}
\nu(w, \rho_t^{\mathrm{MF}}) &= -(I - ww^\top)\nabla_w F_{\mathcal{D}}(w) + (I - ww^\top)\nabla_w \mathbb{E}_{w' \sim \rho_t^{\mathrm{MF}}} K_{\mathcal{D}}(w, w') \\
&= -(I - ww^\top)\mathbb{E}_{x \sim \mathcal{D}_x} f^*(x)\sigma'(w^\top x)x + (I - ww^\top)\mathbb{E}_{w' \sim \rho_t^{\mathrm{MF}}}\mathbb{E}_{x \sim \mathcal{D}_x}\sigma(w'^\top x)\sigma'(w^\top x)x
\end{aligned}$$

Now

$$\begin{aligned}
&P_w^\perp \mathbb{E}_{w' \sim \rho_t^{\mathrm{MF}}}\mathbb{E}_{x \sim \mathcal{D}_x}\sigma(w'^\top x)\sigma'(w^\top x)x \\
&\qquad = P_w^\perp \mathbb{E}_{w' \sim \rho_t^{\mathrm{MF}}}\mathbb{E}_{x \sim \mathcal{D}_x}\sigma(g(w')^\top g(x))\sigma'(g(w)^\top g(x))x \\
&\qquad = P_w^\perp \mathbb{E}_{w' \sim \rho_t^{\mathrm{MF}}}\mathbb{E}_{x \sim \mathcal{D}_x}\sigma(g(w')^\top x)\sigma'(g(w)^\top x)g^{-1}(x) \\
&\qquad = P_w^\perp \mathbb{E}_{w' \sim \rho_t^{\mathrm{MF}}}\mathbb{E}_{x \sim \mathcal{D}_x}\sigma(w'^\top x)\sigma'(g(w)^\top x)g^{-1}(x) \\
&\qquad = (g^{-1}(x) - ww^\top g^{-1}(x))\mathbb{E}_{w' \sim \rho_t^{\mathrm{MF}}}\mathbb{E}_{x \sim \mathcal{D}_x}\sigma(w'^\top x)\sigma'(g(w)^\top x) \\
&\qquad = (g^{-1}(x) - wg(w)^\top x)\mathbb{E}_{w' \sim \rho_t^{\mathrm{MF}}}\mathbb{E}_{x \sim \mathcal{D}_x}\sigma(w'^\top x)\sigma'(g(w)^\top x) \\
&\qquad = g^{-1}(x - g(w)g(w)^\top x)\mathbb{E}_{w' \sim \rho_t^{\mathrm{MF}}}\mathbb{E}_{x \sim \mathcal{D}_x}\sigma(w'^\top x)\sigma'(g(w)^\top x) \\
&\qquad = g^{-1}\left(P_{g(w)}^\perp \nabla_{g(w)}\mathbb{E}_{w' \sim \rho_t^{\mathrm{MF}}} K_{\mathcal{D}}(g(w), w')\right).
\end{aligned}$$

Here (1) is because $z^\top y = z^\top U^\top U y$ for any rotation U and any $y, z \in \mathbb{R}^d$ (2) is because $\mathcal{D}_x$ is invariant with respect to $\mathcal{G}$, (3) is because ρ_t^{MF} is invariant with respect to $\mathcal{G}$ (by induction), (5) again because of the same reason as (1), and (4), (6) and (7) are simple algebraic operations. Similarly,

we can show that

$$\begin{aligned}
P_w^{\perp}\mathbb{E}_{x\sim\mathcal{D}_x} f^*(x)\sigma'(w^\top x)x &= \mathbb{E}_{x\sim\mathcal{D}_x} f^*(x)\sigma'(w^\top x)P_w^{\perp}x \\
&= \mathbb{E}_{x\sim\mathcal{D}_x} f^*(x)\sigma'(w^\top x)g^{-1}(P_{g(w)}^{\perp})g(x) \\
&= \mathbb{E}_{x\sim\mathcal{D}_x} f^*(x)\sigma'(g(w)^\top g(x))g^{-1}(P_{g(w)}^{\perp}g(x)) \\
&= \mathbb{E}_{x\sim\mathcal{D}_x}\mathbb{E}_{w^*\sim\rho^*}\sigma(w^{*\top}x)\sigma'(g(w)^\top g(x))g^{-1}(P_{g(w)}^{\perp}g(x)) \\
&= \mathbb{E}_{x\sim\mathcal{D}_x}\mathbb{E}_{w^*\sim\rho^*}\sigma(w^{*\top}g^{-1}(x))\sigma'(g(w)^\top x)g^{-1}(P_{g(w)}^{\perp}x) \\
&= \mathbb{E}_{x\sim\mathcal{D}_x}\mathbb{E}_{w^*\sim\rho^*}\sigma(g(w^*)^\top x)\sigma'(g(w)^\top x)g^{-1}(P_{g(w)}^{\perp}x) \\
&= \mathbb{E}_{x\sim\mathcal{D}_x} f^*(x)\sigma'(g(w)^\top x)g^{-1}(P_{g(w)}^{\perp}x) \\
&= g^{-1}\left(P_{g(w)}^{\perp}\mathcal{F}_{\mathcal{D}}(g(w))\right).
\end{aligned}$$

Putting these two computations together yields the desired conclusion,

$$\nu(w,\rho_t^{\mathrm{MF}}) = g^{-1}(\nu(g(w),\rho_t^{\mathrm{MF}})).$$

■

Next consider **A2**. Observe that if w is closest to some w^*, then it either is the case that $\xi_t(w^*)$ is always closest to w^*, or at some point there is a tie in the distances $\xi_t(w^*)$ and $\xi_t(w^{*\prime})$. By **A1**, such a tie would imply however that $\|w-w^*\| = \|w-w^{*\prime}\|$, which we have assumed in **I1** is a measure 0 event. The rest follows immediately from the transitivity of $\mathrm{supp}(\rho^*)$.

Finally for **A3**,

$$\begin{aligned}
g(B_\tau) &= \{g(w) : w \in B_\tau\} \\
&= \{g(w) : \min_{w^*\in\mathrm{supp}(w^*)} \|w-w^*\| \le \tau\} \\
&= \{g(w) : \min_{w^*\in\mathrm{supp}(w^*)} \|g(w)-g(w^*)\| \le \tau\} \\
&= \{g(w) : \min_{w^*\in\mathrm{supp}(w^*)} \|g(w)-w^*\| \le \tau\} \\
&= \{w : \min_{w^*\in\mathrm{supp}(w^*)} \|w-w^*\| \le \tau\} \\
&= B_\tau.
\end{aligned}$$

■

Lemma 31 *Suppose $(\rho^*,\mathcal{D}_x,\rho_0)$ is transitive. Then $(H_\infty^{\perp},\rho_0)$ has consistent restricted isometry with $B_\tau^t = \xi_t^{-1}(B_\tau)$ for any $t \le T$, $\tau \ge 0$.*

Proof We will use a series of small claims.

Claim 2 *Fix any t and τ. Let $\tilde{\rho}$ be the distribution of $\xi^\infty(w)$ with $w \sim \rho_0$* conditional *on $\xi_t(w) \in B_\tau$. Then*

$$\tilde{\rho} \sim {\xi^\infty}_{\#}\rho_0.$$

Proof We will show that both $\tilde{\rho}$ and ${\xi^\infty}_\# \rho_0$ are uniform on the support of ρ^*. Fix $w^*, w^{*\prime} \in \text{supp}(\rho^*)$, and let $g \in \mathcal{G}$ be the element in the automorphism group of $(\rho^*, \rho_0, \mathcal{D}_x)$ which takes w^* to $w^{*\prime}$. Now

$$\begin{aligned}
\tilde{\rho}(w^*) &= \mathbb{P}_{w\sim\rho_0}[\xi^\infty(w) = w^* \wedge \xi_t(w) \in B_\tau] \\
&= \mathbb{P}_{w\sim\rho_0}[\xi^\infty(g(w)) = g(w^*) \wedge \xi_t(g(w)) \in g(B_\tau)] \\
&= \mathbb{P}_{w\sim\rho_0}[\xi^\infty(g(w)) = w^{*\prime} \wedge \xi_t(g(w)) \in B_\tau] \\
&= \mathbb{P}_{w\sim\rho_0}[\xi^\infty(w) = w^{*\prime} \wedge \xi_t(w) \in B_\tau] \\
&= \tilde{\rho}(w^{*\prime}).
\end{aligned}$$

Here (1) is by definition, (2) is by **A1**, and **A2**, (3) is by choice of g and **A3**, and (4) is by the symmetry of ρ_0 with respect to $\mathcal{G}$. It follows that $\tilde{\rho}$ is uniform on the support of ρ^*. Now lets check that ${\xi^\infty}_\# \rho_0$ is also uniform on $\text{supp}(\rho^*)$. We have by similar use of **A1** and **A2** that

$$\begin{aligned}
{\xi^\infty}_\#\rho_0(w^*) &= \mathbb{P}_{w\sim\rho_0}[\xi^\infty(w) = w^* \wedge \|\xi^\infty(w), w^*\| \le \|\xi^\infty(w), \tilde{w}^*\| \forall \tilde{w}^* \in \text{supp}(\rho^*)] \\
&= \mathbb{P}_{w\sim\rho_0}[\xi^\infty(g(w)) = g(w^*) \wedge \|\xi^\infty(g(w)), g(w^*)\| \le \|\xi^\infty(g(w)), g(\tilde{w}^*)\| \forall \tilde{w}^* \in \text{supp}(\rho^*)] \\
&= \mathbb{P}_{w\sim\rho_0}[\xi^\infty(g(w)) = w^{*\prime} \wedge \|\xi^\infty(g(w)), w^{*\prime}\| \le \|\xi^\infty(g(w)), \tilde{w}^*\| \forall \tilde{w}^* \in \text{supp}(\rho^*)] \\
&= \mathbb{P}_{w\sim\rho_0}[\xi^\infty(w) = w^{*\prime} \wedge \|\xi^\infty(w), w^{*\prime}\| \le \|\xi^\infty(w), \tilde{w}^*\| \forall \tilde{w}^* \in \text{supp}(\rho^*)] \\
&= {\xi^\infty}_\#\rho_0(w^{*\prime}).
\end{aligned}$$

■

Claim 3 *Let v be an eigenfunction of $(H_\infty^\perp, \rho_0)$, that is $\langle u, v\rangle_{H_\infty^\perp} = \lambda_v \langle u, v\rangle$ for all $u : \mathbb{S}^{d-1} \to \mathbb{R}^d$. Then $v(w) = v'(\xi^\infty(w))$ for some function $v' : \text{supp}(\rho_\infty^{\text{MF}}) \to \mathbb{S}^{d-1}$.*

Proof For all w, we have

$$\lambda_v v(w) = \mathbb{E}_{w'\sim\rho_0} H_\infty^\perp(w, w')v(w') = \mathbb{E}_{w'\sim\rho_0} K'(\xi^\infty(w), \xi^\infty(w'))v'(\xi^\infty(w')).$$

This value only depends on w through $\xi^\infty(w)$. ■

We will now use the previous two claims to show consistency. Fix t and τ, and let v be some eigenfunction of $(H_\infty^\perp, \rho_0)$. Let $v' : \text{supp}(\rho_\infty^{\text{MF}}) \to \mathbb{S}^{d-1}$ be the function guaranteed by the previous claim with $v(w) = v'(\xi^\infty(w))$. Then for all w,

$$\begin{aligned}
&\mathbb{E}_{w'\sim\rho_0} H_\infty^\perp(w, w')v(w')\mathbf{1}(w' \in \xi_t^{-1}(B_\tau)) \\
&\quad= \mathbb{E}_{w'\sim\rho_0} H_\infty^\perp(w, w')v(w')\mathbf{1}(\xi_t(w') \in B_\tau) \\
&\quad= \mathbb{E}_{w'\sim\rho_0} K'(\xi^\infty(w), \xi^\infty(w'))v'(\xi^\infty(w'))\mathbf{1}(\xi_t(w') \in B_\tau) \\
&\quad= \mathbb{P}_{\rho_0}[\xi_t^{-1}(B_\tau)]\mathbb{E}_{w'\sim\rho_0} K'(\xi^\infty(w), \xi^\infty(w'))v'(\xi^\infty(w')) \\
&\quad= \mathbb{P}_{\rho_0}[\xi_t^{-1}(B_\tau)]\mathbb{E}_{w\sim\rho_0} H_\infty^\perp(w, w')v(w') \\
&\quad= \mathbb{P}_{\rho_0}[\xi_t^{-1}(B_\tau)]\lambda_v v(w),
\end{aligned}$$

as desired. Here the third equality follows from Claim 2. ■

F.2.2. CONSTRUCTION OF THE POTENTIAL

Remark 32 *We can verify that the action $\overline{H^{\perp}_{\infty}}$ (from Section 4.1) is well-defined in $\mathcal{Z}$ since $\|\overline{H^{\perp}_{\infty}}v\|_{\mathcal{Z}} \leq \sup_{w,w'} \|H^{\perp}_{\infty}(w,w')\|\|v\|_{\mathcal{Z}}$). We verify that $\overline{H^{\perp}_{\infty}}$ is self-adjoint in $\mathcal{Z}$, ie $\langle v, \overline{H^{\perp}_{\infty}}v'\rangle_{\mathcal{Z}} = \langle \overline{H^{\perp}_{\infty}}v, v'\rangle_{\mathcal{Z}}$. We also verify that the span of $\overline{H^{\perp}_{\infty}}$ is finite-dimensional, thanks to the atomic nature of ρ^*. Indeed, for each $w^* \in supp(\rho^*)$ and $l \in \{1,d\}$, let $\chi_{w^*,l} \in \mathcal{Z}$ be the indicator $\chi_{w^*}(w) = e_l \mathbf{1}(\xi^{\infty}(w) = w^*)$, where e_l is the l-th canonical basis vector. We verify that if $v \perp \mathcal{W} := span(\chi_{w^*,l};\, w^* \in supp(\rho^*), l \in \{1,d\})$, then $\overline{H^{\perp}_{\infty}}v = 0$.*

The following lemma implies Lemma 12. Recall that $C_{\rho^*} = \min\left(|\mathrm{supp}(\rho^*)|, \dim(\rho^*)^{2\,\mathrm{degree}(\sigma)+1}\right)$.

Lemma 33 *Suppose Assumption I2 holds. Then for any μ, there exists an balanced spectral distribution $\mathcal{Q}$ of $(H^{\perp}_{\infty}, \mu)$ which is $\frac{2C_{\rho^*}}{\min_{w^*\in\mathrm{supp}(\rho^*)} \mathbb{P}_{\xi^{\infty}_{\#}\mu}[w^*]}$ balanced. If I1 additionally holds, then there exists an balanced spectral distribution $\mathcal{Q}$ of $(H^{\perp}_{\infty}, \rho_0)$ which is $2C_{\rho^*}$-balanced.*

Proof [Proof of Lemma 33] We will show that the linear operator induced by $(H^{\perp}_{\infty}, \mu)$ has an BSD $\mathcal{Q}$ which is balanced for some constant depending on ρ^*.

Claim 4 *We can write*

$$H^{\perp}_{\infty}(w,w') = M_1(\xi^{\infty}(w), \xi^{\infty}(w'))UU^{\top} + M_2(\xi^{\infty}(w), \xi^{\infty}(w')),$$

where for $w^*, w^{*\prime} \in \mathrm{supp}(\rho^*)$,

$$\begin{aligned} M_1(w^*, w^{*\prime}) &:= \mathbb{E}_{x\sim\mathcal{D}_V} \sigma'(x^{\top}w^*)\sigma'(x^{\top}w^{*\prime}) \\ M_2(w^*, w^{*\prime}) &:= \mathbb{E}_{x\sim\mathcal{D}_V} \sigma'(x^{\top}w^*)\sigma'(x^{\top}w^{*\prime}) P^{\perp}_{w^*} x x^{\top} P^{\perp}_{w^{*\prime}}. \end{aligned}$$

Further, both M_1 and M_2 have rank at most $C_{\rho^*} = \min\left(|\mathrm{supp}(\rho^*)|, \dim(\rho^*)^{2\,\mathrm{degree}(\sigma)+1}\right)$.

Proof Let V be the orthonormal basis spanning $\mathrm{supp}(\rho^*)$, and let U be any orthonormal basis of $\mathbb{R}^d \backslash \mathrm{span}(V)$. Recall that Assumption I2 guarantees that the distribution of x, $\mathcal{D}_x$, can be factorized as $\mathcal{D}_U \otimes \mathcal{D}_V$, where $\mathrm{span}(\mathcal{D}_U) \in \mathrm{span}(U)$, $\mathrm{span}(\mathcal{D}_V) \in \mathrm{span}(V)$, $\mathbb{E}_{x\sim\mathcal{D}_U} x x^{\top} = UU^{\top}$, and $\mathbb{E}_{x\sim\mathcal{D}_U} x = 0$.

Recall that $H^{\perp}_{\infty}(w,w') = \mathbb{E}_{x\sim\mathcal{D}_x} \sigma'(x^{\top}\xi^{\infty}(w))\sigma'(x^{\top}\xi^{\infty}(w')) x x^{\top}$. Observe that for $u, v \in \mathrm{Span}(U)$, we have

$$\begin{aligned} u^{\top} H^{\perp}_{\infty}(w,w') v &= \mathbb{E}_{x\sim\mathcal{D}_x} \sigma'(x^{\top}\xi^{\infty}(w))\sigma'(x^{\top}\xi^{\infty}(w'))(u^{\top}x)(v^{\top}x) \\ &= \mathbb{E}_{x\sim\mathcal{D}_V} \sigma'(x^{\top}\xi^{\infty}(w))\sigma'(x^{\top}\xi^{\infty}(w')) \mathbb{E}_{x\sim\mathcal{D}_U} u^{\top} x x^{\top} v \\ &= \mathbb{E}_{x\sim\mathcal{D}_V} \sigma'(x^{\top}\xi^{\infty}(w))\sigma'(x^{\top}\xi^{\infty}(w')) \mathbb{E}_{x\sim\mathcal{D}_U} u^{\top} v. \end{aligned}$$

If $u \in \mathrm{Span}(U)$, $v \in \mathrm{Span}(V)$, then it is easy to check by the fact that $\mathbb{E}_{x\sim\mathcal{D}_U} x = 0$ that

$$u^{\top} H^{\perp}_{\infty}(w,w') v = \mathbb{E}_{x_V\sim\mathcal{D}} \sigma'(x_V^{\top}\xi^{\infty}(w))\sigma'(x_V^{\top}\xi^{\infty}(w'))(v^{\top}x_V)\mathbb{E}_{x_U\sim\mathcal{D}_U}(u^{\top}x_U) = 0.$$

For $w^*, w^{*\prime} \in \mathrm{supp}(\rho^*)$, let

$$\begin{aligned} M_1(w^*, w^{*\prime}) &:= \mathbb{E}_{x\sim\mathcal{D}_V} \sigma'(x^{\top}w^*)\sigma'(x^{\top}w^{*\prime}) \\ M_2(w^*, w^{*\prime}) &:= \mathbb{E}_{x\sim\mathcal{D}_V} \sigma'(x^{\top}w^*)\sigma'(x^{\top}w^{*\prime}) P^{\perp}_{w^*} x x^{\top} P^{\perp}_{w^{*\prime}} \end{aligned}$$

such that by the above computations,

$$H_\infty^\perp(w,w') = M_1(\xi^\infty(w),\xi^\infty(w'))UU^\top + M_2(\xi^\infty(w),\xi^\infty(w')).$$

The statement about the rank follows from the observations that (1) both M_1 and M_2 are defined on a space of size at most $|\mathrm{supp}(\rho^*)|$, and (2) Alternatively, we can replace the expectation of $x \sim \mathcal{D}_V$ with the expectation over some $x \sim \mathcal{D}'_V$, where $\mathcal{D}'_V$ is supported on at most $\dim(V)^{2\,\mathrm{degree}(\sigma)+1}$ points, and all the moments of $\mathcal{D}'_V$ up to the $\mathrm{degree}(\sigma)$th degree match those of $\mathcal{D}_V$ (as this requires matching at most $\sum_{j=0}^{2\,\mathrm{degree}(\sigma)} \dim(V)^j \le \dim(V)^{2\,\mathrm{degree}(\sigma)+1}$ terms.) ■

We will construct $\mathcal{Q}$ using the eigenfunctions of each of these two parts. Let $\mathcal{F} \subset L^2(\mathrm{supp}(\rho^*), (\xi^\infty)_{\#}\mu, \mathbb{R}^d)$ be an orthonormal basis of eigenfunctions of the linear operator $(M_2, (\xi^\infty)_{\#}\mu)$, that is, we have

$$\sum_{f\in\mathcal{F}} \lambda_f f(w^*)f(w^{*\prime})^\top = M_2(w^*,w^{*\prime})$$
$$\mathbb{E}_{w^{*\prime}\sim(\xi^\infty)_{\#}\mu} M_2(w^*,w^{*\prime})f(w^{*\prime}) = \lambda_f f(w^*),$$

Let $\mathcal{Y} \subset L^2(\mathrm{supp}(\rho^*), (\xi^\infty)_{\#}\mu)$ be an orthonormal basis of eigenfunctions of the linear operator $(M_1, (\xi^\infty)_{\#}\mu)$, that is, we have

$$\sum_{y\in\mathcal{Y}} \lambda_y y(w^*)y(w^{*\prime}) = M_1(w^*,w^{*\prime})$$
$$\mathbb{E}_{w^{*\prime}\sim(\xi^\infty)_{\#}\mu} M_1(w^*,w^{*\prime})y(w^{*\prime}) = \lambda_y y(w^*)$$

Let $\Lambda = \Lambda_1 \cup \Lambda_2$, where

$$\Lambda_2 := \{\lambda_f : f \in \mathcal{F}\} \qquad \Lambda_1 = \{\lambda_y : y \in \mathcal{Y}\}.$$

The following claim is immediate to check from the decomposition of $H_\infty^\perp$ in Claim 4.

Claim 5 *Let $\mathcal{P}_\lambda$ be the projector onto the eigenspace of $H_\infty^\perp$ with eigenvalue λ. Then $\mathcal{P}_\lambda = \overline{P}_\lambda$, where*

$$\begin{aligned} P_\lambda(w,w') &:= \sum_{f\in\mathcal{F}} f(\xi^\infty(w))f(\xi^\infty(w')^\top \mathbf{1}(\lambda_f=\lambda) + UU^\top \sum_{y\in\mathcal{Y}} y(\xi^\infty(w))y(\xi^\infty(w')\mathbf{1}(\lambda_y=\lambda) \\ &= \sum_{v\in\mathcal{B}_\lambda} v(w)v(w')^\top, \end{aligned}$$

where

$$\mathcal{B}_\lambda := \{v^f : \lambda_f = \lambda\}_{f\in\mathcal{F}} \cup \{v^{y,i} : \lambda_y = \lambda\}_{y\in\mathcal{Y}},$$

and

$$v^f(w) := f(\xi^\infty(w));$$
$$v^{y,i}(w) := y(\xi^\infty(w))U_i.$$

It remains to check how balanced this spectral decomposition is. Let $p := \min_{w^* \in \mathrm{supp}(\rho^*)} \mathbb{P}_{\xi^\infty_\# \mu}[w^*]$, and observe that $\max_{w, f \in \mathcal{F}, y \in \mathcal{Y}} (\|f(w)\|, |y(w)|) \leq \frac{1}{\sqrt{p}}$, since the eigenfunctions are orthonormal. Fix $\lambda \in \Lambda$. We have

$$\begin{aligned}\sum_{v \in \mathcal{B}_\lambda} v(w)v(w)^\top &= \sum_{f \in \mathcal{F}} v^f(w)(v^f(w))^\top \mathbf{1}(\lambda_f = \lambda) + \sum_{y \in \mathcal{Y}} \sum_{i=1}^{\dim(U)} v^{y,i}(w)v^{y,i}(w)\mathbf{1}(\lambda_y = \lambda) \\ &= \sum_{f \in \mathcal{F}} f(\xi^\infty(w))(f(\xi^\infty(w)))^\top \mathbf{1}(\lambda_f = \lambda) + \sum_{y \in \mathcal{Y}} y(\xi^\infty(w))y(\xi^\infty(w))UU^\top \mathbf{1}(\lambda_y = \lambda) \\ &\preceq \frac{I}{p}\left(\sum_{f \in \mathcal{F}} \mathbf{1}(\lambda_f = \lambda) + \sum_{y \in \mathcal{Y}} \mathbf{1}(\lambda_y = \lambda)\right).\end{aligned}$$

Thus letting

$$\eta_\lambda^2 := \frac{1}{p}\left(\sum_{f \in \mathcal{F}} \mathbf{1}(\lambda_f = \lambda) + \sum_{y \in \mathcal{Y}} \mathbf{1}(\lambda_y = \lambda)\right),$$

by Claim 4, we have that

$$\sum_{\lambda \in \Lambda} \eta_\lambda^2 = \frac{|\mathcal{F}| + |\mathcal{Y}|}{p} \leq \frac{\mathrm{rank}(M_1) + \mathrm{rank}(M_2)}{p} \leq \frac{2C_{\rho^*}}{p}.$$

Thus $\mathcal{Q} = \{(\mathcal{B}_\lambda, \eta_\lambda)\}_{\lambda \in \Lambda}$ is $\frac{2C_{\rho^*}}{p}$-balanced. This proves the first statement in the lemma.

If $(\rho^*, \rho_0, \mathcal{D}_x)$ is transitive (as per Definition 29), then we can get rid of the denominator and show that almost surely over $w \sim \rho_0$,

$$\sum_{v \in \mathcal{B}_\lambda} v(w)v(w)^\top \preceq I\left(\sum_{f \in \mathcal{F}} \mathbf{1}(\lambda_f = \lambda) + \sum_{y \in \mathcal{Y}} \mathbf{1}(\lambda_y = \lambda)\right)$$

This suffices to prove the lemma.

To do this, let $\mathcal{G}$ be the set of automorphisms of $(\rho^*, \rho_0, \mathcal{D}_x)$ as per Definition 29. For $h \in L^2(\mathbb{S}^{d-1}, \rho_0, \mathbb{R}^d)$, define $g(h)$ by

$$g(h)(w) := g^{-1}(f(g(w))).$$

For convenience, for $y \in \mathcal{Y}$, we will abuse notation and define

$$g(y)(w) := y(g(w)).$$

Claim 6 ($\mathcal{G}$-invariance of Eigenspaces.) *If $f \in \mathcal{F}$ is an eigenfunction of M_2, then $g(f)$ is an eigenfunction of M_2 with the same eigenvalue. Simlary, if $y \in \mathcal{Y}$ is an eigenfunction of M_1, then $g(y)$ is an eigenfunction of M_1 with the same eigenvalue.*

Proof We have

$$\begin{aligned}
M_2(g(f))(w^*) &= \mathbb{E}_{w^{*\prime}\sim(\xi^\infty)_{\#}\rho_0}\mathbb{E}_{x\sim\mathcal{D}_V}\sigma'(x^\top w^*)\sigma'(x^\top w^{*\prime})P^\perp_{w^*}xx^\top P^\perp_{w^{*\prime}}g^{-1}(f(g(w^{*\prime})))\\
&= \mathbb{E}_{w^{*\prime}\sim(\xi^\infty)_{\#}\rho_0}\mathbb{E}_{x\sim\mathcal{D}_V}\sigma'(x^\top w^*)\sigma'(x^\top w^{*\prime})g^{-1}\left(P^\perp_{g(w^*)}g(x)\right)x^\top g^{-1}\left(P^\perp_{g(w^{*\prime})}f(g(w^{*\prime}))\right)\\
&= \mathbb{E}_{w^{*\prime}\sim(\xi^\infty)_{\#}\rho_0}\mathbb{E}_{x\sim\mathcal{D}_V}\sigma'(g(x)^\top g(w^*))\sigma'(g(x)^\top g(w^{*\prime}))g^{-1}\left(P^\perp_{g(w^*)}g(x)\right)g(x)^\top P^\perp_{g(w^{*\prime})}f(g(w^{*\prime}))\\
&= \mathbb{E}_{w^{*\prime}\sim(\xi^\infty)_{\#}\rho_0}\mathbb{E}_{x\sim\mathcal{D}_V}\sigma'(x^\top g(w^*))\sigma'(x^\top w^{*\prime})g^{-1}\left(P^\perp_{g(w^*)}x\right)x^\top P^\perp_{w^{*\prime}}f(w^{*\prime})\\
&= g^{-1}\left(\mathbb{E}_{w^{*\prime}\sim(\xi^\infty)_{\#}\rho_0}\mathbb{E}_{x\sim\mathcal{D}_V}\sigma'(x^\top g(w^*))\sigma'(x^\top w^{*\prime})P^\perp_{g(w^*)}xx^\top P^\perp_{w^{*\prime}}f(w^{*\prime})\right)\\
&= g^{-1}\left(M_2f(g(w^*))\right)\\
&= g^{-1}\left(\lambda_f f(g(w^*))\right)\\
&= \lambda_f g(f)(w^*)
\end{aligned}$$

Here in the second line with used the fact that for any w and z, we have

$$(I-ww^\top)z = z - ww^\top z = z - wg(w)^\top g(z) = g^{-1}\left((I-g(w)g(w)^\top)g(z)\right)$$

If the third line, we just used that for $z, z' \in \mathbb{R}^d$, we have $z^\top z' = g(z)^\top g(z')$. In the fourth line, we used the symmetry of $\mathcal{D}_x$ and $(\xi^\infty)_{\#}\rho_0$ with respect to $\mathcal{G}$ (see **A2**). The proof for that $M_1g(y)(w^*) = \lambda_y g(y)(w^*)$ is similar (but simpler); we omit the computation. ■

Let $\mu_{\mathcal{G}}$ the uniform measure over the group generated by the set of all $g_{w^*,w^{*\prime}} \in \mathcal{G}$ for $w^*, w^{*\prime} \in \mathrm{supp}(\rho^*)$, where $g_{w^*,w^{*\prime}}(w^*) = w^{*\prime}$. Observe that $\mu_{\mathcal{G}}$ a left-invariant measure on $\mathcal{G}$, that is, for any $w^* \in \mathrm{supp}(\rho^*)$, we have that the distribution of $g(w^*)$ is uniform on ρ^* when $g \sim \mu_{\mathcal{G}}$ (that is, it equals ρ^*, since ρ^* is atomic). Also note that for $g \in \mathrm{supp}(\mu_{\mathcal{G}})$ and $v \in \mathrm{span}(V)$, we have that $g(v) \in \mathrm{span}(V)$. Thus for $u \in \mathrm{span}(U)$, we have $g(u) \in \mathrm{span}(U)$, and thus in particular, since g preserves dot products, and thus orthonormality,

$$g^{-1}(U)g^{-1}(U)^\top = UU^\top. \tag{20}$$

Claim 7 *Let $g \in \mathrm{supp}(\mu_{\mathcal{G}})$, and define $g(\mathcal{B}_\lambda) := \{g(v)\}_{v\in\mathcal{B}_\lambda}$. Then $g(\mathcal{B}_\lambda)$ is an orthonormal basis for $\mathcal{P}_\lambda$.*

Proof First we will check that almost surely over w, w',

$$\sum_{f\in\mathcal{F}}\lambda_f g(v^f)(w)g(v^f)(w')^\top + \sum_{y\in\mathcal{Y}}\lambda_y g(v^{y,i})(w)g(v^{y,i})(w')^\top = H^\perp_\infty(w,w'). \tag{21}$$

Using the definition of $g(f)$ and **A2**, almost surely over w, w', we have for $z, z' \in \mathbb{S}^{d-1}$,

$$\begin{aligned} z^\top \sum_{f\in\mathcal{F}} \lambda_f g(v^f)(w) g(v^f)(w')^\top z' &= z^\top \sum_{f\in\mathcal{F}} \lambda_f g^{-1}\left(f(\xi^\infty(g(w)))\right) g^{-1}\left(f(\xi^\infty(g(w')))^\top\right) z' \\ &= z^\top \sum_{f\in\mathcal{F}} \lambda_f g^{-1}\left(f(g(\xi^\infty(w)))\right) g^{-1}\left(f(g(\xi^\infty(w')))^\top\right) z' \\ &= \sum_{f\in\mathcal{F}} \lambda_f g(z)^\top f(g(\xi^\infty(w))) f(g(\xi^\infty(w')))^\top g(z') \\ &= g(z)^\top M_2(g(\xi^\infty(w)), g(\xi^\infty(w')))^\top g(z') \\ &= z^\top M_2(\xi^\infty(w), \xi^\infty(w'))^\top z', \end{aligned} \tag{22}$$

where here in the last line, we used the fact that

$$z^\top M_2(w^*, w^{*\prime})^\top z' = g(z)^\top M_2(g(w^*), g(w^{*\prime}))^\top g(z')$$

for any $g \in \mathcal{G}$, $w^*, w^{*\prime}$. This can verified from the definition of M_2 and the fact that $\mathcal{D}_x$ is invariant with respect to $\mathcal{G}$.

We can perform a similar (much easier) calculation to show that

$$\sum_{y\in\mathcal{Y}} \lambda_y g(y)(\xi^\infty(w)) g(y)(\xi^\infty(w')) = M_1(\xi^\infty(w), \xi^\infty(w'));$$

this arises from the fact that $M_1(w^*, w^{*\prime}) = M_1(g(w^*), g(w^{*\prime}))$ since $\mathcal{D}_x$ is invariant with respect to $\mathcal{G}$. We omit the details. Thus by (20),

$$\begin{aligned} \sum_{y\in\mathcal{Y}} \lambda_y g(v^{y,i})(w) g(v^{y,i})(w')^\top &= M_1(\xi^\infty(w), \xi^\infty(w')) g^{-1}(U) g^{-1}(U)^\top \\ &= M_1(\xi^\infty(w), \xi^\infty(w')) U U^\top. \end{aligned} \tag{23}$$

Employing (23) and (22) yields (21) almost surely as desired.

Now, to prove the claim, we use (1) the fact from Claim 6 guarantees that $g(v)$ is an eigenfunction with the same values as v, and (2) the fact that the set $\{g(v)\}_{v\in\mathcal{B}_\lambda}$ is orthonormal (since dot products are preserved under rotations). These two facts guarantee that $g(\mathcal{B}_\lambda)$ is a basis for $\mathcal{P}_\lambda$. ■

The following claim now suffices to prove the lemma.

Claim 8 *For any $w \in \mathbb{S}^{d-1}$, we have*

$$\sum_{v\in\mathcal{B}_\lambda} v(w) v(w)^\top \preceq I \left(\sum_{f\in\mathcal{F}} \mathbf{1}(\lambda_f = \lambda) + \sum_{y\in\mathcal{Y}} \mathbf{1}(\lambda_y = \lambda) \right).$$

Proof Fix any $w \in \mathbb{S}^{d-1}$, and let $w^* = \xi^\infty(w)$. For $z \in \mathbb{R}^d$, let $\pi_z \in L^2(\mathbb{S}^{d-1}, \rho_0, \mathbb{R}^d)$ be defined by $\pi_z(w') = z\mathbf{1}(\xi^\infty(w') = w^*)$. Then since for $v \in \mathcal{B}_\lambda$, we have $v(w) = v(w')$ if $\xi^\infty(w) = \xi^\infty(w')$, it follows that

$$z^\top P_\lambda(w, w) z = \sum_{v\in\mathcal{B}_\lambda} z^\top v(w) v(w)^\top z = \frac{\langle \overline{P}_\lambda \pi_z, \pi_z \rangle}{(\mathbb{P}_{w'\sim\rho_0}[\xi^{\infty -1}(w^*)])^2} = |\mathrm{supp}(\rho^*)|^2 \langle \overline{P}_\lambda \pi_z, \pi_z \rangle.$$

To see the last equality, observe that $\rho^* = \xi^\infty_\# \rho_0$ by **A2**.

Now recall that by Claim 7, for any $\lambda \in \Lambda$ and $g \in \mathrm{supp}(\mu_\mathcal{G})$, we have that $\{g(v)\}_{v \in \mathcal{B}_\lambda}$ is a basis for $\mathcal{P}_\lambda = \overline{P}_\lambda$, and thus

$$\begin{aligned} z^\top P_\lambda(w,w) z &= |\mathrm{supp}(\rho^*)|^2 \langle \overline{P}_\lambda \pi_z, \pi_z \rangle \\ &= |\mathrm{supp}(\rho^*)|^2 z^\top \mathbb{E}_{g\sim\mu_\mathcal{G}} \sum_{v \in g(\mathcal{B}_\lambda)} \mathbb{E}_{w',w''\sim\rho_0} v(w) v(w')^\top \mathbf{1}(\xi^\infty(w'), \xi^\infty(w'') = w^*) z \end{aligned} \tag{24}$$

$$= z^\top \mathbb{E}_{g\sim\mu_\mathcal{G}} \left(\sum_{f \in \mathcal{F} | \lambda_f = \lambda} g^{-1}(f(g(w^*))) g^{-1}(f(g(w^*)))^\top z + \sum_{y \in \mathcal{F} | \lambda_y = \lambda} |y(g(w^*)|^2 g^{-1}(U) g^{-1}(U)^\top \right) z.$$

Now for any $f \in \mathcal{F}$,

$$\begin{aligned} \mathbb{E}_{g\sim\mu_\mathcal{G}} g^{-1}(f(g(w^*)))(g^{-1}(f(g(w^*))))^\top &\preceq \mathbb{E}_{g\sim\mu_\mathcal{G}} \|f(g(w^*))\|^2 I \\ &= \mathbb{E}_{w^{*\prime}\sim\rho^*} \|f(w^{*\prime})\|^2 I \\ &= I. \end{aligned} \tag{25}$$

Here the second to last inequality holds because we have defined $\mu_\mathcal{G}$ to be a left-invariant measure on $\mathcal{G}$ that induces a uniform measure on $\mathrm{supp}(\rho^*)$. The last equation holds by the fact that $\rho^* = \xi^\infty_\# \rho_0$ (see **A2**) and since f is part of an orthonormal basis, we must have $\mathbb{E}_{w^*\sim\xi^\infty_\#\rho_0} \|f(w^*)\|^2 = 1$. Likewise, for $y \in \mathcal{Y}$, using (20),

$$\begin{aligned} \mathbb{E}_{g\sim\mu_\mathcal{G}} |(g(y))(w^*)|^2 g^{-1}(U) g^{-1}(U)^\top &= \mathbb{E}_{g\sim\mu_\mathcal{G}} |y(g(w^*))|^2 U U^\top \\ &= \mathbb{E}_{w^{*\prime}\sim\rho^*} |y(w^{*\prime})|^2 U U^\top \\ &= U U^\top. \end{aligned} \tag{26}$$

Combining Equations (25) and (26) with (24) yields that

$$P_\lambda(w,w) \preceq I \left(\sum_{f \in \mathcal{F}} \mathbf{1}(\lambda_f = \lambda) + \sum_{y \in \mathcal{Y}} \mathbf{1}(\lambda_y = \lambda) \right),$$

as desired. ■

■

F.2.3. PROPERTIES OF POTENTIAL

To prove our key lemmas 13, 14, 15, we will need several preliminary lemmas.

Lemma 34 *Suppose the high probability event in Lemma 25 holds for $S = B_\tau$ and $v \in L^2(\mathbb{S}^{d-1}, \rho_0, \mathbb{R}^d)$ which is an eigenfunction of $H^\perp_\infty$. Suppose $(H^\perp_\infty, \rho_0)$ has the CRI with respect to $B^t_\tau := \xi_t^{-1}(B_\tau)$. Then with $\|v\|_\infty := \sup w \in \mathbb{S}^{d-1} \|v(w)\|$, we have*

$$\langle \nabla \phi_v(t), \Delta_t \rangle^{B^t_\tau}_{H^\perp_\infty} = \mathbb{P}_{\rho^{\mathrm{MF}}_t}[B_\tau] \lambda_v \phi_v(t) + \mathcal{E} \|v\|_\infty,$$

where

$$\mathcal{E} \leq \epsilon_m^{25}\mathbb{E}_i\|\Delta_t(i)\| + \mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i) \notin B_\tau).$$

Proof First observe that

$$\nabla\phi_v = v\,\mathrm{sign}(\langle v, \Delta_t\rangle),$$

and thus

$$\langle \nabla\phi_v(t), \Delta_t\rangle_{H_\infty^\perp}^{B_\tau^t} = \mathrm{sign}(\langle v, \Delta_t\rangle)\langle v, \Delta_t\rangle_{H_\infty^\perp}^{B_\tau^t}$$

Now by the conclusion of the concentration Lemma 25, we have

$$\langle v, \Delta_t\rangle_{H_\infty^\perp}^{B_\tau^t} = \mathbb{E}_i X(i)\Delta_t(i)\mathbf{1}(\xi_t(w_i) \in B_\tau) \pm \|v\|_\infty \epsilon_m^{25}\mathbb{E}_i\|\Delta_t(i)\|.$$

where $X(i) = \mathbb{E}_{w'\sim\rho_0} H_\infty^\perp(w_i, w')v(w')\mathbf{1}(\xi_t(w') \in B_\tau)$ Now since v is an eigenfunction of $H_\infty^\perp$, by the definition of consistent isometry, we have that

$$X(i) = \lambda_v v(w_i)\mathbb{P}_{\rho_t^{\mathrm{MF}}}[B_\tau].$$

Thus

$$\langle v, \Delta_t\rangle_{H_\infty^\perp}^{B_\tau^t} = \lambda_v \langle v, \Delta_t\rangle^{B_\tau^t}\mathbb{P}_{\rho_t^{\mathrm{MF}}}[B_\tau] \pm \epsilon_m^{25}\|v\|_\infty\mathbb{E}_i\|\Delta_t(i)\|.$$

Now

$$\begin{aligned}\mathrm{sign}(\langle v, \Delta_t\rangle)\langle v, \Delta_t\rangle^{B_\tau^t} &= \mathrm{sign}(\langle v, \Delta_t\rangle)\langle v, \Delta_t\rangle \pm \mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i) \notin B_\tau) \\ &= \phi_v(t) \pm \mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i) \notin B_\tau).\end{aligned}$$

Plugging this back in yields the lemma. ■

Now we prove Lemma 13, which we restate here.

Lemma 35 (Descent with Respect to Interaction Term) *Let $\Phi_{\mathcal{Q}}(t)$ be as defined above, where $\mathcal{Q}$ is a C_b-balanced spectral decomposition of $H_\infty^\perp$. Then for any $\tau > 0$ for which the concentration event of Lemma 25 holds for $S = B_\tau$, we have*

$$\langle \nabla\Phi_{\mathcal{Q}}(t), -H_t^\perp\Delta_t\rangle \leq (1 + C_b)\mathbb{E}_i\|\mathbb{E}_j H_t^\perp(i,j)\Delta_t(j)\|\mathbf{1}(\xi_t(w_i) \notin B_\tau) + \mathcal{E}_{13},$$

where $\mathcal{E}_{13} = O_{C_{reg}, C_b}(\mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i) \notin B_\tau) + (\tau + C_b\epsilon_m^{25})\Omega(t))$.

Proof Let $B_\tau^t := \xi_t^{-1}(B_\tau)$, and let $\bar{B}_\tau^t$ be the complement in $\mathbb{S}^{d-1}$ of B_τ^t. We decompose

$$\langle \nabla\Phi_{\mathcal{Q}}(t), \Delta_t\rangle_{H_t^\perp} = \langle \nabla\Phi_{\mathcal{Q}}(t), \Delta_t\rangle_{H_t^\perp}^{B_\tau^t, B_\tau^t} + \langle \nabla\Phi_{\mathcal{Q}}(t), \Delta_t\rangle_{H_t^\perp}^{B_\tau^t, \bar{B}_\tau^t} + \langle \nabla\Phi_{\mathcal{Q}}(t), \Delta_t\rangle_{H_t^\perp}^{\bar{B}_\tau^t, \mathbb{S}^{d-1}}. \quad (27)$$

Lets start with the first term $\langle \nabla\Phi_{\mathcal{Q}}(t), \Delta_t\rangle_{H_t^\perp}^{B_\tau^t, B_\tau^t} = \langle \nabla\Phi_{\mathcal{Q}}(t), \Delta_t\rangle_{H_t^\perp}^{B_\tau^t}$. Bounding this term is the key part of the lemma.

Claim 9

$$\langle \nabla \Phi_{\mathcal{Q}}(t), \Delta_t \rangle_{H_t^\perp}^{B_\tau^t} \geq -(C_{reg}+1)\mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i) \notin B_\tau) - C_b\epsilon_m^{25}\Omega(t) + |\langle \nabla\Phi(t), G\rangle|\,,$$

where $\mathbb{E}_i\|G(i)\| \leq C_{reg}\tau\Omega(t)$.

Proof We have

$$\langle \nabla \Phi_{\mathcal{Q}}(t), \Delta_t \rangle_{H_t^\perp}^{B_\tau^t} = \langle \nabla \Phi_{\mathcal{Q}}(t), \Delta_t \rangle_{H_\infty^\perp}^{B_\tau^t} + \langle \nabla \Phi_{\mathcal{Q}}(t), G\rangle, \tag{28}$$

where $\|G(i)\| \leq C_{\text{reg}}\tau\mathbb{E}_i\|\Delta_t(i)\|$, since $\|K'(\xi^\infty(w),\xi^\infty(w')) - K'(\xi_t(w),\xi_t(w'))\| \leq C_{\text{reg}}\tau$. This relies on the fact that from the proof of **A2**, almost surely $\|\xi_t(w) - \xi^\infty(w)\| \leq \tau$, because $\|\xi_t(w) - \xi^\infty(w)\| \leq \min_{w^*\in\text{supp}(\rho^*)}\|\xi_t(w) - w^*\| \leq \tau$. Now we will break up $\Phi_{\mathcal{Q}}$ into the $\Psi_{\mathcal{Q}}$ and Ω parts. Starting with the $\Psi_{\mathcal{Q}}$ part, we have

$$\begin{aligned}
\langle \nabla \Psi_{\mathcal{Q}}(t), \Delta_t \rangle_{H_\infty^\perp}^{B_\tau^t} &= \sum_{\lambda\in\Lambda}\eta_\lambda \frac{\sum_{v\in\mathcal{B}_\lambda}\phi_v(t)\langle\nabla\phi_v(t),\Delta_t\rangle_{H_\infty^\perp}^{B_\tau^t}}{\sqrt{\sum_{v\in\mathcal{B}_\lambda}(\phi_v(t))^2}} \\
&= \sum_{\lambda\in\Lambda}\eta_\lambda \frac{\sum_{v\in\mathcal{B}_\lambda}\phi_v(t)\left(\lambda\phi_v(t)\mathbb{P}_{\rho_t^{\text{MF}}}[B_\tau] + \mathcal{E}_{\mathbf{v}}\right)}{\sqrt{\sum_{v\in\mathcal{B}_\lambda}(\phi_v(t))^2}} \\
&= \mathbb{P}_{\rho_t^{\text{MF}}}[B_\tau]\sum_{\lambda\in\Lambda}\eta_\lambda\left(\lambda\sqrt{\sum_{v\in\mathcal{B}_\lambda}(\phi_v(t))^2} + \frac{\sum_{v\in\mathcal{B}_\lambda}\phi_v(t)\mathcal{E}_{\mathbf{v}}}{\sqrt{\sum_{v\in\mathcal{B}_\lambda}(\phi_v(t))^2}}\right) \\
&\geq \mathbb{P}_{\rho_t^{\text{MF}}}[B_\tau]\sum_{\lambda\in\Lambda}\eta_\lambda\lambda\sqrt{\sum_{v\in\mathcal{B}_\lambda}(\phi_v(t))^2} - \mathcal{E}\,,
\end{aligned} \tag{29}$$

where we used Cauchy-Schwartz in the last inequality, the fact that $\sum_\lambda\eta_\lambda = 1$, and $\|\mathcal{E}_{\mathbf{v}}\| \leq \mathcal{E}$, the error term appearing in Lemma 34. **[MG comments:** fix the constant here, may gain C_{b}.**]**

Next consider the $\langle\nabla\Omega(t),\Delta_t\rangle_{H_\infty^\perp}^{B_\tau^t}$ part. Recall from the definition of BSD that $H_\infty^\perp(w,w') = \sum_{v\in\mathcal{Q}}\lambda_v v(w)v(w')^\top$. Let $u_i := \nabla_i\Omega(t) = \frac{\Delta_t(i)}{\|\Delta_t(i)\|}$. We can expand

$$\begin{aligned}
\left|\langle\nabla\Omega(t),\Delta_t\rangle_{H_\infty^\perp}^{B_\tau^t,\mathbb{S}^{d-1}}\right| &= \left|\mathbb{E}_{i,j}\sum_{v\in\mathcal{Q}}\lambda_v u_i^\top v(w_i)v(w_j)^\top\Delta_t(j)\mathbf{1}(w_i\in B_\tau^t)\right| \\
&= \left|\mathbb{E}_i\sum_{v\in\mathcal{Q}}\lambda_v u_i^\top v(w_i)\mathbf{1}(i\in B_\tau^t)\left(\mathbb{E}_j v(w_j)^\top\Delta_t(j)\right)\right| \\
&\leq \sum_{v\in\mathcal{Q}}\lambda_v\phi_v(t)\mathbb{E}_i|u_i^\top v(w_i)|\mathbf{1}(i\in B_\tau^t).
\end{aligned} \tag{30}$$

Now fix i. For any vector $u \in \mathbb{S}^{d-1}$, since $\mathcal{Q} = \{(\mathcal{B}_\lambda, \eta_\lambda)\}_{\lambda \in \lambda}$ is C_{b}-balanced, we have

$$\begin{aligned}
\sum_{v \in \mathcal{Q}} \lambda_v \phi_v(t) |u^\top v(w_i)| &= \sum_{\lambda \in \Lambda} \lambda \sum_{v \in \mathcal{B}_\lambda} \phi_v(t) |u^\top v(w_i)| \\
&\leq \sum_{\lambda \in \Lambda} \lambda \sqrt{\sum_{v \in \mathcal{B}_\lambda} (\phi_v(t))^2} \sqrt{\sum_{v \in \mathcal{B}_\lambda} |u^\top v(w_i)|^2} \\
&= \sum_{\lambda \in \Lambda} \lambda \sqrt{\sum_{v \in \mathcal{B}_\lambda} (\phi_v(t))^2} \sqrt{u^\top \left(\sum_{v \in \mathcal{B}_\lambda} v(w_i) v(w_i)^\top \right) u} \\
&\leq \sum_{\lambda \in \Lambda} \eta_\lambda \lambda \sqrt{\sum_{v \in \mathcal{B}_\lambda} (\phi_v(t))^2}.
\end{aligned}$$

Here the final inequality follows from the definition of a BSD , which states that for any $w \in \mathbb{S}^{d-1}$, $\sum_{v \in \mathcal{B}_\lambda} v(w) v(w)^\top \preceq \eta_\lambda^2 I$. Thus plugging this back into to Equation (30), we have

$$\left| \langle \nabla \Omega(t), \Delta_t \rangle_{H_\infty^\perp}^{B_\tau^t, [m]} \right| \leq \mathbb{P}_i[B_\tau^t] \sum_{\lambda \in \Lambda} \eta_\lambda \lambda \sqrt{\sum_{v \in \mathcal{B}_\lambda} (\phi_v(t))^2}.$$

Now letting $H_i = H_\infty^\perp(w_i, w_j) \mathbb{E}_j \Delta_t(j) \mathbf{1}(w_i \notin B_\tau^t)$, we have

$$\left| \langle \nabla \Omega(t), \Delta_t \rangle_{H_\infty^\perp}^{B_\tau^t} - \langle \nabla \Omega(t), \Delta_t \rangle_{H_\infty^\perp}^{B_\tau^t, [m]} \right| \leq |\langle \nabla \Omega(t), H \rangle| \leq C_{\mathrm{reg}} \mathbb{E}_i \|\Delta_t(i)\| \mathbf{1}(\xi_t(w_i) \notin B_\tau),$$

and thus

$$\left| \langle \nabla \Omega(t), \Delta_t \rangle_{H_\infty^\perp}^{B_\tau^t} \right| \leq \mathbb{P}_i[B_\tau^t] \sum_{\lambda \in \Lambda} \eta_\lambda \lambda \sqrt{\sum_{v \in \mathcal{B}_\lambda} (\phi_v(t))^2} + C_{\mathrm{reg}} \mathbb{E}_i \|\Delta_t(i)\| \mathbf{1}(\xi_t(w_i) \notin B_\tau). \quad (31)$$

Now recall that $\Phi_{\mathcal{Q}}(t) := \Omega(t) + \Psi_{\mathcal{Q}}(t)$. Thus combining Equations (31) and (29), and Equation (28), and plugging in the bound on $\mathcal{E}$ from Lemma 34, we have

$$\langle \nabla \Phi_{\mathcal{Q}}(t), \Delta_t \rangle_{H_t^\perp}^{B_\tau^t} \geq -(C_{\mathrm{reg}} + 1) \mathbb{E}_i \|\Delta_t(i)\| \mathbf{1}(\xi_t(w_i) \notin B_\tau) - C_{\mathrm{b}} \epsilon_m^{25} \Omega(t) + |\langle \nabla \Phi_{\mathcal{Q}}(t), G \rangle|,$$

where $\mathbb{E}_i \|G_i\| \leq C_{\mathrm{reg}} \tau \Omega(t)$. Here we have also used the fact that for all v in the BSD $\mathcal{Q}$, we have that $\|v\|_\infty \leq \sqrt{C_{\mathrm{b}}} \leq C_{\mathrm{b}}$ (this is evident from the definition of BSD). This proves the claim. ■

Next consider the second term $\langle \nabla \Phi_{\mathcal{Q}}(t), \Delta_t \rangle_{H_t^\perp}^{B_\tau^t, \bar{B}_\tau^t}$ in Equation (27). We have

$$\left| \langle \nabla \Phi_{\mathcal{Q}}(t), \Delta_t \rangle_{H_t^\perp}^{B_\tau^t, \bar{B}_\tau^t} \right| = \langle \nabla \Phi_{\mathcal{Q}}(t), H \rangle, \quad (32)$$

where $\|H(i)\| \leq C_{\mathrm{reg}} \mathbb{E}_i \|\Delta_t(i)\| \mathbf{1}(\xi_t(w_i) \notin B_\tau)$.

Finally, for the third term $\langle \nabla \Phi_{\mathcal{Q}}(t), \Delta_t \rangle_{H_t^\perp}^{\bar{B}_\tau^t, \mathbb{S}^{d-1}}$ in Equation (27), we have just write

$$\langle \nabla \Phi_{\mathcal{Q}}(t), \Delta_t \rangle_{H_t^\perp}^{\bar{B}_\tau^t, \mathbb{S}^{d-1}} = \langle \nabla \Phi_{\mathcal{Q}}(t), m_t \rangle^{\bar{B}_\tau^t}, \quad (33)$$

where we recall that $m_t(i) = \mathbb{E}_j H_t^{\perp}(i,j)\Delta_t(j)$.

Combining Equations (32), (33) and Claim 9 into Equation 27, we obtain that

$$\langle \nabla \Phi_{\mathcal{Q}}(t), \Delta_t \rangle_{H_t^{\perp}} \geq -(C_{\text{reg}}+1)\mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i) \notin B_\tau) - C_{\text{b}}\epsilon_m^{25}\Omega(t) + |\langle \nabla \Phi_{\mathcal{Q}}(t), G+H+m_t\rangle|\,,$$

where $\mathbb{E}_i\|G(i)+H(i)\| \leq C_{\text{reg}}\left(\tau\Omega(t) + \mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i)\notin B_\tau)\right)$.

Now we use Lemma 15 to bound

$$\begin{aligned}|\langle \nabla \Phi_{\mathcal{Q}}(t), G+H+m_t\rangle| &\leq \mathbb{E}_i\|G(i)+H(i)+m_t(i)\|\,(1+C_{\text{b}}) \\ &\leq \left(C_{\text{reg}}\left(\tau\Omega(t)+\mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i)\notin B_\tau)\right) + \mathbb{E}_i\|m_t(i)\|\right)(1+C_{\text{b}})\,.\end{aligned}$$

Plugging this back in to the equation above yields

$$\langle \nabla \Phi_{\mathcal{Q}}(t), \Delta_t \rangle_{H_t^{\perp}} \geq -(1+C_{\text{b}})\mathbb{E}_i\|m_t(i)\| - \mathcal{E}_{13},$$

where

$$\begin{aligned}\mathcal{E}_{13} &= (C_{\text{reg}}(2+C_{\text{b}})+1)\mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i)\notin B_\tau) + (C_{\text{b}}\epsilon_m^{25} + (1+C_{\text{b}})C_{\text{reg}}\tau)\Omega(t) \\ &= O_{C_{\text{reg}},C_{\text{b}}}(\mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i)\notin B_\tau) + (\tau + C_{\text{b}}\epsilon_m^{25})\Omega(t)).\end{aligned}$$

This proves the lemma. ■

Now we prove Lemma 14, which we restate here.

Lemma 36 (Descent with Respect to Local Term) *Suppose Assumption LSC holds with (C_{LSC}, τ). Let $\mathcal{Q}$ be a C_b-balanced spectral distribution. Then with $C_{14} = O_{C_{reg},C_b}(1)$, we have*

$$\langle \nabla \Phi_{\mathcal{Q}}(t), D_t^{\perp}\odot\Delta_t\rangle \leq -\left(\frac{c\sqrt{L_{\mathcal{D}}(\rho_t^{\text{MF}})}}{2} - C_{14}\tau\right)\Phi_{\mathcal{Q}}(t) + C_{14}\mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i)\notin B_\tau) + C_b\mathbb{E}_i\|\Delta_t(i)\|^2.$$

Proof Let $\delta := \sqrt{L_{\mathcal{D}}(\rho_t^{\text{MF}})}$. We will show that

$$\langle \nabla \Omega(t), D_t^{\perp}\odot\Delta_t\rangle \leq -\left(C_{\text{LSC}}\delta\right)\Omega(t) + 2C_{\text{reg}}\mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i)\notin B_\tau),$$

and that

$$\begin{aligned}\langle \nabla \Psi_{\mathcal{Q}}(t), D_t^{\perp}\odot\Delta_t\rangle \leq -\left(C_{\text{LSC}}\delta\right)\Psi_{\mathcal{Q}}(t) + \frac{C_{\text{LSC}}\delta + 2C_{\text{b}}C_{\text{reg}}\tau}{2}\Omega(t) \\ 2C_{\text{b}}C_{\text{reg}}\mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i)\notin B_\tau) + C_{\text{b}}\mathbb{E}_i\|\Delta_t(i)\|^2.\end{aligned} \tag{34}$$

The first statement is straightforward. Since $\nabla_i\Omega(t) = \frac{\Delta_t(i)}{\|\Delta_t(i)\|}$, we have

$$\begin{aligned}\langle \nabla \Omega(t), D_t^{\perp}\odot\Delta_t\rangle &\leq \mathbb{E}_i\frac{\Delta_t(i)^{\top}D_t^{\perp}(i)\Delta_t(i)}{\|\Delta_t(i)\|} \\ &= \mathbb{E}_i\frac{\Delta_t(i)^{\top}D_t^{\perp}(i)\Delta_t(i)}{\|\Delta_t(i)\|}\mathbf{1}(\xi_t(w_i)\in B_\tau) + \mathbb{E}_i\frac{\Delta_t(i)^{\top}D_t^{\perp}(i)\Delta_t(i)}{\|\Delta_t(i)\|}\mathbf{1}(\xi_t(w_i)\notin B_\tau) \\ &\leq -C_{\text{LSC}}\delta\mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i)\in B_\tau) + \mathbb{E}_i\|D_t^{\perp}(i)\Delta_t(i)\|\mathbf{1}(\xi_t(w_i)\notin B_\tau) \\ &\leq -C_{\text{LSC}}\delta\mathbb{E}_i\|\Delta_t(i)\| + 2C_{\text{reg}}\mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i)\notin B_\tau),\end{aligned}$$

as desired.

For the second statement, write

$$D_t^{\perp}(i) = D_t^{\text{good}}(i) + D_t^{\text{bad}}(i),$$

where

$$D_t^{\text{good}}(i) = -c_1 P^{\perp}_{\xi^{\infty}(w_i)}(VV^{\top})P^{\perp}_{\xi^{\infty}(w_i)} - c_2(UU^{\top}).$$

By the structured condition in Assumption LSC, we can write such a decomposition where $c_1, c_2 \geq C_{\text{LSC}}\delta$, and for any i such that $\xi_t(w_i) \in B_\tau$, we have $\|D_t^{\text{bad}}(i)\| \leq \frac{C_{\text{LSC}}\delta}{2\sqrt{C_{\text{b}}}} + C_{\text{reg}}\tau$. Note that this decomposition still holds for i where $\xi_t(w_i) \notin B_\tau$, but $\|D_t^{\text{bad}}(i)\|$ can be as large as $2C_{\text{reg}}$.

Claim 10

$$\langle \nabla\phi_v(t), D_t^{good} \odot \Delta_t \rangle \leq -C_{\text{LSC}}\delta\phi_v(t) + \langle \nabla\phi_v(t), G \rangle,$$

where $\|G(i)\| \leq \tau\|\Delta_t(i)\| + 0.5\|\Delta_t(i)\|^2 + \|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i) \notin B_\tau)$.*;*

Proof

Now recall that in the construction for $\mathcal{Q}$ given in Lemma 33, for any $v \in \text{supp}(\mathcal{Q})$, it holds that either $v(w) \in \text{span}(U)$ for all $w \in \mathbb{S}^{d-1}$, or $v(w) \in \text{span}(V)$ for all $w \in \mathbb{S}^{d-1}$. We consider the two cases separately. First suppose $v(w) \in \text{span}(U)$ for all $w \in \mathbb{S}^{d-1}$. Fix w_i with $\xi_t(w_i) \in B_\tau$. For any w, we have

$$v(w)^{\top} D_t^{\text{good}}(i)\Delta_t(i) = -c_2 v(w)^{\top}\Delta_t(i),$$

and thus the desired conclusion holds. Now suppose $v(w) \in \text{span}(V)$. Note that V commutes with $P^{\perp}_{\xi^{\infty}(w_i)}$. Thus any w, we have

$$v(w)^{\top} D_t^{\text{good}}(i)\Delta_t(i) = -c_1 v(w) P^{\perp}_{\xi^{\infty}(w_i)}\Delta_t(i).$$

Now for i with $\xi_t(w) \in B_\tau$, we have $\|\xi_t(w) - \xi^{\infty}(w)\| \leq \tau$ (see the proof of **A2**), and thus, since additionally $|\Delta_t(i)\xi_t(w)| \leq \frac{\|\Delta_t(i)\|2}{2}$ (see (17) in the proof of Lemma 5), we have that

$$\begin{aligned} v(w)^{\top} D_t^{\text{good}}(i)\Delta_t(i) &= -c_1 v(w) P^{\perp}_{\xi^{\infty}(w_i)}\Delta_t(i) \\ &= -c_1 v(w)\Delta_t(i) + O(\tau\|v(w)\| + \|\Delta_t(i)\|^2). \end{aligned}$$

Thus in conclusion, we have that

$$\langle \nabla\phi_v(t), D_t^{\text{good}} \odot \Delta_t \rangle \leq -c_2\delta\phi_v(t) + \langle \nabla\phi_v(t), G \rangle,$$

where $\|G(i)\| \leq \tau\|\Delta_t(i)\| + 0.5\|\Delta_t(i)\|^2 + \|\Delta_t(i)\|\mathbb{1}(\xi_t(w_i) \notin B_\tau)$. This proves the claim. ■

Thus with G as in the claim,

$$\begin{aligned}\langle \nabla \Psi_{\mathcal{Q}}(t), D_t^{\text{good}} \odot \Delta_t - G\rangle &\leq \sum_{\lambda\in\Lambda} \eta_\lambda \frac{\sum_{v\in\mathcal{B}_\lambda} \phi_v(t)\langle \nabla\phi_v(t), D_t^{\text{good}}\odot\Delta_t\rangle}{\sqrt{\sum_{v\in\mathcal{B}_\lambda}(\phi_v(t))^2}} \\ &\leq \sum_{\lambda\in\Lambda} \eta_\lambda \frac{\sum_{v\in\mathcal{B}_\lambda} -C_{\text{LSC}}\delta(\phi_v(t))^2}{\sqrt{\sum_{v\in\mathcal{B}_\lambda}(\phi_v(t))^2}} \\ &= -C_{\text{LSC}}\delta \sum_{\lambda\in\Lambda}\eta_\lambda \sqrt{\sum_{v\in\mathcal{B}_\lambda}(\phi_v(t))^2} \\ &= -C_{\text{LSC}}\delta\Phi_{\mathcal{Q}}(t).\end{aligned}$$

It follows that from the proof of Lemma 15 (see Equation (35)) we have

$$|\langle \nabla\Psi_{\mathcal{Q}}(t), D_t^{\text{good}}\odot\Delta_t - G\rangle| \leq C_{\text{b}}\left(\tau\Omega(t) + 0.5\mathbb{E}_i\|\Delta_t(i)\|^2 + \mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i)\notin B_\tau)\right)$$

Similarly, we have that

$$\begin{aligned}\langle \nabla\Psi_{\mathcal{Q}}(t), D_t^{\text{bad}}\odot\Delta_t\rangle &= \langle \nabla\Psi_{\mathcal{Q}}(t), D_t^{\text{bad}}\odot\Delta_t\rangle_{B_\tau^t} + \langle \nabla\Psi_{\mathcal{Q}}(t), D_t^{\text{bad}}\odot\Delta_t\rangle_{\bar{B}_\tau^t} \\ &\leq C_{\text{b}}\left(\frac{C_{\text{LSC}}\delta}{2\sqrt{C_{\text{b}}}} + C_{\text{reg}}\tau\right)\mathbb{E}_i\|\Delta_t(i)\| + C_{\text{b}}(2C_{\text{reg}})\mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i)\notin B_\tau),\end{aligned}$$

and so

$$\langle \nabla\Psi_{\mathcal{Q}}(t), D_t^{\perp}\odot\Delta_t\rangle \leq -C_{\text{LSC}}\delta\Psi_{\mathcal{Q}}(t) + \left(\frac{C_{\text{LSC}}\delta + 2C_{\text{b}}C_{\text{reg}}\tau}{2}\Omega(t)\right) + 3C_{\text{b}}C_{\text{reg}}\mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i)\notin B_\tau).$$

This yields (34), which proves the lemma.

■

We now prove Lemma 15, which we restate here.

Lemma 37 (L1 Perturbation Lemma) *Let $\mathcal{Q}$ be a C_b-balanced spectral distribution. Let $G : [m] \to \mathbb{R}^d$. Then $|\langle \nabla\Phi_{\mathcal{Q}}(t), G\rangle| \leq (1 + C_b)\mathbb{E}_i\|G(i)\|$.*

Proof [Proof of Lemma 15] First observe that $\langle \nabla\Phi_{\mathcal{Q}}(t), G\rangle \leq \mathbb{E}_i\|G(i)\|$, since $\nabla_i\Omega(t) = \frac{\Delta_t(i)}{\|\Delta_t(i)\|}$, which has norm 1. Now for any $v \in \text{supp}(\mathcal{Q})$, we have

$$|\langle \nabla\phi_v(t), G\rangle| \leq \mathbb{E}_i|G(i)^\top v(w_i)|,$$

and so

$$
\begin{aligned}
|\langle \nabla \Psi_{\mathcal{Q}}(t), G\rangle| &\leq \sum_{\lambda\in\Lambda} \eta_\lambda \frac{\sum_{v\in\mathcal{B}_\lambda}\phi_v(t)|\langle\nabla\phi_v(t), G\rangle}{\sqrt{\sum_{v\in\mathcal{B}_\lambda}(\phi_v(t))^2}} \\
&\leq \mathbb{E}_i\left[\sum_{\lambda\in\Lambda}\eta_\lambda \frac{\sum_{v\in\mathcal{B}_\lambda}\phi_v(t)|G(i)^\top v(w_i)|}{\sqrt{\sum_{v\in\mathcal{B}_\lambda}(\phi_v(t))^2}}\right] \\
&\leq \mathbb{E}_i\left[\sum_{\lambda\in\Lambda}\eta_\lambda \frac{\sqrt{\sum_{v\in\mathcal{B}_\lambda}(\phi_v(t))^2}\sqrt{\sum_{v\in\mathcal{B}_\lambda}|G(i)^\top v(w_i)|^2}}{\sqrt{\sum_{v\mathcal{B}_\lambda}(\phi_v(t))^2}}\right] \\
&= \mathbb{E}_i\left[\sum_{\lambda\in\Lambda}\eta_\lambda \sqrt{G(i)^\top\left(\sum_{v\in\mathcal{B}_\lambda} v(w_i)v(w_i)^\top\right)G(i)}\right] \\
&\leq \mathbb{E}_i\sum_{\lambda\in\Lambda}\eta_\lambda^2\|G(i)\| = C_{\mathrm{b}}\mathbb{E}_i\|G(i)\|.
\end{aligned}
\tag{35}
$$

Here in the third inequality, we used Cauchy-Schwartz. It follows that

$$
|\langle\nabla\Phi_{\mathcal{Q}}(t), G\rangle| \leq |\langle\nabla\Omega(t), G\rangle| + |\langle\nabla\Psi_{\mathcal{Q}}(t), G\rangle| \leq (1+C_{\mathrm{b}})\mathbb{E}_i\|G(i)\|,
$$

as desired. ■

F.3. Dynamics of the Potential

Before proving our main theorem on the dynamics of the potential, we need the following lemma, which gathers all the required concentration events.

Lemma 38 *Fix some δ. With high probability as $d, m, n \to \infty$, the events in all concentration lemmas (Lemma 23,Lemma 27, Lemma 24 and Lemma 25) hold, where we apply Lemma 24 and Lemma 25 for $S = B_\tau$ for all*

$$
\tau \in \left\{\frac{C_{\mathrm{LSC}}\cdot rd(e)}{8(C_{13}+C_{14})}\right\}_{e\in[\delta,1]},
$$

where $rd(z)$ is a rounding of z to its first non-zero decimal, in binary (so $rd(z)\in[z/2, z]$). We also apply Lemma 25 for all eigenfunctions v in the BSD $\mathcal{Q}$.

Proof The set $\left\{\frac{c\cdot \mathrm{rd}(e)}{8(C_{13}+C_{14})}\right\}_{e\in[\delta,1]}$ has size at most $O_{C_{13},C_{14}}(\log_2(1/\delta))$, so we can take a union bound over the result in Lemma 24 for all B_τ. Similarly, since there are $O(dC_{\rho^*})$ eigenfunctions in $\mathcal{Q}$ (see the proof of Lemma 33), we take a union bound of Lemma 25 over all these eigenfunctions. (Note that the "with high probability" is explicitly $o(1/d)$ there). The rest follows immediately from the three concentration lemmas. ■

For the remainder of the text, we assume the following assumptions hold up to time T (if relevant): Assumptions Regularity,Stability,LSC,Symmetry. Let (C_{LSC}, τ) denote the parameters of the local strong convexity (we will use the parameter τ differently later). We also assume that $\mathcal{Q}$ is a C_{b}-balanced BSD , where by Lemma 12, we have that $C_{\text{b}} = C_{\rho^*}$.

Theorem 39 (Main Potential Dynamics Theorem) *Let $\delta := \sqrt{L_{\mathcal{D}}(\rho_T^{\text{MF}})}$, and let C be a constant depending on $C_{\text{LSC}}, \tau, \delta$ and C_b. Condition on the event that the high probability event in Lemma 38 holds for δ. Let $\epsilon_{n,m} := \epsilon_n + \epsilon_m^{23} + \epsilon_m^{24} + \epsilon_m^{25}$ from the concentration lemmas. Suppose n and m are large enough such that $J_{max}^4 T^3(\epsilon_n + \epsilon_m) \le 1/C$. Suppose that*

$$J_{max}^2 \left(\int_{s=0}^{t} \Phi_{\mathcal{Q}}(s)^2 ds \right) \le \epsilon_{n,m}.$$

and $J_{max}^2 t^2 \epsilon_{n,m} \le \frac{1}{64}$. Then for some $C = O_{C_{reg}, C_b}(1)$ and $\tau = \Omega_{C_{reg}, C_b}(\delta)$, for all $t \le T$, we have

$$\frac{d}{dt}\Phi_{\mathcal{Q}}(t) \le -\frac{C_{\text{LSC}}\delta}{C}\Phi_{\mathcal{Q}}(t) + C J_{avg}(\tau) \int_{s=0}^{t} \Phi_{\mathcal{Q}}(s) ds + C J_{max} t \epsilon_{n,m}.$$

Corollary 40 (Solution to Potential Dynamics) *Suppose that for some $\tau = \Omega_{C_{reg}, C_b}(\delta)$,*

$$4 J_{max}^4 C^2 T^3 \exp(2C J_{avg}(\tau) t / (C_{\text{LSC}}\delta)) \epsilon_{n,m} \le 1.$$

Then for some $C = O_{C_{reg}, C_b}(1)$ we have

$$\Phi_{\mathcal{Q}}(T) \le \exp(CT J_{avg}(\tau)/(C_{\text{LSC}}\delta)) C J_{max} T \epsilon_{n,m}.$$

Proof We will use real induction (see eg. [Cla12, Theorem 2]). Our inductive hypothesis will be that for some t,

$$J_{\text{max}}^2 \left(\int_{s=0}^{t} \Phi_{\mathcal{Q}}(s)^2 ds \right) \le \frac{1}{2}\epsilon_{n,m}.$$

Note that is implies the assumption in Equation 39. Clearly this holds for $t = 0$. Since $\Phi_{\mathcal{Q}}(s)$ is continuous, if Equation 39 holds for all $s < t$, it also holds for t. This is the continuity assumption. Finally, for the inductive step, we will show that if Equation 39 holds for some s, then for some ι small enough, it holds at $s + \iota$. To show this, first we use Lemma 42 (which bounds the solution of the ODE given in Theorem 39), to show that for all $s' \le s$,

$$\Phi_{\mathcal{Q}}(s') \le \exp(Cs' J_{\text{avg}}(\tau)/(C_{\text{LSC}}\delta)) C s J_{\text{max}} \epsilon_{n,m} + \epsilon_{n,m} \le \left(\exp(Cs J_{\text{avg}}(\tau)/(C_{\text{LSC}}\delta)) C s J_{\text{max}}\right) \epsilon_{n,m}.$$

Note that $\Phi_{\mathcal{Q}}(t)$ is continuous. Thus for ι small enough, we have $\Phi_{\mathcal{Q}}(t) \le \Phi_{\mathcal{Q}}(s) + \epsilon_{n,m}$ for all $t \in [s, s + \iota]$. It follows that for ι small enough,

$$\int_{s'=0}^{t} (\Phi_{\mathcal{Q}}(s'))^2 ds' \le (Ct J_{\text{max}} \epsilon_{n,m})^2 \int_{s'=0}^{s} \exp(2Cs J_{\text{avg}}(\tau)/(C_{\text{LSC}}\delta)) ds' + \int_{s'=s}^{t} (\Phi_{\mathcal{Q}}(s) + \epsilon_{n,m})^2 ds'$$

$$\le 2(Ct J_{\text{max}} \epsilon_{n,m})^2 t \exp(2C J_{\text{avg}}(\tau) t/(C_{\text{LSC}}\delta))$$

Now using the assumption in the corollary that

$$4J_{\max}^4 C^2 t^3 \exp(2CJ_{\text{avg}}(\tau)t/(C_{\text{LSC}}\delta))\epsilon_{n,m} \leq 1,$$

it follows that $\int_{s=0}^{s'}(\Phi_{\mathcal{Q}}(s))^2 ds \leq \frac{\epsilon_{n,m}}{2J_{\max}^2}$.

This proves the inductive step. Thus by real induction, the hypothesis holds up to time T. The result of the lemma then holds by applying Lemma 42 to the result of Theorem 39 at time T. ■

Proof [Proof of Theorem 39] Recall from Lemma 5 that

$$\frac{d}{dt}\Delta_t(i) = D_t^{\perp}(i)\Delta_t(i) - \mathbb{E}_j H_t^{\perp}(i,j)\Delta_t(j) + \boldsymbol{\epsilon}_{t,i},$$

where

$$\|\boldsymbol{\epsilon}_{t,i}\| \leq 2\epsilon_{n,m} + 2C_{\text{reg}}\left(\|\Delta_t(i)\|^2 + \mathbb{E}_j\|\Delta_t(j)\|^2\right).$$

Now we have

$$\begin{aligned}\frac{d}{dt}\Phi_{\mathcal{Q}}(t) &\leq \langle \nabla\Phi_{\mathcal{Q}}(t), \frac{d}{dt}\Delta_t\rangle \\ &= -\langle \nabla\Phi_{\mathcal{Q}}(t), H_t^{\perp}\Delta_t\rangle + \langle \nabla\Phi_{\mathcal{Q}}(t), D_t^{\perp}\odot\Delta_t\rangle + \langle \nabla\Phi_{\mathcal{Q}}(t), \mathcal{E}\rangle,\end{aligned}$$

where $\mathcal{E}(i) = \boldsymbol{\epsilon}_{t,i}$. We will consider the terms in order. Let

$$\tau := \frac{C_{\text{LSC}}\cdot \text{rd}(\delta)}{8(C_{13}+C_{14})},$$

where $\text{rd}(z)$ is a rounding of z to its first non-zero decimal, in binary (so $\text{rd}(z)\in[z/2, 2z]$).

Now by Lemma 13, we have

$$-\langle \nabla\Phi_{\mathcal{Q}}(t), H_t^{\perp}\Delta_t\rangle = -\langle \nabla\Phi_{\mathcal{Q}}(t), \Delta_t\rangle_{H_t^{\perp}} \leq (1+C_{\text{b}})\mathbb{E}_i\|m_t(i)\|\mathbf{1}(\xi_t(w_i)\notin B_\tau) + \mathcal{E}_{13},$$

where $m_t(i) = \mathbb{E}_j H_t^{\perp}(i,j)\Delta_t(j)$, and

$$\mathcal{E}_{13} = C_{13}(\mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i)\notin B_\tau) + (\tau + C_{\text{b}}\epsilon_m^{25})\Omega(t)).$$

Next by Lemma 14, we have

$$\langle \nabla\Phi_{\mathcal{Q}}(t), D_t^{\perp}\odot\Delta_t\rangle \leq -\left(\frac{C_{\text{LSC}}\delta}{2} - \tau C_{14}\right)\Phi_{\mathcal{Q}}(t) + C_{14}\mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i)\notin B_\tau) + C_{\text{b}}\mathbb{E}_i\|\Delta_t(i)\|^2.$$

Here we have used the fact that since the loss is decreasing, the loss in Lemma 14 is less than the loss δ^2 at time T.

Putting these together, and employing Lemma 15, yields

$$\begin{aligned}\frac{d}{dt}\Phi_{\mathcal{Q}}(t) \leq &\left(-\frac{C_{\text{LSC}}\delta}{4}\right)\Phi_{\mathcal{Q}}(t) \\ &+ (C_{13}+C_{14})\mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i)\notin B_\tau) \\ &+ (1+C_{\text{b}})\mathbb{E}_i\|m_t(i)\|\mathbf{1}(\xi_t(w_i)\notin B_\tau) \\ &+ (1+2C_{\text{b}})\mathbb{E}_i\|\epsilon_{t,i}\|,\end{aligned} \tag{36}$$

where here we used that τ was chosen such that $(C_{14} + C_{13})(\tau + C_{\mathrm{b}}\epsilon_m^{25}) \leq \frac{C_{\mathrm{LSC}}\delta}{8}$, and trivially, $\Omega(t) \leq \Phi_{\mathcal{Q}}(t)$. We also bounded $\mathbb{E}_i\|\Delta_t(i)\|$ by $\mathbb{E}_i\|\epsilon_{t,i}\|$.

Now let us consider the term $\mathbb{E}_i\|m_t(i)\|\mathbf{1}(\xi_t(w_i) \notin B_\tau)$. Using Lemma 26, we have

$$\mathbb{E}_i\|m_t(i)\|\mathbf{1}(\xi_t(w_i) \notin B_\tau) \leq (1 + C_{\mathrm{b}})\left(\epsilon_{n,m} + J_{\mathrm{avg}}(\tau)\right)\Phi_{\mathcal{Q}}(t).$$

Now let use consider the term $\mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i) \notin B_\tau)$. Recall from Equation (3) that

$$\Delta_t(i) = -\int_{s=0}^{t} J_{t,s}(i)m_s(i)ds + \int_{s=0}^{t} J_{t,s}(i)\boldsymbol{\epsilon_{s,i}}ds.$$

Thus by Lemma 26, we have

$$\begin{aligned}\mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i) \notin B_\tau) \leq (1 + C_{\mathrm{b}})\left(\epsilon_{n,m} + J_{\mathrm{avg}}(\tau)\right)\int_{s=0}^{t}\Phi_{\mathcal{Q}}(s)ds \\ + \int_{s=0}^{t}\mathbb{E}_i\|J_{t,s}(i)\boldsymbol{\epsilon_{s,i}}\|\mathbf{1}(\xi_t(w_i) \notin B_\tau)ds.\end{aligned}$$

Plugging this back into Equation (36) yields

$$\begin{aligned}\frac{d}{dt}\Phi_{\mathcal{Q}}(t) &\leq -\frac{C_{\mathrm{LSC}}\delta}{5}\Phi_{\mathcal{Q}}(t) + (C_{13} + 4\sqrt{C_{\mathrm{b}}}C_{\mathrm{reg}})(1 + C_{\mathrm{b}})(\epsilon_{n,m} + J_{\mathrm{avg}}(\tau))\int_{s=0}^{t}\Phi_{\mathcal{Q}}(s)ds \\ &\quad + (1 + C_{\mathrm{b}})\mathbb{E}_i\|\boldsymbol{\epsilon}_{t,i}\| + (1 + C_{\mathrm{b}})(C_{13} + 4\sqrt{C_{\mathrm{b}}}C_{\mathrm{reg}})\int_{s=0}^{t}\mathbb{E}_i\|J_{t,s}(i)\boldsymbol{\epsilon_{s,i}}\|ds \\ &\leq -\frac{C_{\mathrm{LSC}}\delta}{5}\Phi_{\mathcal{Q}}(t) + CJ_{\mathrm{avg}}(\tau)\int_{s=0}^{t}\Phi_{\mathcal{Q}}(s)ds \\ &\quad + (1 + C_{\mathrm{b}})\mathbb{E}_i\|\boldsymbol{\epsilon}_{t,i}\| + C\int_{s=0}^{t}\mathbb{E}_i\|J_{t,s}(i)\boldsymbol{\epsilon_{s,i}}\|ds,\end{aligned}$$

where $C = O_{C_{\mathrm{reg}},C_{\mathrm{b}}}(1)$. Let us simplify the error terms. Appealing to Lemma 41, we have for all i, $\|\Delta_t(i)\|^2 \leq 4\epsilon_{n,m}$ and $E_{t,i} := \int_{s=0}^{t}\|J_{t,s}(i)\boldsymbol{\epsilon_{s,i}}\|ds \leq 8J_{\max}t\epsilon_{n,m}$.

Thus

$$\mathbb{E}_i\|\boldsymbol{\epsilon}_{t,i}\| \leq 2\epsilon_{n,m} + 4C_{\mathrm{reg}}\mathbb{E}_i\|\Delta_t(i)\|^2 \leq 18C_{\mathrm{reg}}\epsilon_{n,m},$$

and

$$\int_{s=0}^{t}\mathbb{E}_i\|J_{t,s}(i)\boldsymbol{\epsilon_{s,i}}\|ds = \mathbb{E}_iE_{t,i} \leq 8J_{\max}t\epsilon_{n,m}.$$

Thus plugging this back into the bound on the dynamics, we have

$$\frac{d}{dt}\Phi_{\mathcal{Q}}(t) \leq -\frac{C_{\mathrm{LSC}}\delta}{5}\Phi_{\mathcal{Q}}(t) + CJ_{\mathrm{avg}}\int_{s=0}^{t}\Phi_{\mathcal{Q}}(s)ds + CJ_{\max}t\epsilon_{n,m}ds,$$

where $C = O_{C_{\mathrm{reg}},C_{\mathrm{b}}}(1)$. ■

Lemma 41 (Inductive Squared Error Bound.) *Suppose Assumption Stability hold with value J_{max}. Suppose for all $t' \leq t$, we have*

$$J_{max}^2 \left(\int_{s=0}^{t'} \Phi_{\mathcal{Q}}(s)^2 ds \right) \leq \epsilon_{n,m}.$$

and $J_{max}^2 t^2 \epsilon_{n,m} \leq \frac{1}{64}$. Then for all i and $t' \leq t$, we have

$$\|\Delta_{t'}(i)\|^2 \leq 4\epsilon_{n,m}$$
$$E_{t,i} := \int_{s=0}^{t} \|J_{t,s}(i)\boldsymbol{\epsilon_{s,i}}\| ds \leq 8 J_{max} t \epsilon_{n,m},$$

where $\boldsymbol{\epsilon_{s,i}}$ is defined in Lemma 5.

Proof It suffices to prove the statement just for the final time t, because we could always apply the lemma with a smaller value of t.

Recall that

$$\boldsymbol{\epsilon_{s,i}} \leq 2\epsilon_{n,m} + 2C_{\text{reg}} \left(\|\Delta_t(i)\|^2 + \mathbb{E}_j \|\Delta_t(j)\|^2 \right).$$

Since

$$\mathbb{E}_i \|\boldsymbol{\epsilon}_{t,i}\| \leq 2\epsilon_{n,m} + 4C_{\text{reg}} \mathbb{E}_i \|\Delta_t(i)\|^2,$$

by Equation 3, we have

$$\begin{aligned}
\|\Delta_t(i)\| &\leq \int_{s=0}^{t} J_{t,s}(i)(m_s(i) + \boldsymbol{\epsilon}_{s,i}) ds \\
&\leq \int_{s=0}^{t} \|J_{t,s}(i) m_s(i) ds\| + \int_{s=0}^{t} \|J_{t,s}(i)\boldsymbol{\epsilon}_{s,i} ds\| \\
&= \int_{s=0}^{t} \|J_{t,s}(i) m_s(i) ds\| + E_{t,i} \\
&\leq \sqrt{\int_{s=0}^{t} \|J_{t,s}(i)\|^2 ds} \sqrt{\int_{s=0}^{t} \|m_s(i)\|^2 ds} + E_{t,i} \\
&\leq J_{\max} \sqrt{\int_{s=0}^{t} \|m_s(i)\|^2 ds} + E_{t,i} \\
&\leq J_{\max} \sqrt{\int_{s=0}^{t} \Phi_{\mathcal{Q}}(s)^2 ds} + E_{t,i} \\
&\leq \sqrt{\epsilon_{n,m}} + E_{t,i},
\end{aligned}$$

Here in the second last inequality, we used the fact that $\|m_s(i)\| \leq \Phi_{\mathcal{Q}}(s)$ for any i, **[MG comments:** Might want to have a formal lemma for this? I think I used it in the potential proof too.] and in the last line, we used assumption of the lemma. Note that this same calculation holds for all $s \leq t$, so we have

$$\|\Delta_s(i)\| \leq \sqrt{\epsilon_{n,m}} + E_{t,i}.$$

Now lets bound $E_{t,i}$:

$$E_{t,i} := \int_{s=0}^{t} \|J_{t,s}(i)\boldsymbol{\epsilon_{s,i}}\| ds \leq \int_{s=0}^{t} \|J_{t,s}(i)\| \left(2\epsilon_{n,m} + 4C_{\text{reg}} \max_j \|\Delta_s(j)\|^2\right) ds$$
$$\leq J_{\max} \int_{s=0}^{t} \left(2\epsilon_{n,m} + \max_j \left(2\epsilon_{n,m} + 2E_{t,j}^2\right)\right) ds,$$

where in the second line, we plugged in the bound on $\Delta_s(i)$.

Thus letting $E_t := \max_j E_{t,j}$, we have

$$E_t \leq 2J_{\max} t \left(2\epsilon_{n,m} + E_t^2\right)$$

Now assuming the discriminant $1 - 32J_{\max}^2 t^2 \epsilon_{n,m} > 0$, this equation has two sets of disjoint solutions, one small (including 0) and one large:

$$E_t \in \left[-\infty, \frac{1 - \sqrt{1 - 32J_{\max}^2 t^2 \epsilon_{n,m}}}{4J_{\max} t}\right] \cup \left[\frac{1 + \sqrt{1 - 32J_{\max}^2 t^2 \epsilon_{n,m}}}{4J_{\max} t}, \infty\right]$$

Note that since at time $t = 0$, we have $E_t = 0$, and E_t is continuous, it must be that if the discriminant is positive up to time t, the solution is always in the first set. Indeed, since an assumption of the lemma is that $J_{\max}^2 t^2 \epsilon_{n,m} \leq \frac{1}{64}$.

Thus we have

$$E_t \leq \frac{1 - \sqrt{1 - 32J_{\max}^2 t^2 \epsilon_{n,m}}}{4J_{\max} t} \leq 8J_{\max} t \epsilon_{n,m}.$$

Plugging this back above into our bound on $\Delta_t(i)$ yields that for all i,

$$\|\Delta_t(i)\|^2 \leq 4\epsilon_{n,m}.$$

■

Lemma 42 (ODE Analysis) *Suppose we have a differential equation of the form*

$$\frac{d}{dt} X_t \leq -aX_t + b \int_{s=0}^{t} X_s ds + \epsilon.$$

with initial condition $X_0 = 0$ and $a, b \geq 0$. Then

$$X_t \leq \exp(bt/a) \frac{\epsilon}{\sqrt{a^2 + 4b}}.$$

Proof Let Y_t solve the ODE

$$\frac{d}{dt} Y_t = -aY_t + b \int_{s=0}^{t} Y_s ds + 2\epsilon,$$

with initial condition $Y_0 = 0$, and let $Z_t = X_t - Y_t$. We will show that Z_t never goes above 0.

Observe that Z_t solves the differential equation

$$\frac{d}{dt}Z_t \leq -aZ_t + b\int_{s=0}^{t} Z_s ds - \epsilon,$$

with initial condition $Z_t = 0$. One can check by the *real induction* that $Z_t \leq 0$. Indeed, if $Z_s \leq 0$ for all $s < t$, then we have $Z_t \leq 0$. Further, since Z_t is continuous, if the hypothesis $Z_t \leq 0$ holds up to time s, we can show that it holds at time $s + \iota$ for some $\iota > 0$. Indeed, for ι small enough (in terms of b and ϵ), for all $r \in [s, s + \iota]$, we have $Z_r \leq \frac{\epsilon}{b}$. Thus for $r \in [s, s + \iota]$, we have $\frac{d}{dr}Z_r \leq -aZ_r + b\iota\left(\frac{\epsilon}{b}\right) - \epsilon \leq -aZ_r$ for $\iota \leq 1$. Then Gronwall's inequality gives that $Z_{s+\iota} \leq Z_s \leq 0$, which is the inductive step. This yields the claim that $Z_t \leq 0$ for all $t > 0$.

Now we just need to solve the differential equation for Y_t. Taking a second derivative, we have

$$Y_t'' = -aY_t' + bY_t.$$

A standard second order ODE analysis yields that

$$Y_t = C_1 \exp(r_1 t) + C_2 \exp(r_2 t),$$

where r_1 and r_2 are the roots of $x^2 + ax - b = 0$, that is,

$$(r_1, r_2) = \frac{-a \pm \sqrt{a^2 + 4b}}{2}$$

Checking the initial condtitions of Y_0 and Y_0' yields

$$Y_t = \left(\frac{\epsilon}{\sqrt{a^2 + 4b}}\right)\left(\exp(r_1 t) - \exp(r_2 t)\right),$$

where r_1 is the larger root. Since $r_1 \leq \frac{b}{a}$, we obtain the lemma. ■

Appendix G. Applications to Learning a Single-index Model

G.1. Setting

We will study the setting of learning a well-specified even single index function $f^*(x) = \sigma(x^\top w^*)$, where $w^* \in \mathbb{S}^{d-1}$, and $\sigma(z) = \sum_{k=k^*}^{K} c_k \mathrm{He}_k(z)$, where:

1. $k^* \geq 4$, and $\frac{1}{C_{\text{SIM}}} \leq c_{k^*} \leq C_{\text{SIM}} \max_k c_k$.
2. For all k, $c_k \geq 0$.
3. All k with $c_k \neq 0$ are even. (That is, σ is an even function).

We assume the initial distribution ρ_0 of the neurons is uniform on $\mathbb{S}^{d-1}$, and the data is drawn i.i.d from the distribution $\mathcal{D}$, which has Gaussian covariates, and subGaussian label noise: that is,

$$x \sim \mathcal{N}(0, I_d) \sim \mathcal{D}_x$$
$$y = f^*(x) + \zeta(x),$$

where $\zeta(x)$ has mean 0 and is 1-subGaussian.

We will prove the following theorem, which we restate from Theorem 9 in the main body.

Theorem 9 (PoC in Single-Index Model) *Fix any $\delta > 0$, and suppose d is large enough in terms of δ, C_{SIM} and K. Let $T(\delta) := \arg\min\{t : \|f_{\rho_t^{\mathrm{MF}}} - f^*\|^2 \le \delta^2\}$. Then $T(\delta) = O_{K,C_{SIM}}(\sqrt{d}^{k^*-2}\delta^{-(k^*-1)})$. If $n \ge d^{11k^*}$ and $m \ge d^{13k^*}$, then with high probability, for all $t \le T(\delta)$,*

$$\|f_{\rho_t^{\mathrm{MF}}} - f_{\rho_t^m}\|^2 \le \frac{O_{K,\delta}(d^{3k^*})}{\min(\sqrt{m}, \sqrt{n})} \le 3\delta^2 .$$

We will prove Theorem 9 by (1) analyzing the MF dynamics to show the convergence of ρ_t^{MF}, and then (2) checking the assumptions of Theorem 7 hold, and applying it to show the convergence of ρ_t^m.

Notation Define $\alpha(w) := |w^\top w^*|$. Let $v(\alpha, t)$ denote the velocity of a particle w with $\alpha(w) = \alpha$ in the $w^* \operatorname{sign}(w^\top w^*)$ direction. Formally, we have

$$v(\alpha, t) := \langle w^*, \nu(w, \rho_t^{\mathrm{MF}})\rangle \operatorname{sign}(w^\top w^*),$$

for any w with $\alpha(w) = \alpha$. We will often use the notation $\alpha \sim \rho$ or $\alpha' \sim \rho$ to denote the distribution of $\alpha(w)$ with $w \sim \rho$. We use $\alpha_t(w) := \alpha(\xi_t(w))$. We use $\xi_{t,s}(w)$ denote the location of the particle at time t which is initialized at w at time s. In this language, we have that $\xi_t(w) = \xi_{t,0}(w)$. We similarly define $\alpha_{t,s}(\beta)$ to be $\alpha(\xi_{t,s}(w))$ for any w with $\alpha(w) = \beta$.

We will use q_σ to denote the polynomial with kth coefficient $k!c_k^2$, where $\sum c_k \mathrm{He}_k(z)$ is the Hermite decomposition of σ. Similarly, we denote $q_{\sigma'}(z) = \sum_{k=k^*-1}^{K-1} c_{k+1}^2(k+1)(k+1)!z^k$. From the Hermite polynomial identity that $\mathbb{E}_x \mathrm{He}_k(w^\top x)\mathrm{He}_j(v^\top x) = k!\delta_{jk}(w^\top v)^k$, we have

$$\mathbb{E}_x \sigma(w^\top x)\sigma(v^\top x) = q_\sigma(w^\top v).$$
$$\mathbb{E}_x \sigma'(w^\top x)\sigma'(v^\top x) = q_{\sigma'}(w^\top v).$$

G.2. Bounds on the Velocity and its Derivative

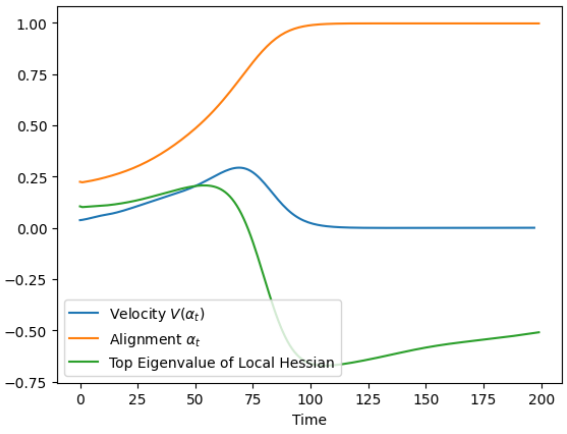


Figure 6: Self-Concordance Property: the top eigenvalue of the Local Hessian is Bounded by $\frac{k-1}{\alpha_t}\nu(\alpha_t)$

The key ingredients in both the MF convergence analysis, the perturbation analysis (bounding $J_{\max}$ and J_{avg}), and in showing local strong convexity, is obtaining a lower bound on the particle

velocity, and bounds on the local Hessian, $D_t^{\perp}(w)$. It turns out, it is much easier to bound these quantities under a certain inductive assumption (which in our MF analysis we will prove holds). We define the inductive property with parameter ι to hold at time t if

$$\mathbb{P}_{w\sim\rho_t^{\mathrm{MF}}}\left[\alpha(w)\in[\iota,1-\iota]\right]\leq\iota. \tag{$\star$}$$

Eventually, we will choose ι to be some small constant dependent on the desired final loss δ.

Lemma 43 (Lower Bound of Velocity) *Let $\delta := \sqrt{\mathbb{E}_x(f_{\rho_t^{\mathrm{MF}}}(x)-f^*(x))^2}$. Suppose ($\star$) holds at time t for $\iota \leq \min\left(\Theta_K(1), \delta^{6K^2}\right)$. Then*

$$v(\alpha,t)\geq q_{\sigma'}(\alpha)(1-\alpha^2)(1-r_t)-O_K\left((1-\alpha^2)\mathcal{R}_\alpha\right),$$

where $\mathcal{R}_\alpha = O_K(\iota(\alpha\sqrt{d}^{-\max(2,k^*-2)}+\alpha^{\max(1,k^*-3)}\sqrt{d}^{-2})+\alpha\sqrt{d}^{-k^*})$, and $r_t = \mathbb{E}_{\alpha'\sim\rho_t^{\mathrm{MF}}}(\alpha')^{k^*} = \Omega_K(\delta)$. In particular, if $\alpha \geq \frac{\delta^{3K}}{\sqrt{d}}$, for d large enough (in terms of δ, K), we have that

$$v(\alpha,t)\geq q_{\sigma'}(\alpha)(1-\alpha^2)(1-r_t)(1-\sqrt{\iota}).$$

Proof Let us expand the velocity by expressing $v(\alpha,t)$ as a polynomial in terms of α. Fix w with $\alpha(w)=\alpha$ and without loss of generality assume $w^\top w^*>0$. For $w'\in\mathbb{S}^{d-1}$, we denote $w'=\alpha' w^*+y$, where $y'\in\sqrt{1-\alpha'^2}\mathbb{S}^{d-2}$, which we will use to denote the sphere perpendicular to w^* of radius $\sqrt{1-\alpha'^2}$. We expand

$$\begin{aligned}
\nu(w,\rho_t^{\mathrm{MF}})^\top w^* &= \mathbb{E}_x(f^*(x)-f_{\rho_t^{\mathrm{MF}}}(x))\sigma'(w^\top x)x^\top P_w^\perp w^* \\
&= \mathbb{E}_x\sigma(w^{*\top}x)\sigma'(w^\top x)x^\top P_w^\perp w^* - \mathbb{E}_{w'\sim\rho_t^{\mathrm{MF}}}\mathbb{E}_x\sigma(w'^\top x)\sigma'(w^\top x)x^\top P_w^\perp w^* \\
&= q_{\sigma'}(w^\top w^*)w^{*\top}P_w^\perp w^* - \mathbb{E}_{w'\sim\rho_t^{\mathrm{MF}}}q_{\sigma'}(w^\top w')(w')^\top P_w^\perp w^* \\
&= q_{\sigma'}(\alpha)(1-\alpha^2) - \mathbb{E}_{w'\sim\rho_t^{\mathrm{MF}}}q_{\sigma'}(w^\top w')(w')^\top P_w^\perp w^* \\
&= q_{\sigma'}(\alpha)(1-\alpha^2) - \mathbb{E}_{\alpha'\sim\rho_t^{\mathrm{MF}}}\mathbb{E}_{y'\sim\sqrt{1-\alpha'^2}\mathbb{S}^{d-2}}q_{\sigma'}(\alpha\alpha'+y'^\top w)\left(\alpha'(1-\alpha^2)-\alpha y'^\top w\right).
\end{aligned} \tag{37}$$

Here in the fifth equality, we used the rotational symmetry of ρ_t^{MF} about the w^* axis.

Lets break down this expression. Let

$$r_{t,k} := \mathbb{E}_{\alpha\sim\rho_t^{\mathrm{MF}}}\alpha^k.$$

Fix a (necessarily odd) coefficient $k^*-1\leq k\leq K-1$ of the polynomial $q_{\sigma'}(z) := \sum q_k z^k$, and consider all terms in the above equation arising from that order term:

$$\begin{aligned}
&q_k\alpha^k(1-\alpha^2) - q_k\sum_{j=0}^{k}\binom{k}{j}(\alpha\alpha')^j\mathbb{E}_{y'\sim\sqrt{1-\alpha'^2}\mathbb{S}^{d-2}}(y'^\top w)^{k-j}(\alpha'(1-\alpha^2)-\alpha y'^\top w) \\
&\quad= q_k\alpha^k(1-\alpha^2)\left(1-r_{t,k+1}\right) + \mathbb{E}_{\alpha'\sim\rho_t^{\mathrm{MF}}}\mathcal{E}_{\alpha,\alpha',k},
\end{aligned}$$

where

$$\mathcal{E}_{\alpha,\alpha',k} = \begin{cases} O_k\left((1-\alpha^2)(\alpha')^2(1-\alpha'^2)(\alpha\sqrt{d}^{-(k-1)} + \alpha^{k-2}\sqrt{d}^{-2}) + \alpha(1-\alpha^2)\sqrt{d}^{-(k+1)}\right) & k \geq 3 \\ 0 & k = 1 \end{cases}.$$

Note here that we have used the fact that k is even and $\mathbb{E}_{y'}(y'^\top w)^j = O_j(\big((1-\alpha'^2)(1-\alpha^2)d^{-1}\big)^{j/2})$, and is 0 for odd j. The final error terms arises from the fact that we have only counted the terms in the binomial expansion which could be most significant — depending on the relative size of $\alpha\alpha'$ and $\sqrt{(1-\alpha^2)(1-\alpha'^2)}/\sqrt{d}$. Now plugging in the hypothesis ($\star$), we have that $\mathbb{E}_{\alpha'\sim\rho_t^{\mathrm{MF}}}(\alpha')^2(1-\alpha'^2) \leq 2\iota$, so for all k,

$$\mathcal{E}_{\alpha,\alpha',k} = O_k\left((1-\alpha^2)\iota(\alpha\sqrt{d}^{-(k-1)} + \alpha^{k-2}\sqrt{d}^{-2}) + \alpha(1-\alpha^2)\sqrt{d}^{-(k+1)}\right) \leq (1-\alpha^2)\mathcal{R}_\alpha$$

Summing over all odd $k^* - 1 \leq k \leq K-1$ yields that

$$\begin{aligned} v(\alpha,t) &= \sum_{k=k^*-1}^{K-1} q_k\alpha^k(1-\alpha^2)(1-r_{t,k+1}) + (1-\alpha^2)\mathcal{R}_\alpha \\ &\geq q_{\sigma'}(\alpha)(1-\alpha^2)(1-r_t) - (1-\alpha^2)\mathcal{R}_\alpha, \end{aligned} \tag{38}$$

where here in the inequality, we used the fact that $r_t = \mathbb{E}_{\alpha'\sim\rho_t^{\mathrm{MF}}}(\alpha')^{k^*} \geq \mathbb{E}_{\alpha'\sim\rho_t^{\mathrm{MF}}}(\alpha')^k = r_{t,k}$ for all $k \geq k^*$. Now for $\alpha \geq \frac{\delta^{3K}}{\sqrt{d}}$, we have

$$\begin{aligned} v(\alpha,t) \geq\ & q_{\sigma'}(\alpha)(1-\alpha^2)(1-r_t) \\ & - O_K\left(\iota(1-\alpha^2)(\alpha^{k^*-1}\delta^{-3K(k^*-2)} + \alpha^{k^*-1}\delta^{-6K}) + (1-\alpha^2)\alpha^{k^*-1}\delta^{-3K(k^*-2)}/d\right), \end{aligned}$$

Since by Lemma 46, we have $(1-r_t) = \Omega(\delta)$, it follows that

$$v(\alpha,t) \geq q_{\sigma'}(\alpha)(1-\alpha^2)(1-r_t)(1-\sqrt{\iota}).$$

■

In the following lemma, we analyze $\frac{d}{d\alpha}v(\alpha,t)$. As will be shown in Section G.4, bounding $\frac{d}{d\alpha}v(\alpha,t)$ is useful in bounding $D_t^\perp(w)$. The second part of this lemma will also be instrumental in proving local strong convexity (Definition LSC).

Lemma 44 *Let* $\delta := \sqrt{\mathbb{E}_x(f_{\rho_t^{\mathrm{MF}}}(x) - f^*(x))^2}$. *Suppose* ($\star$) *holds at time* t *for* $\iota \leq \min\left(\Theta_K(1), \delta^{6K^2}\right)$. *Then*

$$\frac{d}{d\alpha}v(\alpha,t) \begin{cases} = \frac{k^*-1}{\alpha}v(\alpha,t) + \mathcal{E}_\alpha & \alpha \leq 1; \\ \leq -\frac{\alpha}{1-\alpha^2}v(\alpha,t) - \Omega_K(\delta) & \alpha \geq 1 - \frac{1}{5K}, \end{cases}$$

where $\mathcal{E}_\alpha := \Theta_K\left(\alpha^{k^*} + \iota(\sqrt{d}^{-(k^*-2)} + \alpha^{k^*-4}\sqrt{d}^{-2}) + \sqrt{d}^{-(k^*-2)}\right)$.

Proof First we compute $\frac{d}{d\alpha}v(\alpha,t)$. Fix a coefficient $k^*-1 \le k \le K-1$ of the polynomial $q_{\sigma'}$, and consider all terms in the the derivative of Equation (37) arising from that order term:

$$\begin{aligned}
&q_k k\alpha^{k-1}\left(1-\frac{k+2}{k}\alpha^2\right)\\
&\quad - q_k\sum_{j=0}^{k}\binom{k}{j}j(\alpha\alpha')^{j-1}\mathbb{E}_{y'\sim\mathbb{S}^{d-2}_{\sqrt{1-\alpha'^2}}}(y'^\top w)^{k-j}\left(\alpha'\left(1-\frac{j+2}{j}\alpha^2\right)+\frac{j+1}{j}\alpha y'^\top w\right)\\
&= kq_k\alpha^{k-1}\left(1-\frac{k+2}{k}\alpha^2\right)(1-r_{t,k+1})+\mathcal{E}_{\alpha,k},
\end{aligned}$$

where $\mathcal{E}_{\alpha,k}=\Theta_k(\iota(\sqrt{d}^{-(k-1)}+\alpha^{k-3}\sqrt{d}^{-2})+\sqrt{d}^{-(k+1)})$, and $r_{t,k}=\mathbb{E}_{\alpha'\sim\rho_t^{\mathrm{MF}}}(\alpha')^k$.

Here we have used the same computations as in the proof of Lemma 43. Summing over all odd $k^*-1\le k\le K-1$ yields

$$\begin{aligned}
\frac{d}{d\alpha}v(\alpha,t) &= \sum_{k=k^*-1}^{K} q_k k\alpha^{k-1}\left(1-\frac{k+2}{k}\alpha^2\right)(1-r_{t,k+1})+\mathcal{E}_{\alpha,k} \qquad (39)\\
&= (k^*-1)q_{k^*-1}(1-r_t)\alpha^{k^*-2}+\Theta_K\left(\alpha^{k^*}+\iota(\sqrt{d}^{-(k^*-2)}+\alpha^{k^*-4}\sqrt{d}^{-2})+\sqrt{d}^{-(k^*-2)}\right)
\end{aligned}$$

Combining Lemma 43 with the previous equation, we obtain

$$\frac{d}{d\alpha}v(\alpha,t)=\frac{k^*-1}{\alpha}v(\alpha,t)+\Theta_K\left(\alpha^{k^*}+\iota(\sqrt{d}^{-(k^*-2)}+\alpha^{k^*-4}\sqrt{d}^{-2})+\sqrt{d}^{-(k^*-2)}\right).$$

This yields the first case in the conclusion of the lemma.

For the case that $\alpha\ge 1-\frac{1}{5K}\ge\sqrt{\frac{k}{k+0.5}}$ for all $k\le K$, we have

$$k\left(1-\frac{k+2}{k}\alpha^2\right)\le -1.5\alpha^2$$

We will compare the terms with coefficient q_k in the first line of Equation (39) and the first line of Equation (38). Let

$$v_k(\alpha,t):=q_k\alpha^k(1-\alpha^2)(1-r_{t,k+1}),$$

such that Equation (38) gives

$$v(\alpha,t)=\sum_{k=k^*-1}^{K-1}v_k(\alpha,t)+(1-\alpha^2)\mathcal{R}_\alpha,$$

where $\mathcal{R}_\alpha$ is as in Lemma 43. Thus the first line of Equation (39) gives

$$\begin{aligned}\frac{d}{d\alpha}v(\alpha,t) &= \sum_k (v_k(\alpha,t))\,\frac{1}{\alpha(1-\alpha^2)}k\left(1-\frac{k+2}{k\alpha^2}\right)+\mathcal{E}_{\alpha,k}\\ &\le \sum_k (v_k(\alpha,t))\,\frac{-1.5\alpha}{(1-\alpha^2)}+\mathcal{E}_{\alpha,k}\\ &= \frac{-1.5\alpha}{(1-\alpha^2)}\left(v(\alpha,t)-(1-\alpha^2)\mathcal{R}_\alpha\right)+\sum_k \mathcal{E}_{\alpha,k}\\ &\le -\frac{\alpha}{1-\alpha^2}v(\alpha,t)-\Omega_K(\delta).\end{aligned}$$

Here in the first inequality, we used the fact that all the c_k (and hence all the q_k and $v_k(\alpha,t)$) are non-negative. **[MG comments:** If we want to remove the non-negativity on the c_k assumption, will need to:

1. Leverage ($\star$) to show that $v_k(\alpha,t) \approx q_k\alpha^k(1-\alpha^2)(1-r)$ for some r.
2. Observe that at $\alpha = 1$ the desired conclusion holds (with stronger factor of 2 instead of 1.5); use continuity that it holds for α large enough for 1.5.

] In the last inequality, we have used the bounds on $\mathcal{R}_\alpha$ and $\mathcal{E}_{\alpha,k}$, along with the fact from Lemma 43 that $v(\alpha,t) = \Omega_K((1-\alpha^2)\delta)$. This yields the desired conclusion.

■

A key part of both our MF convergence analysis, and the perturbation analysis is understanding the stability of the $\alpha_t(w)$ with respect to small changes in $\alpha_s(w)$. The following lemma controls this derivative. Define

$$\ell_{t,s}(w) := \left.\frac{d\alpha_{t,s}(\beta)}{d\beta}\right|_{\beta=\alpha_s(w)}$$

Lemma 45 *Suppose that for all $s \le t$, we have $\sqrt{\mathbb{E}_x(f_{\rho_s^{\mathrm{MF}}}(x)-f^*(x))^2} \ge \delta$. Suppose $\iota \le \min\left(\Theta_K(1), \delta^{6K^2}\right)$, and $t \le \frac{\sqrt{d}^{k^*-2}}{\iota}$. Finally suppose ($\star$) holds for all $s \le t$. Then for and $\tau \le 1/2$ and any w for which $\alpha_t(w) \le 1-\tau$, we have*

$$\ell_{t,s}(w) := \left.\frac{d\alpha_{t,s}(\beta)}{d\beta}\right|_{\beta=\alpha_s(w)} = \left(\frac{\alpha_t(w)}{\alpha_s(w)}\right)^{k^*-1}\exp\left(O_K\left(\frac{\log(1/\tau)}{\delta}\right)\right).$$

Proof Observe that $\ell_{t,s}(w)$ satisfies the differential equation

$$\begin{aligned}\frac{d}{dt}\ell_{t,s}(w) &= \left(\frac{d}{d\alpha_t(w)}v(\alpha_t(w),t)\right)\ell_{t,s}(w);\\ \ell_{s,s}(w) &= 1.\end{aligned}$$

From Lemma 44, we have that

$$\frac{d}{dt}\ell_{t,s}(w) = \left((k^*-1)\frac{v(\alpha_t(w),t)}{\alpha_t(w)} + \mathcal{E}_\alpha\right)\ell_{t,s}(w);$$
$$\frac{d}{dt}\alpha_t(w) = \frac{v(\alpha_t(w),t)}{\alpha_t(w)}\alpha_t(w),$$

where we recall that

$$\mathcal{E}_\alpha = O_K\left(\alpha^{k^*} + \sqrt{d}^{-k^*} + \iota\left(\sqrt{d}^{-(k^*-2)} + \alpha_t^{(k^*-4)}\sqrt{d}^{-2}\right)\right)$$

Equivalently, taking logs, we have

$$\frac{d}{dt}\frac{\log(\ell_{t,s}(w))}{k^*-1} = \frac{v(\alpha_t(w),t)}{\alpha_t(w)} + \mathcal{E}_\alpha;$$
$$\frac{d}{dt}\log(\alpha_t(w)) = \frac{v(\alpha_t(w),t)}{\alpha_t(w)}.$$

Let us split up the time interval into (at most 3) intervals: $[s,t_1],[t_1,t_2],[t_2,t]$, where t_1 is first moment at which $\alpha_{t_1} \geq \frac{1}{\sqrt{d}}$, and α_{t_2} is the first moment at which $\alpha_{t_2} = 0.5$. In the first interval, we have $\mathcal{E}_\alpha \leq O_K(\iota\sqrt{d}^{(k^*-2)})$. In the second interval, by Lemma 43, we have $\mathcal{E}_\alpha \leq O_K\left(\sqrt{\iota}\frac{v(\alpha,t)}{\alpha_t^3}\sqrt{d}^{-2} + v(\alpha,t)\alpha/\delta\right)$.

For the first interval, since $t \leq \frac{\sqrt{d}^{k^*-2}}{\iota}$, we have

$$\int_{r=s}^{t_1}\mathcal{E}_{\alpha_r}dr \leq O_K(\iota\sqrt{d}^{-(k^*-2)})(t_1-s) \leq O_K(1).$$

For the second interval, using u-substitution, we have

$$\begin{aligned}\int_{r=t_1}^{t_2}\mathcal{E}_{\alpha_r}dr &\leq \frac{O_K(\sqrt{\iota})}{d}\int_{r=t_1}^{t_2}\frac{v(\alpha_r,r)}{(\alpha_r)^3}dr + \int_{r=t_1}^{t_2}O_K(v(\alpha_r,r)\alpha_r/\delta)dr\\ &= \frac{O_K(\sqrt{\iota})}{d}\int_{\alpha=\alpha_{t_1}}^{\alpha_{t_2}}\frac{1}{\alpha^3}d\alpha + \int_{\alpha=\alpha_{t_1}}^{\alpha_{t_2}}O_K(\alpha^2/\delta)d\alpha + O_K(\alpha_{t_2}^2)\\ &= \frac{O_K(\sqrt{\iota})}{d}\left(\frac{1}{2\alpha_{t_1}^2} - \frac{1}{2\alpha_{t_2}^2}\right) \leq O_K(1/\delta).\end{aligned}$$

For the third interval, observe from Lemma 43 that during the duration of this interval, $1-\alpha_r(w)$ decays exponentially with rate $O_K(\delta)$. Thus, the length of this interval is at most $O_K\left(\frac{\log(1/\tau)}{\delta^2}\right)$, so

$$\int_{r=t_2}^{t}\mathcal{E}_{\alpha_r}dr \leq O_K\left(\frac{\log(1/\tau)}{\delta}\right).$$

Thus integrating, we obtain

$$\frac{\log(\ell_{t,s}(w)) - \log(\ell_{s,s}(w))}{k^*-1} = \int_{r=s}^{t}\frac{v(\alpha_r(w),t)}{\alpha_r(w)}dr + O_K(\log(1/\tau)/\delta).$$

Plugging in the integration of the differential equation for $\log(\alpha_t(w))$ yields

$$\frac{\log(\ell_{t,s}(w))}{k^*-1} = \log\left(\frac{\alpha_t(w)}{\alpha_s(w)}\right) + O_K(\log(1/\tau)/\delta).$$

Multiplying both sides by k^*-1 and exponentiating yields

$$\ell_{t,s}(w) = \left(\frac{\alpha_t(w)}{\alpha_s(w)}\right)^{k^*-1} \exp\left(O_K\left(\frac{\log(1/\tau)}{\delta}\right)\right)$$

as desired. ■

Lemma 46 *For d large enough in terms of $\delta = \sqrt{\mathbb{E}_x(f_{\rho_t^{\mathrm{MF}}}(x) - f^*(x))^2}$, we have*

$$1 - \mathbb{E}_{\alpha'\sim\rho_t^{\mathrm{MF}}}(\alpha')^{k^*} \geq \Omega_{K,C_{reg}}(\delta).$$

Proof Observe that

$$\mathbb{E}_x(f^*(x))^2 = \mathbb{E}_x\sigma(w^{*\top}x)\sigma(w^{*\top}x) = q_\sigma(1).$$

$$\mathbb{E}_x f_{\rho_t^{\mathrm{MF}}}(x) f^*(x) = \mathbb{E}_x\mathbb{E}_{w'\sim\rho_t^{\mathrm{MF}}}\sigma(w'^{\top}x)\sigma(w^{*\top}x) = \mathbb{E}_{\alpha'\sim\rho_t^{\mathrm{MF}}} q_\sigma(\alpha').$$

Further

$$\mathbb{E}_x(f_{\rho_t^{\mathrm{MF}}}(x))^2 = \mathbb{E}_x\mathbb{E}_{w,w'\sim\rho_t^{\mathrm{MF}}}\sigma(w^\top x)\sigma(w'^\top x) = \mathbb{E}_{w,w'\sim\rho_t^{\mathrm{MF}}} q_\sigma(w^\top w').$$

Now for even k, we have

$$\mathbb{E}_{w,w'\sim\rho_t^{\mathrm{MF}}}(w^\top w')^k = \mathbb{E}_{\alpha,\alpha'\sim\rho_t^{\mathrm{MF}}}\mathbb{E}_\zeta(\alpha\alpha' + \sqrt{(1-\alpha^2)(1-\alpha')^2}\zeta)^k,$$

where ζ is $\frac{1}{\sqrt{d}}$-subGaussian. Thus by Minowski's inequality, we have

$$\begin{aligned}
\mathbb{E}_{w,w'\sim\rho_t^{\mathrm{MF}}}(w^\top w')^k &\leq \left(\left(\mathbb{E}_{\alpha,\alpha'\sim\rho_t^{\mathrm{MF}}}(\alpha\alpha')^k\right)^{1/k} + \frac{O_K(1)}{\sqrt{d}}\right)^k \\
&\leq \mathbb{E}_{\alpha,\alpha'\sim\rho_t^{\mathrm{MF}}}(\alpha\alpha')^k + \frac{O_K(1)}{\sqrt{d}} \\
&= \left(\mathbb{E}_{\alpha\sim\rho_t^{\mathrm{MF}}}\alpha^k\right)^2 + \frac{O_K(1)}{\sqrt{d}}.
\end{aligned}$$

It follows that with $q_\sigma(z) = \sum_k q_k z^k$, we have

$$\begin{aligned}
\mathbb{E}_x(f_{\rho_t^{\mathrm{MF}}}(x) - f^*(x))^2 &= \mathbb{E}_x(f^*(x))^2 + \mathbb{E}_x(f_{\rho_t^{\mathrm{MF}}}(x))^2 - 2\mathbb{E}f^*(x)f_{\rho_t^{\mathrm{MF}}}(x) \\
&= \sum_{k=k^*}^{K} q_k\left(1^k + \left(\mathbb{E}_{\alpha\sim\rho_t^{\mathrm{MF}}}\alpha^k\right)^2 - 2\mathbb{E}_{\alpha\sim\rho_t^{\mathrm{MF}}}\alpha^k\right) + \frac{O_K(1)}{\sqrt{d}} \\
&= \sum_{k=k^*}^{K} q_k\left(1 - \mathbb{E}_{\alpha\sim\rho_t^{\mathrm{MF}}}\alpha^k\right)^2 + \frac{O_K(1)}{\sqrt{d}}
\end{aligned}$$

Now for all $k > k^*$, with $1 - s := r := \mathbb{E}_{\alpha\sim\rho_t^{\text{MF}}}(\alpha)^{k^*}$, using ($\star$), we have

$$r^{\frac{k}{k^*}} \leq \mathbb{E}_{\alpha'\sim\rho_t^{\text{MF}}}(\alpha')^k,$$

so

$$\begin{aligned}\left(1 - \mathbb{E}_{\alpha\sim\rho_t^{\text{MF}}}\alpha^k\right)^2 &\leq \left(1 - r^{k/k^*}\right)^2 = \left(1 - (1-s)^{k/k^*}\right)^2 \\ &\leq (1 - (1 - sk/k^*)) = O_K(s^2).\end{aligned}$$

So

$$\mathbb{E}_x(f_{\rho_t^{\text{MF}}}(x) - f^*(x))^2 = \mathbb{E}_x(f^*(x))^2 = O_{K,C_{\text{reg}}}(1 - \mathbb{E}_{\alpha\sim\rho_t^{\text{MF}}}(\alpha)^{k^*}) + \frac{O_K(1)}{\sqrt{d}},$$

and thus for d large enough in terms of $\delta = \sqrt{\mathbb{E}_x(f_{\rho_t^{\text{MF}}}(x) - f^*(x))^2}$, we have $1 - \mathbb{E}_{\alpha\sim\rho_t^{\text{MF}}}(\alpha)^{k^*} = O_{K,C_{\text{reg}}}(\delta)$ as desired.

■

G.3. MF Convergence Analysis

Proposition 47 (Convergence of $f_{\rho_t^{\text{MF}}}$ to f^*) *Fix any δ small enough, and let $\iota = \delta^{6K^2}$. For d large enough, we have*

$$T(\delta) := \arg\min\{t : \mathbb{E}_x(f_{\rho_t^{\text{MF}}}(x) - f^*(x))^2 \leq \delta^2\} = O_K(\sqrt{d}^{k^*-2}\delta^{-(k^*-1)}).$$

We also have the following implication (which we will use for the analysis of J_{max} and J_{avg}) for any $t \leq T(\delta)$ and for any $\tau > 0$:

$$\mathbb{E}_{w\sim\rho_t^{\text{MF}}}[(\alpha(w))^{k^*-1}\mathbf{1}(\alpha(w) \leq 1 - \tau)] \leq \sqrt{d}^{-(k^*-2)}O_{K,\delta}\left(\frac{1}{\tau^{O_K(1)}}\right).$$

Proof First we need to prove by induction on t that for all $t \leq T(\delta)$, the hypothesis ($\star$) holds. First observe that it holds at time 0, because

$$\mathbb{P}_{w\sim\mathbb{S}^{d-1}}[\alpha(w) \geq \iota] \leq \exp(\Theta(d/\iota^2)) \leq \iota$$

for d large enough. Suppose the hypothesis holds up to some time s. We need to show that it holds at time $s + \epsilon$ for some ϵ. First note that for ϵ small enough, by the continuity of $v(\alpha, t)$ and $\frac{d}{d\alpha}v(\alpha, t)$, the conclusion of Lemma 43 and Lemma 44 still hold up to time t. To prove the hypothesis holds at time t, our approach will be to non-constructively bound the interval of $I \subset [0, 1]$ for which $\alpha_0(w) \notin I$ implies $\alpha_t(w) \notin [\iota, 1 - \iota]$. We will use the following claim.

Claim 11 *Suppose ($\star$) holds up to time t. For any $\tau \leq 1/2$ and $\gamma \leq \frac{1-\tau}{2}$, we have*

$$\mathbb{P}_{w\sim\rho_t^{\text{MF}}}[\alpha(w) \in [\gamma, 1 - \tau]] \leq \frac{2}{\gamma^{k^*-2}}\sqrt{d}^{-(k^*-2)}\exp\left(O_K\left(\frac{\log(1/\tau)}{\delta}\right)\right)$$

Proof We will show that

$$\mathbb{P}_{w\sim\rho_t^{\mathrm{MF}}}\left[\alpha(w)\in[\gamma,2\gamma]\right]\leq\frac{1}{\gamma^{k^*-2}}\sqrt{d}^{-(k^*-2)}\exp\left(O_K\left(\frac{\log(1/\tau)}{\delta}\right)\right)$$

The claim will then follow by summing this bound over $\log_2((1-\tau)/\gamma)$ intervals.

Suppose we have some w and w' with $\alpha_t(w),\alpha_t(w')\in[\gamma,2\gamma]$. Since the conditions of Lemma 45 hold up to time t for any particle $\tilde{w}$ with $\alpha_0(\tilde{w})$ initialized between $\alpha_0(w)$ and $\alpha_0(w')$, by the mean value theorem, we have that

$$\alpha_t(w)-\alpha_t(w')\geq\left|\alpha_0(w)-\alpha_0(w')\right|\min_{\tilde{w}:\alpha_0(\tilde{w})\in[\alpha_0(w'),\alpha_0(w')]}\left(\frac{\alpha_t(\tilde{w})}{\alpha_0(\tilde{w})}\right)^{k^*-1}\exp\left(O_K\left(\frac{\log\left(\tau/(k^*-1)\right)}{\delta}\right)\right)$$

$$\geq\left|\alpha_0(w)-\alpha_0(w')\right|\left(\frac{\gamma}{\alpha_0(w')}\right)^{k^*-1}\exp\left(O_K\left(\frac{\log\left(\tau\right)}{\delta}\right)\right),$$

Thus since $|\alpha_t(w)-\alpha_t(w')|\leq\gamma$, we have that

$$\left|\alpha_0(w)-\alpha_0(w')\right|\leq\frac{1}{\gamma^{k^*-2}}\left(\alpha_0(w')\right)^{k^*-1}\exp\left(O_K\left(\frac{\log\left(\tau\right)}{\delta}\right)\right).$$

We need to upper bound the probability over ρ_0 of the set in which $\alpha_0(w')$ and $\alpha_0(w)$ can lie. By the above calculation, the set which $\alpha_0(w')$ and $\alpha_0(w)$ lies in is contained in

$$I_\lambda:=\left[\frac{\lambda}{\sqrt{d}},\frac{\lambda}{\sqrt{d}}+\frac{1}{\gamma^{k^*-2}}\left(\frac{\lambda}{\sqrt{d}}\right)^{k^*-1}\exp\left(O_K\left(\frac{\log\left(\tau\right)}{\delta}\right)\right)\right]$$

for some λ. Recall that the distribution of $\alpha_0(w)$ under $w\sim\rho_0$ is $\frac{1}{\sqrt{d}}$-subGaussian. Thus

$$\begin{aligned}\mathbb{P}_{w\sim\rho_0}[\alpha_0(w)\in I_\lambda]&\leq\frac{\lambda^{k^*-1}}{\gamma^{k^*-2}}\sqrt{d}^{-(k^*-2)}\exp\left(O_K\left(\frac{\log(1/\tau)}{\delta^2}\right)\right)\left(\exp(-\lambda^2)\right)\\&\leq\frac{1}{\gamma^{k^*-2}}\sqrt{d}^{-(k^*-2)}\exp\left(O_K\left(\frac{\log(1/\tau)}{\delta}\right)\right).\end{aligned}$$

This proves the claim. ■

Plugging $\gamma=\iota$ and $\tau=\iota$ into this claim yields that

$$\mathbb{P}_{w\sim\rho_t^{\mathrm{MF}}}[\alpha(w)\in[\iota,1-\iota]]\leq\frac{2}{\iota^{k^*-2}}\sqrt{d}^{-(k^*-2)}\exp\left(O_K\left(\frac{\log(1/\tau)}{\delta}\right)\right)\leq\iota,$$

where the second inequality holds for d large enough in terms of δ. This proves the inductive step.

Now to prove the convergence guarantee, a standard analysis of the ODE for α (see eg. [DNGL23]) now yields that, for any w with $\alpha_0(w)\geq\frac{\delta^2}{\sqrt{d}}$, we have that

$$\alpha_t(w)\geq1-\frac{1}{2K}$$

for $t \geq \frac{\Theta(1)}{\delta^2(\alpha_0(w))^{k^*-2}}$. This arises directly from the fact that Lemma 43 guarantees that for $\alpha \geq \frac{\delta^2}{\sqrt{d}}$,

$$v(\alpha, t) \geq \Theta_K(\delta\alpha^{k^*-1}(1-\alpha^2)).$$

After that, it is clear that $1-\alpha_t(w)$ decays exponentially fast (with rate $\Omega(\delta)$), so for $t \geq \frac{\Theta(1)}{\delta(\alpha_0(w))^{k^*-2}} + O_K(\log(1/\delta)) = \frac{\Theta(1)}{\delta(\alpha_0(w))^{k^*-2}}$, we have $1-\alpha_t(w) \leq \delta/4$.

Now using the initial distribution of $\alpha_0(w)$ with $w \sim \rho_0$, we have that an at least $1-\delta/4$ fraction of particles have initialization $\alpha_0(w) \geq O_K(\frac{\delta}{\sqrt{d}})$. Clearly once all these particles achieve $1-\alpha_t(w) \leq 1-\delta/4$, we will have loss at most δ. Thus occurs at some time at most

$$\frac{\Theta_K(1)}{\delta(\delta\sqrt{d}^{-1})^{(k^*-2)}} = O_K(\sqrt{d}^{k^*-2}\delta^{-(k^*-1)}).$$

This proves the main statement of the proposition. To prove the additional clause, fix τ. We have

$$\begin{aligned}
\mathbb{E}_{w\sim\rho_t^{\mathrm{MF}}}[(\alpha(w))^{k^*-1}\mathbf{1}(\alpha(w) \leq 1-\tau)] &= \int_{\beta=0}^{1-\tau} \mathbb{P}_{w\sim\rho_t^{\mathrm{MF}}}[(\alpha(w))^{k^*-1} \in [\beta, (1-\tau)]d\beta. \\
&= \int_{\gamma=0}^{1-\tau} \mathbb{P}_{w\sim\rho_t^{\mathrm{MF}}}[\alpha(w) \in [\gamma^{\frac{1}{k^*-1}}, (1-\tau)^{\frac{1}{k^*-1}}]d\gamma. \\
&\leq \int_{\gamma=0}^{1-\tau} \frac{2}{\gamma^{\frac{k^*-2}{k^*-1}}}\sqrt{d}^{-(k^*-2)} \exp\left(O_K\left(\frac{\log(1/\tau)}{\delta}\right)\right) d\gamma \\
&= \sqrt{d}^{-(k^*-2)} \exp\left(O_K\left(\frac{\log(1/\tau)}{\delta}\right)\right) \int_{\gamma=0}^{1-\tau} \frac{2}{\gamma^{\frac{k^*-2}{k^*-1}}} d\gamma \\
&= \sqrt{d}^{-(k^*-2)} \exp\left(O_K\left(\frac{\log(1/\tau)}{\delta}\right)\right) 2(k^*-1)\gamma^{\frac{1}{k^*-1}}\Big|_0^{1-\tau} \\
&= \sqrt{d}^{-(k^*-2)} \exp\left(O_K\left(\frac{\log(1/\tau)}{\delta}\right)\right).
\end{aligned}$$

Here the inequality follows from Claim 11 and the fact that $(1-\tau)^{\frac{1}{k^*-1}} \geq 1 - \frac{\tau}{k^*-1}$. This proves the additional clause.

■

G.4. Proving Assumptions in Theorem 7 for Single-index Model

We need to check that the problem $(f^*, \mathcal{D}_x, \rho_0)$ introduced in Section G.1 satisfies the Assumptions of Theorem 7. Fix a desired loss δ, and let $T(\delta)$ be as in Proposition 47.

Local Strong Convexity.

Lemma 48 (Local Strong Convexity for SIM) *If d is large enough, then for any $t \leq T(\delta)$, we have for any w with $|\xi_t(w) - w^* \operatorname{sign}(\xi_t(w)^\top w^*)| \leq \frac{1}{5K}$,*

$$D_t^{\perp}(w) \preceq -\Omega_{K,C_{reg}}\left(\sqrt{L(\rho_t^{\mathrm{MF}})}\right).$$

Proof For simplicity, let $w_t := \xi_t(w)$, let $\alpha := \alpha(w_t)$. Assume that $\alpha \neq 1$; if $\alpha = 1$, we can take the limit of the calculations below.

Recall that

$$D_t^{\perp}(w) = \nabla_w \nu(w_t, \rho_t^{\mathrm{MF}})$$

It is evident that $\nu(w_t, \rho_t^{\mathrm{MF}})$ is in the direction $\tilde{w} := \sqrt{1-\alpha} w^* - \alpha w_{\perp}$, where $w_{\perp} = \frac{P_{w^*}^{\perp} w_t}{\|P_{w^*}^{\perp} w_t\|}$, and thus

$$\nu(w_t, \rho_t^{\mathrm{MF}}) = v(\alpha, t) \frac{\tilde{w}}{\sqrt{1-\alpha^2}}.$$

We will consider the quadratic form $y^\top D_t^{\perp}(w) y$ for $y \in \operatorname{span} \tilde{w}$ and for $y \perp \operatorname{span}(\xi_t(w), w^*)$. It suffices to show that for both such vectors we have $y^\top D_t^{\perp}(w) y \leq -\Omega_{K, C_{\mathrm{reg}}}\left(\sqrt{L(\rho_t^{\mathrm{MF}})}\right) \|y\|^2$.

Lets start with the first, letting $y = \tilde{w}$. We have

$$\begin{aligned} D_t^{\perp}(w) y &= \frac{d\nu(w, \rho_t^{\mathrm{MF}})}{d(y^\top w)} \\ &= \frac{v(\alpha, t)}{\sqrt{1-\alpha^2}} \frac{d\tilde{w}}{d(y^\top w_t)} + v(\alpha, t) \tilde{w} \frac{d(1-\alpha^2)^{-1/2}}{d(y^\top w_t)} + \left(\frac{\tilde{w}}{\sqrt{1-\alpha^2}}\right) \frac{dv(\alpha, t)}{d(y^\top w_t)} \end{aligned}$$

Now

$$\left(\frac{\tilde{w}}{\sqrt{1-\alpha^2}}\right) \frac{dv(\alpha, t)}{d(y^\top w_t)} = \left(\frac{\tilde{w}}{\sqrt{1-\alpha^2}}\right) \frac{dv(\alpha, t)}{d\alpha} \frac{d\alpha}{d(y^\top w_t)} = \tilde{w} \frac{dv(\alpha, t)}{d\alpha}.$$

Next,

$$\begin{aligned} \frac{d(1-\alpha^2)^{-1/2}}{d(y^\top w_t)} &= \frac{d(1-\alpha^2)^{-1/2}}{d\alpha} \frac{d\alpha}{d(y^\top w_t)} \\ &= \frac{-\alpha}{(1-\alpha^2)^{3/2}} \frac{1}{\sqrt{1-\alpha^2}} \\ &= \frac{\alpha}{(1-\alpha^2)}. \end{aligned}$$

Finally,

$$\frac{d\tilde{w}}{d(y^\top w_t)} = 0$$

Thus in summary, putting these three terms together we have

$$y^\top D_t^{\perp}(w) y = v(\alpha, t) \frac{\alpha}{(1-\alpha^2)} + \frac{dv(\alpha, t)}{d\alpha}.$$

By Lemma 44, we have for $y = \tilde{w}$,

$$y^\top D_t^{\perp}(w) y \leq -\Omega_{K, C_{\mathrm{reg}}}\left(\sqrt{L(\rho_t^{\mathrm{MF}})}\right).$$

Now we consider $y \perp \tilde{w}, w_t$. We have

$$\begin{aligned} y^\top \frac{d\nu(w_t, \rho_t^{\mathrm{MF}})}{d(y^\top w_t)} &= y^\top \tilde{w} \frac{d\left(\frac{v(\alpha,t)}{\sqrt{1-\alpha^2}}\right)}{dy^\top w} + \frac{v(\alpha,t)}{\sqrt{1-\alpha^2}} y^\top \frac{d\tilde{w}}{d(y^\top w_t)} \\ &= 0 + \frac{v(\alpha,t)}{\sqrt{1-\alpha^2}} y^\top \frac{d\tilde{w}}{d(y^\top w_t)} \\ &= -\alpha \frac{v(\alpha,t)}{\sqrt{1-\alpha^2}} y^\top \frac{dw_\perp}{d(y^\top w_t)} \\ &= -\alpha \frac{v(\alpha,t)}{\sqrt{1-\alpha^2}} y^\top \frac{y}{\sqrt{1-\alpha^2}} \\ &= -\frac{\alpha v(\alpha,t)}{1-\alpha^2} \|y\| \\ &\leq -\Omega_{K, C_{\mathrm{reg}}}\left(\sqrt{L(\rho_t^{\mathrm{MF}})}\right). \end{aligned}$$

Here the final inequality follows from Lemma 43. ■

Proving Assumption Stability for SIM. First we will need the following lemma. Recall that $\mathbf{J}_h$ denotes the Jacobian of a multivariate function h.

Lemma 49 *For any w and $s \leq t \leq T(\delta)$, we have*

$$\left\|\mathbf{J}_{\xi_{t,s}}(w_s)\right\| \leq O_K\left(\left(\frac{\alpha_t(w)}{\alpha_s(w)}\right)^{k^*-1}\right) \exp\left(O_K\left(\frac{1}{\delta}\right)\right).$$

Proof It suffices to check that this holds for times where $\alpha_t(w) \leq \frac{1}{5K}$, because after that, by Lemma 44, $D_t^\perp(w)$ is negative definite, and so $\left\|\mathbf{J}_{\xi_{t,s}}(\xi_s(w))\right\|$ can only decrease.

Claim 12 *In the setting of of the lemma, for any w with $\alpha_t(w) \leq \frac{1}{5K}$, we have*

$$\left\|\mathbf{J}_{\xi_{t,s}}(\xi_s(w))\right\| \leq O_K\left(\left.\frac{d\alpha_{t,s}(z)}{dz}\right|_{z=\alpha_s(w)}\right) + 1.$$

Proof Let $w_s = \xi_s(w)$. Without loss of generality assume $w_s^\top w^* > 0$ such that $\alpha_s(w) = \xi_s(w)^\top w^*$. Let $w_\perp := \frac{P_{w^*}^\perp w_s}{\|P_{w^*}^\perp w_s\|}$. We have

$$\xi_{t,s}(w_s) = \alpha_{t,s}(w_s) w^* + \sqrt{1-\alpha_{t,s}(w_s)^2} w_\perp.$$

Thus

$$\mathbf{J}_{\xi_{t,s}}(w_s) = \mathbf{J}_{\alpha_{t,s}}(w_s)(w^*)^\top + \frac{-\alpha_{t,s}(w_s)}{\sqrt{1-\alpha_{t,s}(w_s)^2}} \mathbf{J}_{\alpha_{t,s}}(w_s)(w_\perp)^\top + \frac{\sqrt{1-\alpha_{t,s}(w_s)^2}}{\sqrt{1-\alpha_s(w_s)^2}} P_{w^*}^\perp,$$

and so, since $\alpha_r(w)$ is increasing for $s \le r \le t$ if $\alpha_s(w) \ge \frac{1}{\sqrt{d}}$ (see Lemma 43) and $\alpha_t(w) \le 1 - \frac{1}{5K}$, we have

$$\left\|\mathbf{J}_{\xi_{t,s}}(w_s)\right\| \le O_K\left(\left\|\mathbf{J}_{\alpha_{t,s}}(w_s)\right\|\right) + 1.$$

■

The conclusion now follows from combining this claim and Lemma 45. ■

We are now ready to bound $J_{\max}$ and J_{avg}.

Lemma 50 *For any $t \le T(\delta)$, we have*

$$J_{max} \le O_{K,\delta}(\sqrt{d}^{2(k^*-1)})$$
$$J_{avg}(\tau) \le O_{K,\tau,\delta}(1/T(\delta)).$$

Proof By Lemma 49, for all w, we have

$$\|J_{t,s}^{\perp}(w)\| = O_{K,\delta}\left(\left(\frac{\alpha_t(w)}{\alpha_s(w)}\right)^{k^*-1}\right) \tag{40}$$

We bound this in two cases. Let $\iota = \delta^{6K^2}$. In the first case, if $\alpha_s(t) \ge \frac{\iota}{\sqrt{d}}$, then this is at most $O_{K,\delta}(\sqrt{d}^{k^*-1})$ as desired. In the second case, if $\alpha_s(w) \le \frac{\iota}{\sqrt{d}}$, then we can show that $\alpha_t(w)$ never exceeds $2\alpha_s(w)$. Indeed, one can inductively show by Equation (38) that for $s \le r \le t$, we have $v(\alpha_r, r) \le \iota^2 \sqrt{d}^{-(k^*-1)}$. Since $T(\delta) \le \frac{1}{\iota}\sqrt{d}^{k^*-2}$, we have $\alpha_t(w) \le 2\alpha_s(w)$. Thus in either case, we have $\|J_{t,s}^{\perp}(w)\| = O_{K,\delta}(\sqrt{d}^{k^*-1})$. The desired bound on $J_{\max}$ is immediate.

To bound J_{avg} we have to be more careful, and we will use an additional averaging lemma (Lemma 51) which allows us to show that when a set of neurons w are well-dispersed on the sphere at some time s, then on average over w, $H^{\perp}(w, w')$ is small for any w'.

$$\begin{aligned}
&\mathbb{E}_{w\sim\rho_0}\|J_{t,s}(w)H_s^{\perp}(w,w')v\|\mathbf{1}(\xi_t(w) \notin B_\tau) \\
&\quad= \mathbb{E}_{\alpha\sim\rho_s^{\text{MF}}}\mathbb{E}_{w\sim\rho_0|\alpha_s(w)=\alpha}\|J_{t,s}(w)H_s^{\perp}(w,w')v\|\mathbf{1}(\xi_t(w) \notin B_\tau) \\
&\quad\le \mathbb{E}_{\alpha\sim\rho_s^{\text{MF}}}\mathbf{1}(\alpha_{t,s}(\alpha) \le 1-\tau) \sup_{w|\alpha_s(w)=\alpha} \|J_{t,s}(w)\|\mathbb{E}_{w\sim\rho_0|\alpha_s(w)=\alpha}\|H_s^{\perp}(w,w')v\| \\
&\quad\le \mathbb{E}_{\alpha\sim\rho_s^{\text{MF}}}\mathbf{1}(\alpha_{t,s}(\alpha) \le 1-\tau) O_{K,\delta}\left(\frac{\alpha_t(w)}{\alpha_s(w)}\right)^{k^*-1}\left(\frac{\alpha_t(w)}{\alpha_s(w)}\right)^{k^*-1}\left(\alpha_s(w)^{k^*-1} + \sqrt{d}^{-(k^*-1)}\right)
\end{aligned}$$

Here the first inequality follows from the fact that the event $\xi_t(w) \notin B_\tau$ is equivalent to the even $\alpha_{t,s}(\alpha_s(w)) \le 1-\tau$. The second inequality is derived from (40) and Lemma 51.

Now to bound this expectation, recall the two cases from earlier in the lemma: $\alpha_s(w) \le \frac{\iota}{\sqrt{d}}$, and $\alpha_s(w) \ge \frac{\iota}{\sqrt{d}}$. Recall that in the first case, $\alpha_t(w) \le 2\alpha_s(w)$. Thus we have

$$\begin{aligned}
&\mathbb{E}_{\alpha\sim\rho_s^{\text{MF}}}\mathbf{1}(\alpha_{t,s}(\alpha) \le 1-\tau) O_{K,\delta}\left(\frac{\alpha_t(w)}{\alpha_s(w)}\right)^{k^*-1}\left(\frac{\alpha_t(w)}{\alpha_s(w)}\right)^{k^*-1}\left(\alpha_s(w)^{k^*-1} + \sqrt{d}^{-(k^*-1)}\right) \\
&\quad\le O_{K,\delta}\left(\sqrt{d}^{(k^*-1)}\right) + \mathbb{E}_{w\sim\rho_t^{\text{MF}}} O_{K,\delta}\left(\alpha(w)^{k^*-1}\right)\mathbf{1}(\alpha(w) \le 1-\tau).
\end{aligned}$$

The additional implication in Proposition 47 bounds this second term, yielding

$$\mathbb{E}_{w\sim\rho_0}\|J_{t,s}(w)H_s^{\perp}(w,w')v\|\mathbf{1}(\xi_t(w)\notin B_\tau) \quad \leq O_{K,\delta}\left(\sqrt{d}^{(k^*-1)}\right)+\sqrt{d}^{-(k^*-2)}O_{K,\delta}\left(\frac{1}{\tau^{O_K(1)}}\right)$$

$$= O_{K,\delta,\tau}\left(1/T(\delta)\right).$$

This proves the lemma. ■

Lemma 51 *For any distribution μ over w, for and $w', v \in \mathbb{S}^{d-1}$, with $w_s := \xi_s(w)$, we have*

$$\sup_{w',v}\mathbb{E}_{w\sim\mu}\|H_s^{\perp}(w,w')v\| \lessapprox \sup_{\|u\|=1}\sqrt{\mathbb{E}_{w\sim\mu}({w_s}^{\top}u)^{2(k^*-1)}}\|v\| + \sup_{\|u\|=1}\sqrt{\mathbb{E}_{w\sim\mu}({w_s}^{\top}u)^{2(k^*-2)}({w_s}^{\top}v)^2}.$$

In particular, if the distribution of w_s is rotationally symmetric in some set of dimensions, and has norm at most α if the remaining dimensions, then

$$\sup_{w',v}\mathbb{E}_{w\sim\mu}\|H_s^{\perp}(w,w')v\| \leq O_K\left(\alpha^{k-1}+\sqrt{d}^{-(k^*-1)}\right).$$

Proof [Proof of Lemma 51] By Cauchy-Schwartz,

$$\mathbb{E}_{w\sim\mu}\|H_s^{\perp}(w,w')v\| \leq \sqrt{\mathbb{E}_{w\sim\mu}v(H_s^{\perp}(w,w'))^{\top}H_s^{\perp}(w,w')v}.$$

Let us expand $H_s^{\perp}(w,w')$. With $w_s := \xi_s(w)$ and $u := \xi_s(w')$, we have

$$H_s^{\perp}(w,w') = \sum_{k=k^*-1}^{K-1} P_{w_s}^{\perp}\left(c(w,w')({w_s}^{\top}u)^k I + c'(w,w')({w_s}^{\top}u)^{k-1}u{w_s}^{\top}\right)P_u^{\perp},$$

where $c(w,w'), c'(w,w') \leq C_{\text{reg}}$. Thus we have

$$\begin{aligned}
&H_s^{\perp}(w,w')^{\top}H_s^{\perp}(w,w') \\
&\quad\preceq \sum_k 2C_{\text{reg}}({w_s}^{\top}u)^{2k}P_u^{\perp} \\
&\qquad + 2C_{\text{reg}}({w_s}^{\top}u)^{2(k-1)}P_u^{\perp}w_s u^{\top}P_{w_s}^{\perp}u{w_s}^{\top}P_u^{\perp} \\
&\quad\preceq 2C_{\text{reg}}({w_s}^{\top}u)^{2(k^*-1)}I \\
&\qquad + 2C_{\text{reg}}({w_s}^{\top}u)^{2(k^*-2)}w_s{w_s}^{\top},
\end{aligned}$$

and thus

$$\mathbb{E}_{w\sim\mu}v(H_s^{\perp}(w,w'))^{\top}H_s^{\perp}(w,w')v \leq 2C_{\text{reg}}({w_s}^{\top}u)^{2(k^*-1)}\|v\|^2 + 2C_{\text{reg}}({w_s}^{\top}u)^{2(k^*-1)}(v^{\top}w_s)^2.$$

Taking a square root yields the desired result. The second statement follows observing that $\mathbb{E}_w[(u^\top w_s)^k] = O_k(\sqrt{d}^{-k})$ if u is in the span of the rotationally invariant directions, because $u^\top w_s$ $\frac{1}{\sqrt{d}}$- subGaussian. ■

Proof [Proof of Theorem 9] Fix a desired loss δ, and let $T(\delta) = O_K(\sqrt{d}^{k^*-2}\delta^{-(k^*-1)})$ be as in Proposition 47, such that

$$\mathbb{E}_x(f_{\rho_t^{\text{MF}}}(x) - f^*(x))^2 \le \delta^2. \tag{41}$$

Let us check the conditions of Theorem 7. First, the regularity conditions in Assumption Regularity trivially hold for $C_{\text{reg}} = O_{C_{\text{SIM}}}(1)$ by our choice of Gaussian data and σ.

By Lemma 50, up to time $T(\delta)$, $(f^*, \rho_0, \mathcal{D}_x)$ satisfies Assumption Stability with $J_{\max} = O_{K,\delta}(d^{2(k^*-1)})$ and $J_{\text{avg}}(\tau) = O_{K,\delta,\tau}(1/T(\delta))$.

Observe that by Lemma 48, $(f^*, \rho_0, \mathcal{D}_x)$ is (c, τ) local strongly convex up to time $T(\delta)$ for $c = \Omega_{K,C_{\text{reg}}}(1)$, $\tau = \frac{1}{5K}$. Further, since the problem has rotational symmetry in all directions orthogonal to the w^* axis, the *structured* condition holds because by the smoothness of $\nabla_w \nu(w, \rho_t^{\text{MF}}) P_w^\perp$ in w, and the fact that at $\nabla_{\xi^\infty(w_i)} \nu(\xi^\infty(w_i), \rho_t^{\text{MF}}) P^\perp_{\xi^\infty(w_i)}$ (which approximates $D_t^\perp(i)$ to $C_{\text{reg}}\tau$ error) must be completely in the space orthogonal to w^*, and is rotationally symmetric in that space. Thus Assumption LSC holds.

Finally, the symmetry conditions in Assumption Symmetry trivially hold because the data is Guassian, and there is a reflection symmetry between w^* and $-w^*$.

Now suppose $n \ge d^{11k^*} \ge J_{\max}^8(T(\delta))^6 d^4$ and $m \ge d^{13k^*} \ge J_{\max}^{10}(T(\delta))^6 d^4$ such that

$$\epsilon_n + \epsilon_m = \frac{\log(n) d^{3/2}}{\sqrt{n}} + \frac{\log(mT)\max(d^{1/2}J_{\max}, d^{3/2})}{\sqrt{m}} \le \frac{1}{d J_{\max}^4 T^3}.$$

Thus for d large enough, the condition on $\epsilon_n + \epsilon_m$ in Theorem 7 holds. Thus all the assumptions of Theorem 7 hold, and the result guarantees that for $t \le T(\delta)$, with high probability over the draw of the data and of the neural network initialization, we have with $\lambda = \min(\tau, \delta)$,

$$\begin{aligned}\mathbb{E}_x(f_{\rho_t^{\text{MF}}}(x) - f_{\rho_t^m}(x))^2 &\le t J_{\max}(\epsilon_m + \epsilon_n) \exp\left(\frac{O(t J_{\text{avg}}(\lambda))}{c\lambda - \Omega(J_{\text{avg}}/\lambda)}\right) \\ &\le t d^{2(k^*-1)}(\epsilon_n + \epsilon_m) O_{K,\delta}(1).\end{aligned}$$

Combining this with Equation (41), we have that

$$\begin{aligned}\mathbb{E}_x(f^*(x) - f_{\rho_t^m}(x))^2 &\le 2\mathbb{E}_x(f_{\rho_t^{\text{MF}}}(x) - f_{\rho_t^m}(x))^2 + 2\mathbb{E}_x(f_{\rho_t^{\text{MF}}}(x) - f^*(x))^2 \\ &\le 2\delta^2 + 2t d^{2(k^*-1)}(\epsilon_n + \epsilon_m) O_{K,\delta}(1) \le 3\delta^2.\end{aligned}$$

This proves the theorem. ■

Appendix H. Full Details of Simulations

Name	Target Function	Activation/Network Design	LSC?	Symmetric?	J_{avg} assm?	C_{ρ^*}
He_4	$\mathsf{He}_4(x^\top e_1)$	$\sigma = \mathsf{He}_4$	Yes	Yes	Yes	1
Circle	$\mathbb{E}_{w\sim\mathbb{S}^1}\mathsf{He}_4(x^\top w)$	$\sigma = \mathsf{He}_4$	No	Yes	Yes	$\approx 2^4$
Misspecified	$0.8\mathsf{He}_4(x^\top e_1) + 0.6\mathsf{He}_6(x^\top e_1)$	$\sigma = \mathsf{He}_4 + \mathsf{He}_6$	No	No	Yes	$\approx d^4$
$\mathsf{Random}_{6,6}$	He_4 link, 6 random teachers in $\mathbb{R}^6$	$\sigma = \mathsf{He}_4$	Yes	No	Yes?	6
Staircase	$0.25x_1 + 0.75\mathsf{XOR}_4(x_{[4]})$	$\sigma =$ SoftPlus, 2nd layer ± 8	Yes	No	No	$\approx 2^8$
XOR_4	$\mathsf{XOR}_4(x_{[4]})$	$\sigma =$ SoftPlus, 2nd layer ± 8	Yes	No	?	$\approx 2^4$

Table 1: List of problem settings we empirically investigated.

H.1. Experimental Design

For each problem of interest, we simulated the training dynamics for several different widths $m \in [2^{12}, 2^{15}])$. We let M be twice the largest value of m. Crucially, *we initialized all the networks to be a subnetwork of the largest width network.* Further, we used the same training data and training procedure (hyperparameters, batch size, batch selection, etc.) for all values of m and M. We used the width M network as a proxy for the mean-field limit, and studied how the neurons in the smaller networks differed in their trajectories from their counterparts in the largest network. All experiments are repeated for 3 times. Source code is available at https://github.com/margalitglasgow/prop-chaos.

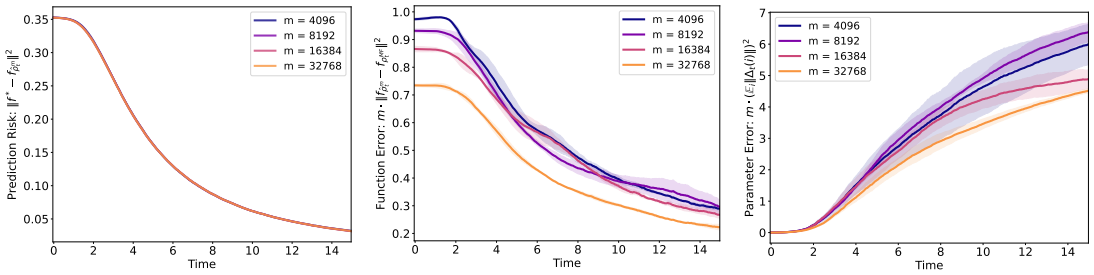
Training procedure. We optimized the neural network as follows.

1. We trained the models via mini-batch SGD with $n = 2^{16}$ total data points, and a batch size of 8196.
2. We used a step size of 0.01 (or occasionally smaller) for the problems with Gaussian data, and 0.05 for the problems with Boolean data. This was mainly because the Gaussian data had higher moments, and hence the loss occasionally exploded under large step size.
3. For the Gaussian single-/multi-index problems we used a Hermite activation function and all-1 second-layer weights, whereas in the Boolean experiments we used the SoftPlus activation with temperature 16 (which is a smooth approximation of ReLU), and we fixed the 2nd layer weights to $\pm C_k$ with equal probability, where $C_k = 2^k/\sqrt{k}$ for the k-parity problem.

Analysis procedure. We made the following measurements along the training dynamics.

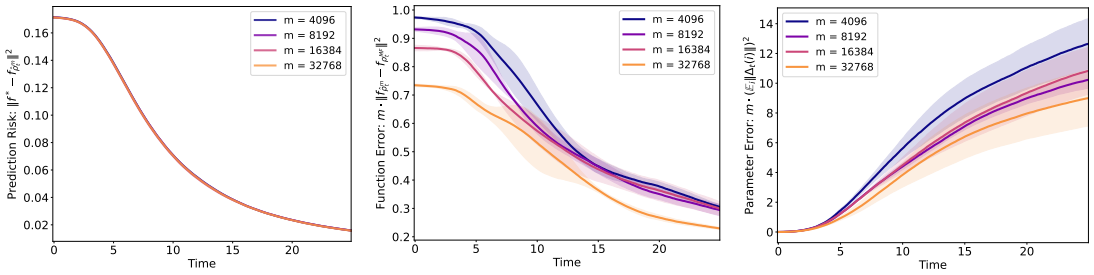
1. At each epoch, we computed the function error between the networks of width m and M, using a randomly sampled dataset of size n.
2. For each neuron i in the width-m network, we computed $\|\hat{\Delta}_t(i)\|$ as the norm of the difference between the neuron in the width-m network and the corresponding neuron in the width-M network.
3. We plot (a) the prediction risk curves, (b) the function error over time, and (c) $\mathbb{E}_i\|\hat{\Delta}_t(i)\|$ over time. In all the plots of the function and parameter error, we scaled up the error by the width m for better visualization.

H.2. Additional Experimental Results



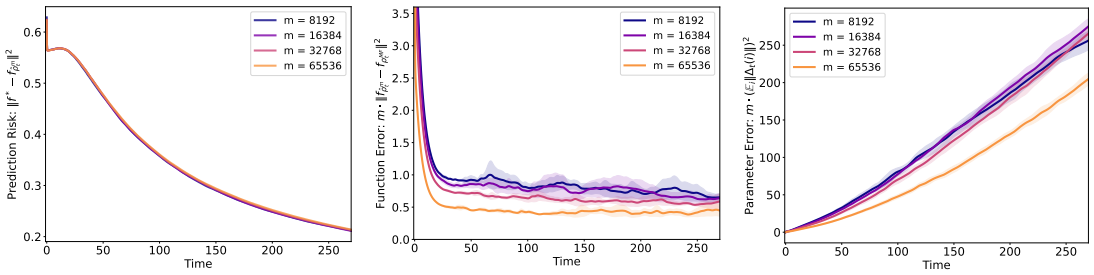
(a) prediction risk $\|f_{\rho_t^m} - f^*\|^2$. (b) scaled function error $m\|f_{\rho_t^m} - f_{\rho_t^{\mathrm{MF}}}\|^2$. (c) scaled parameter coupling error $m(\mathbb{E}_i\|\hat{\Delta}_t(i)\|)^2$.

Figure 7: Manifold (Circle) target function $f^*(x) = \mathbb{E}_{w\sim\mathbb{S}^1}\mathrm{He}_4(x^\top w), x \sim \mathcal{N}(0, I_d)$, and $\sigma = \mathrm{He}_4$ (ρ^* is distributed on a circle in 2 dimensions). We set $d = 64$ and learning rate $\eta = 0.01$.



(a) prediction risk $\|f_{\rho_t^m} - f^*\|^2$. (b) scaled function error $m\|f_{\rho_t^m} - f_{\rho_t^{\mathrm{MF}}}\|^2$. (c) scaled parameter coupling error $m(\mathbb{E}_i\|\hat{\Delta}_t(i)\|)^2$.

Figure 8: Additive ($\mathsf{Random}_{6,6}$) target function $f^*(x) = \frac{1}{6}\sum_{i=1}^{6}\mathrm{He}_4(x^\top w_i), x \sim \mathcal{N}(0, I_d)$, and $\sigma = \mathrm{He}_4$. We set $d = 64$ and learning rate $\eta = 0.01$.



(a) prediction risk $\|f_{\rho_t^m} - f^*\|^2$. (b) scaled function error $m\|f_{\rho_t^m} - f_{\rho_t^{\mathrm{MF}}}\|^2$. (c) scaled parameter coupling error $m(\mathbb{E}_i\|\hat{\Delta}_t(i)\|)^2$.

Figure 9: Staircase target function $f^*(x) = 0.25[x]_1 + 0.75\prod_{j\le 4}[x]_j, [x]_i \sim \mathrm{Unif}\{1, -1\}$, and $\sigma =$ SoftPlus with temperature 16. We set $d = 64$ and learning rate $\eta = 0.025$.