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

Margalit Glasgow, Denny Wu, Joan Bruna

February 18, 2025

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

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

$$f(x) = \frac{1}{m}\sum_{i=1}^{m}\sigma(\langle x, w_i\rangle), \quad w_1, w_2, ..., w_m \in \mathbb{S}^{d-1}, \tag{1.1}$$

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. However, two recent 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, DZPS19, 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].

In contrast, the mean-field analysis lifts (1.1) into the (infinite-dimensional) space of measures by considering the empirical distribution of neurons $\rho^m = m^{-1}\sum_{i=1}^m \delta_{w_i}$. Under certain regularity conditions, one can establish weak convergence of the empirical distribution to the limiting mean-field measure as the number of neurons tend to infinity: $\rho^m \overset{m\to\infty}{\to} \rho^{\mathrm{MF}}$, and the trajectory of 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, and global convergence

can be established in the mean-field limit ($m \to \infty$) by exploiting the convexity of the loss function (see the review paper [BC21]).

The goal of this work is to relate properties of the mean-field limit to a finite-width neural network, the learning dynamics of which 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 fall short of delivering this non-asymptotic control. On the one hand, *Exponential-in-time Grönwall-type* estimates leverage the regularity of the dynamics to propagate the Monte-Carlo error at initialisation (at scale $O(1/m)$) to obtain an estimate of the form $\sup_{t\in[0,T]}(f_{\rho_t}(x) - f_{\rho_t^m}(x))^2 \lesssim \exp T \cdot (m^{-1} \wedge \eta)$ 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 staircase functions [AAM22] and $T = O_d(\log d)$ for learning certain quartic polynomials [MZD+23]. Alternatively, *Uniform-in-time propagation of chaos* [HRSS19, NWS22, Chi22a] considered adding Gaussian noise to the gradient update (i.e., noisy GD) which gives the *mean-field Langevin dynamics* (MFLD). The previous exponential dependency on time can be removed under a uniform logarithmic Sobolev inequality [CRW22, SWN23, KZC+24, Nit24], but this ultimately transfers the exponential dependency to the runtime [SWON23, WMHC24, MHWE24]. Finally, [Chi22b, CRBVE20] establish uniform-in-time results, but in the asymptotic width limit.

Consequently, despite the feature learning advantage, the function class that can be learned by 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 $\log d$) 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 (i.e., multi-index model), 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, ABAM23]. We therefore ask the question

Can we identify sufficient and verifiable conditions under which the mean-field limit is well-approximated by $m = poly(d)$ neurons up to $T = poly(d)$ time horizon?

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 4), 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 2) 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 1 that if these conditions hold (along with several other regularity and technical conditions), then for $t \le T$,

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

This means that 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. We remark that we do not expect propogation of chaos to hold in the non-spherical setting, even for learning simple functions (see Remark 3).

In Theorem 2, we apply our result to a setting of learning a single index model (SIM) with information exponent $k^* \geq 4$, for which gradient flow converges in time $T = d^{\Theta(k^*)}$. First, we prove that in this setting, the limiting mean-field network, trained on the population, can learn the target function at time T. Then we use Theorem 1 to show that with $m, n = d^{\Theta(k^*)}$, at time T, $W_1(\rho_t^{\text{MF}}, \rho_t^m)$ is small, and thus the finite-width model ρ_t^m also achieves small population loss.

Notation $\mathcal{P}(\Omega)$ denotes the space of probability distributions over Ω. We use $W_1(\rho, \rho')$ to denote the 1-Wasserstein distance between two distributions ρ and ρ'. When $\hat{\mu}$ is an empirical measure of the form $\hat{\mu} = \frac{1}{m}\sum_i \delta_{w_i}$ which is clear from context, 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 use $P_w^{\perp} := (I - ww^T)$. 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 $H\Lambda(w) := \mathbb{E}_{w'\sim\mu} H(w, w')\Lambda(w')$. We use $D \odot \Lambda(w) = D(w)\Lambda(w)$.

Throughout this paper, we will 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 use "with high probability" to mean that the probability approaches 1 as m or n approaches 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

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

for a link function σ. We require that σ satisfies the regularity conditions in Assumption 1.

A 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 square 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:

$$\tfrac{d}{dt}w = V_{\mathcal{D}}(w, \rho) := -(I - ww^T)\nabla_w F_{\mathcal{D}}(w) + (I - ww^T)\nabla_w \mathbb{E}_{w'\sim\rho} K_{\mathcal{D}}(w, w') \tag{2.1}$$

where

$$F_{\mathcal{D}}(w) := \mathbb{E}_{(x,y)\sim\mathcal{D}} y\sigma(w^T x) \qquad \text{and} \qquad K_{\mathcal{D}}(w, w') := \mathbb{E}_{(x,y)\sim\mathcal{D}}\sigma(w'^T x)\sigma(w^T x).$$

In the case where we are training with 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 a setting, and when $\mathcal{D}_x$ is clear from context, we will use $V(w, \rho)$ (without any distribution subscripted) to denote the case where $x \sim \mathcal{D}_x$ and $y = f^*(x)$ deterministically. Whenever an expectation over x appears in this paper without any 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 $f^*(x) = \mathbb{E}_{w^*\sim\rho^*}\sigma(x^T w^*)$. 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 $\rho_t^{\text{MF}} \in \Delta(\mathbb{S}^{d-1})$, 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 $V(w, \rho_t^{\text{MF}}) \in T_w\mathbb{S}^{d-1}$. $\xi_t(w) \in \mathbb{S}^{d-1}$ denotes the characteristic of a particle initialized at w and evolved under the mean-field dynamics,

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

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

Finite-width, finite-data dynamics. Let ρ_t^m denote the dynamics of a distribution supported on the 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 the finite dynamics evolves according to the empirical velocity $V_{\hat{\mathcal{D}}}(w, \rho_t^m)$. We use $\hat{\xi}_t(w_i)$ to denote the location at time t of the particle initialized at w_i whose dynamics are given by

$$\tfrac{d}{dt}\hat{\xi}_t(w_i) = V_{\hat{\mathcal{D}}}(\hat{\xi}_t(w_i), \rho_t^m) \qquad \hat{\xi}_0(w_i) = w_i \ .$$

We will study the setting where the training data are drawn i.i.d. from an subgaussian distribution with subgaussian label noise (See 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 $V(\cdot, \rho_t^{\text{MF}})$. 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 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. *With high probability over the draw ρ_0^m, we have*

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

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 term 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*

$$D_t^\perp(w) := \left(\frac{d}{d\xi_t(w)}V(\xi_t(w), \rho_t^{\text{MF}})\right)(I - \xi_t(w)\xi_t(w)^T).$$

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

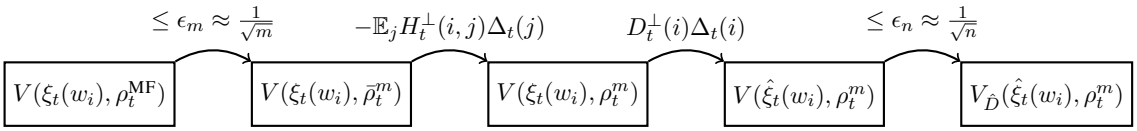

Figure 1: Decomposing $\frac{d}{dt}\Delta_t(i) = V(\xi_t(w_i), \rho_t^{\mathrm{MF}}) - V_{\hat{\mathcal{D}}}(\hat{\xi}_t(w_i), \rho_t^m)$. The approximate differences between the terms in the rectangles are given above the arrows.

Remark 1. *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 3 (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*

$$H_t^\perp(w, w') := \big(I - \xi_t(w)\xi_t(w)^T\big)\nabla_{\xi_t(w')}\nabla_{\xi_t(w)}K(\xi_t(w), \xi_t(w'))\big(I - \xi_t(w')\xi_t(w')^T\big),$$

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

Fact 4. *For any w, w', $H_t^\perp(w, w')$ is a positive semi-definite kernel.*

Proof. By definition of K in Equation 2.1, one can check that $H_t^\perp(w, w') = \mathbb{E}_x \phi_x(w)\phi_x(w')^T$, where we define the feature map $\phi_x(w) := (I - \xi_t(w)\xi_t(w)^T)\sigma'(\xi_t(w)^T x)x$ □

We assume the following basic regularity assumption on the activation function and the data.

Assumption 1 (Regularity Assumption)**.** ***R1*** *For a constant C_{reg}, the activation σ satisfies that for $j = 0, 1, 2, 3$ and any subgaussian variable X, we have $\mathbb{E}_X|\sigma^{(j)}(X)|^5 \leq C_{reg}/11$, where $\sigma^{(j)}$ denotes the jth derivative of σ.*

R2 *The distribution $\mathcal{D}_x$ on the covariates is subgaussian, and the noise has covariance at most 1, that is $\mathbb{E}_{y\sim\mathcal{D}|x}(y - f^*(x))^2 \leq 1$.*

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 17 that with high probability, the error $\|V(\xi_t(w_i), \rho_t^{\mathrm{MF}}) - V(\xi_t(w_i), \bar{\rho}_t^m)\|$ due to sampling only m neurons is uniformly (over i and t) bounded by ϵ_m. Similarly, we will show in Lemma 21 that the error $\|V_{\hat{\mathcal{D}}}(\hat{\xi}_t(w_i), \rho_t^m) - V(\hat{\xi}_t(w_i), \rho_t^m)\|$ due to using the empirical data distribution $\mathcal{D}$ is uniformly bounded by ϵ_n.

Lemma 5 (Parameter-Space Error Dynamics)**.** *Suppose Assumption 1 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 $\|\boldsymbol{\epsilon}_{t,i}\| \leq 2\epsilon_m + \epsilon_n + 2C_{reg}\big(\|\Delta_t(i)\|^2 + \mathbb{E}_j\|\Delta_t(j)\|^2\big)$.

We prove Lemma 5 by decomposing $\frac{d}{dt}\Delta_t(i) = V(\xi_t(w_i), \rho_t^{\mathrm{MF}}) - V(\hat{\xi}_t(w_i), \rho_t^m)$ 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 $J^{\perp}_{t,s}(w)$ to be the matrix that solves*

$$\tfrac{d}{dt}J^{\perp}_{t,s}(w) = D^{\perp}_t(w)J^{\perp}_{t,s}(w); \qquad J^{\perp}_{s,s}(w) = (I - \xi_s(w)\xi_s(w)^T).$$

We call this the local stability matrix, because $J^{\perp}_{t,s}(w) = \frac{d}{d\xi_s(w)}\xi_{t,s}(\xi_s(w))$, 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. We use the shorthand $J_{t,s}(i) := J_{t,s}(w_i)$.

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

$$\Delta_t(i) = \int_0^t J^{\perp}_{t,s}(i)\Big(-\mathbb{E}_j H^{\perp}_s(i,j)\Delta_t(j) + \boldsymbol{\epsilon_{s,i}}\Big)ds. \tag{2.2}$$

3 Main Result: Propagation of Chaos

3.1 Intuition and Key Challenges

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

$$\tfrac{d}{dt}\Delta_t(i) = D^{\perp}_t(i)\Delta_t(i) - \mathbb{E}_{j\sim[m]}H^{\perp}_t(i,j)\Delta_t(j) + \boldsymbol{\epsilon}_{t,i} \qquad \|\boldsymbol{\epsilon}_{t,i}\| \le \epsilon. \tag{3.1}$$

One might hope to leverage the linearity of (3.1) to solve this ODE in closed form, but unfortunately, the time-dependent coefficient matrix, $\mathrm{diag}(D^{\perp}_t) - H^{\perp}_t$, does not commute at different times t.

Going Beyond Grönwall. The conventional approach (see eg [MMN18]), uses the maximum Lipschitzness of $V(w,\rho)$ (in our spherical case, this translates to a bound on $\sup_{i,j,t}\|D^{\perp}_t\|, \|H^{\perp}_t(i,j)\|$) to bound the RHS of (3.1) as

$$\frac{d}{dt}\|\Delta_t(i)\| \le 2\,\mathrm{Lip}_{\max}\sup_{j\in[m]}\|\Delta_t(j)\| + \epsilon. \tag{3.2}$$

In standard settings, this maximum Lipschitzness is a constant, so this method can achieve no better than the bound $W_1(\rho^m_t, \rho^{\mathrm{MF}}_t) \le \exp(\Theta(t))\epsilon$. The work of [MZD+23] goes further to bound (3.2) using a tight time-dependent Lipschitz constant, yielding propagation of chaos for $\log(d)$ time. However, for problems with polynomial-in-d time to convergence, such as learning a SIM with a high information exponent, the approach in (3.2) 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 Fig. 2).

Equation (2.2) gives us an alternative way to approach (3.1) which can leverage the non-uniform Lipschitzness. Ignoring for a moment the interaction terms in Equation (2.2), we have $\|\Delta_t(i)\| \approx \int_0^t J^{\perp}_{t,s}(i)\boldsymbol{\epsilon_{s,i}}ds$, where we recall that the perturbation matrix $J^{\perp}_{t,s}(i)$ measures of the stability of $\xi_t(w)$ with respect to perturbations at time s. Naively, $J^{\perp}_{t,s}(w_i)$ appears to grow at an exponential rate whenever the local landscape of the linearized loss around $\xi_t(w_i)$ (see Remark 1) is non-convex.

A key observation of our work is that when w_i escapes certain higher-order saddles, $\|J^{\perp}_{t,s}(i)\|$ will be bounded polynomially in $t-s$. We achieve this by showing a certain *self-concordance*-like property which upper bounds $D^{\perp}_t(i)$ using the velocity (which is small near the saddle). Thus one part of our assumptions will be a worst-case polynomial bound on $\|J^{\perp}_{t,s}(w)\|$ (see Assumption 2).

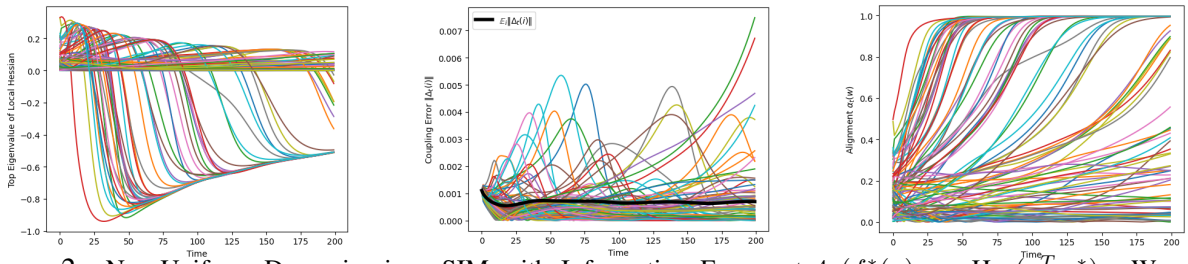

Figure 2: Non-Uniform Dynamics in a SIM with Information Exponent 4 ($f^*(x) = \mathrm{He}_4(x^T w^*)$. We plot $D_t^\perp(i), \|\Delta_t(i)\|, \alpha_t(w_i) = |w^* \xi_t(w_i)|$ for each neuron. Left: Top eigenvalue of the local hessians $D_t^\perp(i)$. Center: $\|\Delta_t(i)\|$, Right: Alignment $\alpha_t(w_i)$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

The Interaction Term: A Blessing and a Curse. At first glance, the presence of the PSD interaction term $H_t^\perp$ in (3.1) 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 $w_i^T w_j$ – 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 (2.2), 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 2 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 2 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 Appendix A. 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 2 (Worst-Case and Average Stability Assumption)**.** *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 $T > 0$,

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

Next, we will state our local strong convexity assumption. We remark that such an assumption can only hold when ρ^* is atomic (see Remark 2).

Assumption 3 (Local Strong Convexity (Abbreviated; see Assumption 4))**.** *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*

$$D^{\perp}_t(w) \preceq -C_{\mathrm{LSC}} P^{\perp}_{\xi_t(w)} \sqrt{\mathbb{E}_x (f_{\rho^{\mathrm{MF}}_t}(x) - f^*(x))^2}.$$

Both Assumption 2 and 3 are verifiable via solving the deterministic mean-field dynamics ρ^{MF}_t. For technical reasons, our result requires depends on two additional conditions. First, our theorem depends on the rank of the interaction Hessian as $\rho^{\mathrm{MF}}_t \to \rho^*$ 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:

$$C_{\rho^*} := \min\Big(|\mathrm{supp}(\rho^*)|, \dim(\mathrm{supp}(\rho^*))^{\mathrm{degree}(\sigma)} \Big).$$

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 Remark 4.

Second, we require a symmetry condition stated in Assumption 5 (in Appendix A). 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 position, or a ring of evenly spaced neurons in a circle. See Remark 6 for further discussion.

We are now ready to state the main theorem.

Theorem 1. *Suppose the Assumptions 1,2,4,5 hold up to time T (if relevant). Let C be a constant depending on $C_{\mathrm{LSC}}, \tau, \delta$ and C_{ρ^*}. Suppose n and m are large enough such that $J^4_{max} T^3 (\epsilon_n + \epsilon_m) \le 1/C$. Then with high probability over the draw ρ^m_0, for all $t \le T$,*

$$\mathbb{E}_x (f_{\rho^{\mathrm{MF}}_t}(x) - f_{\rho^m_t}(x))^2 \le 2 (W_1(\rho^m_t, \rho^{\mathrm{MF}}_t))^2 + \frac{2 C_{reg} \log(m)}{m} \le (C J_{max} t (\epsilon_m + \epsilon_n))^2.$$

where $\epsilon_m = \frac{\log(mT) \max(d^{1/2} J_{max}, d^{3/2})}{\sqrt{m}}, \epsilon_n = \frac{\sqrt{d} \log^2(n)}{\sqrt{n}}$, *and* $\delta := \sup_{s \le t} \sqrt{\mathbb{E}_x (f_{\rho^{\mathrm{MF}}_s}(x) - f^*(x))^2}$.

Theorem 1 follows directly from Lemma 1 and Corollary 42 in Section D. In Theorem 2, 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 2 (Local Strong Convexity)**.** *Our local strong convexity is similar to assumptions appearing in several other works on MF neural networks [Chi22c, Assumption A5][CRBVE20, Lemma D.9]. In comparison to the assumption 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 sucked into w^*, and thus any small perturbations are dampened. Local strong convexity can only *hold when ρ^* is atomic. Properties similar to local strong convexity have been shown for various sparse optimization problem over measures (eg. [FDGW21, PKP23]).*

3.3 Application to Single Index Model with High Information Exponent.

We will study the setting of learning a well-specified even single index function $f^*(x) = \sigma(x^T w^*)$, where $w^* \in \mathbb{S}^{d-1}$, and $\sigma(z) = \sum_{k=k^*}^{K} c_k \mathrm{He}_k(z)$, where (a) $k^* \geq 4$, and $\frac{1}{C_{\mathrm{SIM}}} \leq c_{k^*} \leq C_{\mathrm{SIM}} \max_k c_k$, (b) For all k, $c_k \geq 0$ (c) σ is an even function. We restrict to the case when $k^* = 4$, because the analysis for the setting where $k^* = 2$ has notable differences; namely, the escape times of the neurons is no longer non-uniform as in Figure 2(right). 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,

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

Theorem 2. *Fix any δ, and suppose d is large enough in terms of δ, C_{SIM} and K. Let $T(\delta) := \arg\min\{t : \mathbb{E}_x(f_{\rho_t^{\mathrm{MF}}}(x) - f^*(x))^2 \leq \delta^2\}$. Then $T(\delta) = O_{K,C_{SIM}}(\sqrt{d}^{k^*-2}\delta^{k^*-1})$. If $n \geq d^{11k^*}$ and $m \geq d^{13k^*}$, then with high probability, for all $t \leq T(\delta)$,*

$$\mathbb{E}_x(f_{\rho_t^{\mathrm{MF}}}(x) - f_{\rho_t^m}(x))^2 \leq \frac{O_{K,\delta}(d^{3k^*})}{\min(\sqrt{m}, \sqrt{n})}.$$

Thus, $\mathbb{E}_x(f_{\rho_t^{\mathrm{MF}}}(x) - f_{\rho_t^m}(x))^2 \leq 3\delta^2$.

4 Overview of Proof Ideas

4.1 Potential-Based Analysis to Prove Theorem 1

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 the 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 for the 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 $\rho_t^{\mathrm{MF}} \approx \rho^*$, the $D_t(i)$ are negative definite, so $\frac{d}{dt}\mathbb{E}_i\|\Delta_t(i)\|^2 \approx -\Delta_t^T H_t^\perp \Delta_t - 2\mathbb{E}_i\Delta_t(i)^T D_t(i)\Delta_t(i) \leq 0$. However, such a function doesn't satisfy **P3** whenever there is a lot of variance among the $\|\Delta_t(i)\|$. This turns out to be a major issue.

To achieve **P3**, intuitively, the potential function 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. As an example, consider the case where $\rho^* = \delta_{w^*}$, and thus near convergence, $H_t^\perp \approx \mathbf{1}\mathbf{1}^T \otimes P_{w^*}^\perp$, 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 example 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 7), which closely approximates $H_t^\perp$ on neurons in B_τ and avoids tracking the temporal evolution of the eigendecomposition. This ultimately allows us to leverage the PSD structure of $H_t^\perp$.

Definition 7. *Define $H_\infty^\perp$ explicitly in the following way.*

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

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 $\langle f, g\rangle_{\mathcal{Z}} = \mathbb{E}_{w\sim\rho_0} f(w)^T g(w)$. 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 D.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 $\{\varphi_j\}_{j\le J}$:

$$\overline{H_\infty^\perp} = \sum_{j\le J} \lambda_j \varphi_j \otimes \varphi_j \ , \ \lambda_j \in \mathbb{R} \ , \ \varphi_j \in \mathcal{Z} \ , \tag{4.1}$$

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 8 (Balanced Spectral Decomposition of $H_\infty^\perp$ (WED))**.** *We say that the spectral decomposition (4.1) 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

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

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 $\Phi_{\mathcal{Q}}(t) = \Phi_{\mathcal{Q}}(\Delta_t)$.

Lemma 9 (Balanced Spectral Decomposition)**.** *Suppose Assumption 5 holds. Then there exists an spectral distribution $\mathcal{Q}$ which is $C_{\rho^*} = \min\big(|\operatorname{supp}(\rho^*)|, \dim(\operatorname{supp}(\rho^*))^{\operatorname{degree}(\sigma)}\big)$-balanced.*

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

Lemma 10 (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 19 holds for $S = B_\tau$, we have*

$$\langle \nabla\Phi_{\mathcal{Q}}(t), -H_t^\perp \Delta_t\rangle \le (1 + C_b)\mathbb{E}_i\|\mathbb{E}_j H_t^\perp(i,j)\|\mathbf{1}(\xi_t(w_i) \notin B_\tau) + \mathcal{E}_{10},$$

where $\mathcal{E}_{10} = O_{C_{reg},C_b}(\mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i) \notin B_\tau) + (\tau + C_b\epsilon_m^{19})\Omega(t))$.

Lemma 11 (Descent with Respect to Local Term)**.** *Suppose Assumption 4 holds with* (C_{LSC}, τ)*. Let* $\mathcal{Q}$ *be a* C_b*-balanced spectral distribution. Then with* $C_{11} = O_{C_{reg}, C_b}(1)$*, we have*

$$\langle \nabla \Phi_{\mathcal{Q}}(t), D_t^{\perp} \odot \Delta_t \rangle \le -\left(\frac{c\sqrt{L_{\mathcal{D}}(\rho_t^{\mathrm{MF}})}}{2} - C_{11}\tau \right) \Phi_{\mathcal{Q}}(t) + C_{11}\mathbb{E}_i \|\Delta_t(i)\| \mathbf{1}(\xi_t(w_i) \notin B_\tau) + C_b \mathbb{E}_i \|\Delta_t(i)\|^2.$$

Lemma 12 (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 2 allows us to bound the dynamics of the potential function in the following way (formalized in Theorem 3):

$$\frac{d}{dt}\Phi_{\mathcal{Q}}(t) \le -\frac{C_{\mathrm{LSC}}\sqrt{L(\rho_t^{\mathrm{MF}})}}{C}\Phi_{\mathcal{Q}}(t) + C J_{\mathrm{avg}} \int_{s=0}^{t} \Phi_{\mathcal{Q}}(s) ds + C J_{\max}(\epsilon_m + \epsilon_n), \tag{4.2}$$

where $C = O_{C_{\rho^*}, C_{\mathrm{reg}}}(1)$. Theorem 1 follows by analyzing this differential equation. We leverage Assumption 2 to prove (4.2), by bounding the term $\mathbb{E}_i \|\Delta_t(i)\| \mathbf{1}(\xi_t(w_i) \notin B_\tau)$ which arises from Lemmas 10 and 11. Indeed, using the closed form for $\Delta_t(i)$ given in Equation 2.2, we can expand

$$\begin{aligned}
\mathbb{E}_i \|\Delta_t(i)\| \mathbf{1}(\xi_t(w_i) \notin B_\tau) &\approx \mathbb{E}_i \int_{s=0}^{t} \|J_{t,s}^{\perp}(i) \mathbb{E}_j H_s^{\perp}(i,j) \Delta_s(j)\| \mathbf{1}(\xi_t(w_i) \notin B_\tau) ds \\
&\le \int_{s=0}^{t} \mathbb{E}_j \|\Delta_s(j)\| \sup_{v \in \mathbb{S}^{d-1}} \mathbb{E}_i \int_{s=0}^{t} \|J_{t,s}^{\perp}(i) H_s^{\perp}(i,j) v\| \mathbf{1}(\xi_t(w_i) \notin B_\tau) ds \\
&\lessapprox J_{\mathrm{avg}} \int_{s=0}^{t} \mathbb{E}_j \|\Delta_s(j)\| ds \\
&\le J_{\mathrm{avg}} \int_{s=0}^{t} \Phi_{\mathcal{Q}}(s) ds.
\end{aligned}$$

Here the third line follows from Assumption 2 (along with a concentration argument in Lemma 18).

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

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

Observation 13. *When the velocity* $V(w, \rho_t^{\mathrm{MF}})$ *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 $f^*(x) = \sigma(x^T w^*)$ 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 $\alpha_t := w_t^T w^*$ is small (and assume for simplicity that α_t is positive), we have that

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

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

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

Recalling the differential equation of $J_{t,s}^{\perp}$, we have just shown that $\frac{d}{dt}\|J_{t,s}^{\perp}\| \le \frac{(k-1)}{\alpha_t} V(\alpha_t) \|J_{t,s}^{\perp}\|$. 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 $\|J_{t,s}^{\perp}\| \le \left(\frac{|\alpha_t|}{|\alpha_s|}\right)^{k-1}$; see Lemma 54.

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

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

where $T = \Theta(d^{(k-2)/2})$ is the desired training time. We briefly give some intuition for why this holds in the single-index model $f^*(x) = \mathrm{He}_k(x^T w^*)$, which requires $T = \Theta(d^{(k-2)/2})$. To tightly bound $J_{\text{avg}}(\tau)$, we need 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 $\alpha_t(w) := |w^{*T}\xi_t(w)|$) we have

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

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

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

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 49).

5 Conclusion and Discussion

We have studied propagation-of-chaos in the context of two-layer neural network training. By leveraging several key geometric assumptions of the optimization landscape, we have 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 models. 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 case of learning a misspecified SIM. Another interesting question is how to go beyond the Monte-Carlo scale of fluctuations, which is known to hold asymptotically under certain conditions [CRBVE20, Theorem 3.5].

Remark 3 (Spherical Constraint)**.** *We remark that when the weights of the neural network are not constrained to the sphere, propagation of chaos fails even in simple settings: to see this, consider the case of learning a SIM with information exponent $k > 2$. With polynomial width, we require the standard $T \approx d^{(k-2)/2}$ time. However, with infinite neurons, only a $o(1)$ fraction of neurons are required to learn the feature since they can become a disproportionate mass of the network, so we can achieve $T = d^{(k-1)/(k+2)}$ convergence time, by leveraging only the neurons with initial alignment $\ge d^{\frac{-2}{k+2}}$.*

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.

[ABAM23] Emmanuel Abbe, Enric Boix-Adsera, and Theodor Misiakiewicz. Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics. *arXiv preprint arXiv:2302.11055*, 2023.

[AZLS19] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A Convergence Theory for Deep Learning via Over-Parameterization. 2019.

[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.

[BC21] Francis Bach and Lénaïc Chizat. Gradient descent on infinitely wide neural networks: Global convergence and generalization. *arXiv preprint arXiv:2110.08084*, 2021.

[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. *arXiv preprint arXiv:2205.01445*, 2022.

[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.

[Chi22a] Lénaïc Chizat. Mean-field langevin dynamics: Exponential convergence and annealing. *arXiv preprint arXiv:2202.01009*, 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.

[COB19] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On Lazy Training in Differentiable Programming. 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.

[CRW22] Fan Chen, Zhenjie Ren, and Songbo Wang. Uniform-in-time propagation of chaos for mean field langevin dynamics. *arXiv preprint arXiv:2212.03050*, 2022.

[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.

[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.

[DZPS19] Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient Descent Provably Optimizes Over-parameterized Neural Networks. 2019.

[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.

[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.

[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.

[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.

[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. 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.

[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.

[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.

[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.

[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.

[WMHC24] Guillaume Wang, Alireza Mousavi-Hosseini, and Lénaïc Chizat. Mean-field langevin dynamics for signed measures via a bilevel approach. *arXiv preprint arXiv:2406.17054*, 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

A Full Statement of Assumptions and Remarks

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

$$C_{\rho^*} := \min\Big(|\text{supp}(\rho^*)|, \dim(V)^{\text{degree}(\sigma)}\Big).$$

Assumption 4 (Local Strong Convexity (Full Version of Assumption 3))**.** *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*

$$D_t^\perp(w) \preceq -C_{\text{LSC}} P^\perp_{\xi_t(w)} \sqrt{\mathbb{E}_x (f_{\rho_t^{\text{MF}}}(x) - f^*(x))^2}.$$

Further, the strong convexity is structured*, meaning there exist values* $c_t^1, c_t^2 \ge c$ *such that for any* w *with* $\xi_t(w) \in B_\tau$*, we have*

$$\|c_t^1 V V^T P^\perp_{\xi^\infty(w)} V V^T + c_t^2 U U^T - D_t^\perp(w)\| \le \left(\frac{C_{\text{LSC}} \sqrt{\mathbb{E}_x (f_{\rho_t^{\text{MF}}}(x) - f^*(x))^2}}{2\sqrt{C_{\rho^*}}} + C_{reg}\tau \right).$$

Assumption 5 (Symmetries of ρ^*)**.** *The* automorphism group $\mathcal{G}$ *of a problem* $(\rho^*, \mathcal{D}_x, \rho_0)$ *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*** $\text{supp}(\rho^*)$ *is* transitive *under* $\mathcal{G}$*, that is, for any* $w^*, w^{*\prime} \in \text{supp}(\rho^*)$*, there exists* $g \in \mathcal{G}$ *such that* $g(w^*) = w^{*\prime}$*. Further,* $\mathbb{P}_{w\sim\rho_0}[\{\|w - w^*\| = \|w - w^{*\prime}\| \exists w^*, w^{*\prime} \in \text{supp}(\rho^*)\}] = 0$*.*
- ***I2*** *Let* $V = \text{span}(\text{supp}(\rho^*))$ *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* $\mathbb{E}_{x_U\sim\mathcal{D}_U} x x^T = UU^T$*.*

Remark 4 (Dependence on C_{ρ^*}, and *structured* assumption.)**.**

Remark 5 (The *structured* condition in Assumption 4)**.**

Remark 6 (Symmetry Assumption)**.**

B Proofs of Lemmas from Basic Setup

B.1 Notation.

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

$$F(w) := \mathbb{E}_{x\sim\mathcal{D}_x} f^*(x)\sigma(w^T x)$$
$$F'(w) := (I - ww^T)\nabla_w F(w)$$

and

$$K(w, w') := \mathbb{E}_{x\sim\mathcal{D}_x} \sigma(w'^T x)\sigma(w^T x)$$
$$K'(w, w') := (I - ww^T)\nabla_w K(w, w').$$

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

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

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}:

$$\bar{D}_t(w) := \frac{d}{d\xi_t(w)} V(\xi_t(w), \bar{\rho}_t^m) = \nabla_{\xi_t(w)} F'(\xi_t(w)) - \mathbb{E}_{w' \sim \bar{\rho}_t^m} \nabla_{\xi_t(w)} K'(\xi_t(w), w').$$
$$D_t^\perp(w) = \frac{d}{d\xi_t(w)} V(\xi_t(w), \rho_t^{\mathrm{MF}}) = \nabla_{\xi_t(w)} F'(\xi_t(w)) - \mathbb{E}_{w' \sim \rho_t^{\mathrm{MF}}} \nabla_{\xi_t(w)} K'(\xi_t(w), w').$$

B.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 14. *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 $\left\|\frac{d}{dw} K'(w, w')\right\| \le C_{reg}$ *and* $\left\|\frac{d}{dw} F'(w)\right\| \le C_{reg}$

S2 $\left\|\frac{d^2}{dw'^2} K'(w, w')\right\| \le C_{reg}$

S3 $\left\|\frac{d^2}{dw^2} K'(w, w')\right\| \le C_{reg}$

S4 $\left\|\frac{d}{dw'}\frac{d}{dw} K'(w, w')\right\| \le C_{reg}$

S5 $\left\|\frac{d^2}{dw^2} F'(w)\right\|_{op} \le C_{reg}$

S6 *For any distribution $\rho \in \Delta(\mathbb{S}^{d-1})$, we have* $\left\|\frac{d^2}{dw^2} V(w, \rho)\right\|_{op} \le C_{reg}$

S7 $\mathbb{E}_{x \sim \mathcal{D}_x}\left[\mathrm{Lip}(\sigma(\langle \cdot, x\rangle))^2\right] \le C_{reg}$

Proof. [Proof of Lemma 14] These are straightforward to check from the definitions. First note that the operator norm of the first and second derivatives of $I - ww^T$ is at most 2. Thus for any function $G(w)$, by chain rule, we have

$$\left\|\frac{d}{dw}(I - ww^T)G(w)\right\| \le \left\|\frac{d}{dw}G(w)\right\| + 2\|G(w)\|$$
$$\left\|\frac{d^2}{dw^2}(I - ww^T)G(w)\right\| \le 3\left\|\frac{d^2}{dw^2}G(w)\right\| + 8\left\|\frac{d}{dw}G(w)\right\|.$$

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

$$\|\nabla_w F(w)\|, \left\|\frac{d^2}{dw^2}F(w)\right\|, \left\|\frac{d^3}{dw^3}F(w)\right\|,$$

and

$$\|\nabla_w K(w,w')\|, \left\|\frac{d^2}{dw^2}K(w,w')\right\|, \left\|\frac{d^3}{dw^3}K(w,w')\right\|, \left\|\frac{d^2}{dwdw'}\nabla_w K(w,w')\right\|, \left\|\frac{d^3}{dw^2dw'}\nabla_w K(w,w')\right\|$$

As an example, for **S2**, we have

$$\begin{aligned}\left\|\frac{d^2}{dw'^2}K'(w,w')\right\|_{op} &\leq \sup_{v_2,v_2,v_3\in\mathbb{S}^{d-1}} \mathbb{E}_x \sigma(w^T x)\sigma'''(w'^T x)v_1^T(I-ww^T)x(v_2^T x)(v_3^T x)\\ &\leq \sup_{z,z'\in B_2^d}\left(\mathbb{E}_x|\sigma(z^T x)|^5\right)^{1/5}\left(\mathbb{E}_x|\sigma'''(z'^T x)|^5\right)^{1/5}\sup_{v\in\mathbb{S}^{d-1}}\left(\mathbb{E}_x|(v^T x)|^5\right)^{3/5}\\ &\leq C_{\text{reg}}/11,\end{aligned}$$

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 $(I-ww^T)$, 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 identitcal, 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 thrid order term. We have

$$\begin{aligned}\left\|\frac{d^3}{dw^3}F(w)\right\| &\leq \sup_{v_1,v_2,v_3\in\mathbb{S}^{d-1}} \mathbb{E}_x|\sigma^{(3)}(w^T x)(v_1^T x)(v_2^T x)(v_3^T x)f^*(x)|\\ &\leq \sup_{z,z'\in B_2^d}\left(\mathbb{E}_x|\sigma^{(3)}(z^T x)|^5\right)^{1/5}\sup_{v\in\mathbb{S}^{d-1}}\left(\mathbb{E}_x|(v^T x)|^5\right)^{3/5}\left(\mathbb{E}_x(f^*(x))^5\right)^{1/5}\\ &\leq C_{\text{reg}}/11.\end{aligned}$$

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

□

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

Lemma 15. *With high probability over the draw ρ_0^m, we have*

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

Proof. [Proof of Lemma 1] Given coupling $\pi \in \Pi(\rho_t^{\text{MF}}, \rho_t^m)$, we may write

$$f_{\rho_t^{\text{MF}}}(x) - f_{\rho_t^m}(x) = \mathbb{E}_{w,w'\sim\pi}[\sigma(x^\top w) - \sigma(x^\top w')].$$

Jensen's inequality on the square yields

$$\mathbb{E}_x[(f_{\rho_t^{\text{MF}}}(x) - f_{\rho_t^m}(x))^2] \leq \mathbb{E}_{w,w'\sim\pi}\mathbb{E}_x[(\sigma(x^\top w) - \sigma(x^\top w'))^2].$$

We proceed to control the inner expectation via Taylor expansion,

$$\mathbb{E}_x[(\sigma(x^\top w) - \sigma(x^\top w'))^2] \leq \mathbb{E}_x\left[((w-w')^\top x)^2\int_0^1(\sigma'(tw+(1-t)w')^\top x)^2 dt\right] \leq C\|w-w'\|^2,$$

for some constant $C > 0$, where the last inequality follows from Assumption **R1** and Cauchy-Schwarz. Now due to the spherical constraint, $\mathbb{E}_{w,w'\sim\pi}\|w-w'\|^2 \leq 4\mathbb{E}_{w,w'\sim\pi}\|w-w'\|$, and thus taking an infimum over coupling yields

$$\mathbb{E}_z[(f_{\rho_t^{\text{MF}}}(x) - f_{\rho_t^m}(x))^2] \leq C\cdot W_1(\rho_t^{\text{MF}}, \rho_t^m).$$

We conclude the proof by Lemma 14.

□

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

Lemma 16 (Parameter-Space Error Dynamics)**.** *Suppose Assumption 1 holds. With high probability, for all $t \le T$ 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 $\|\boldsymbol{\epsilon}_{t,i}\| \le 2\epsilon_m + \epsilon_n + 2C_{reg}\big(\|\Delta_t(i)\|^2 + \mathbb{E}_j\|\Delta_t(j)\|^2\big)$.

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

$$\begin{aligned}
\frac{d}{dt}\Delta_t(i) &= V(\xi_t(w_i), \rho_t^{\mathrm{MF}}) - V(\hat{\xi}_t(w_i), \rho_t^m) \\
&= \big(V(\xi_t(w_i), \rho_t^{\mathrm{MF}}) - V(\xi_t(w_i), \bar{\rho}_t^m)\big) + \big(V(\xi_t(w_i), \bar{\rho}_t^m) - V(\xi_t(w_i), \rho_t^m)\big) \\
&\quad + \Big(V(\xi_t(w_i), \rho_t^m) - V(\hat{\xi}_t(w_i), \rho_t^m)\Big) + \Big(V(\xi_t(w_i), \rho_t^m) - V_{\hat{\mathcal{D}}}(\hat{\xi}_t(w_i), \rho_t^m)\Big).
\end{aligned}$$

By Lemma 17 and Lemma 21, we can bound the first and fourth terms respectively with high probability:

$$\begin{aligned}
\|V(\xi_t(w_i), \rho_t^{\mathrm{MF}}) - V(\xi_t(w_i), \bar{\rho}_t^m)\|_2 &\le \epsilon_m. \\
V(\xi_t(w_i), \rho_t^m) - V_{\hat{\mathcal{D}}}(\hat{\xi}_t(w_i), \rho_t^m) &\le \epsilon_n.
\end{aligned} \tag{B.1}$$

For the second term, we have

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

where $\|\mathbf{v}_j\| \le C_{\mathrm{reg}}\|\Delta_t(j)\|^2$. Indeed we can plug Lemma 14 **S2** into the Lagrange error bound

$$\|K'(w,w') - K'(w,w'+\Delta) - \frac{d}{dw'}K'(w,w')\Delta\| \le \|\Delta\|^2 \sup_{w':\|w'\|\le 1}\left\|\frac{d^2}{dw'^2}K'(w,w')\right\|.$$

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

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

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

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

Finally for the third term, we have

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

where by **S6**,

$$\|\mathbf{v}\| \leq \|\Delta_t(i)\|^2 \left\| \frac{d^2}{dw^2} V(w, \rho_t^m) \right\|_{op} \leq C_{\text{reg}} \|\Delta_t(i)\|^2$$

Recall that we have defined

$$\bar{D}_t(w) := \frac{d}{d\xi_t(w)} V(\xi_t(w), \bar{\rho}_t^m) = \nabla_{\xi_t(w)} F'(\xi_t(w)) - \mathbb{E}_{w' \sim \bar{\rho}_t^m} \nabla_{\xi_t(w)} K'(\xi_t(w), w').$$

Now

$$\begin{aligned}
\frac{d}{d\xi_t(w_i)} V(\xi_t(w_i), \rho_t^m) &= \frac{d}{d\xi_t(w_i)} F'(\xi_t(w_i)) - \mathbb{E}_j \frac{d}{d\xi_t(w_i)} K'(\xi_t(w_i), \hat{\xi}_t(w_j)) \\
&= \frac{d}{d\xi_t(w_i)} F'(\xi_t(w_i)) - \mathbb{E}_j \frac{d}{d\xi_t(w_i)} (K'(\xi_t(w_i), \xi_t(w_j)) + \mathbf{M}_j) \\
&= \bar{D}_t(i) - \mathbb{E}_j \mathbf{M}_j.
\end{aligned}$$

where by **S4**,

$$\|\mathbf{M}_j\|_{op} \leq \|\Delta_t(j)\| \sup_{w,w'} \left\| \frac{d}{dw} \frac{d}{dw'} K'(w, w') \right\|_{op} \leq C_{\text{reg}} \|\Delta_t(j)\|.$$

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 (B.2) and **S1** to swap $D_t(i)\Delta_t(i)$ for $D_t^\perp(i)\Delta_t(i)$ with an error term of magnitude $0.5 C_{\text{reg}} \|\Delta_t(i)\|^2$, we have that

$$V(\xi_t(w_i), \rho_t^m) - V(\hat{\xi}_t(w_i), \rho_t^m) = D_t^\perp(i)\Delta_t(i) + \mathbf{v}_3, \tag{B.4}$$

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 (B.1), (B.3), and (B.4), 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

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

□

C Proof of Concentration Lemmas

Lemma 17 (Uniformly Bounded Sampling Error)**.** *With probability* $1 - o(1)$ *over the initialization, for all* $t \le T$ *and* $i \in [m]$*, the following holds with* $\epsilon_m = \frac{d^{3/2}\log(Tm)}{\sqrt{m}}$.

$$\|V(\xi_t(w_i), \rho_t^{\text{MF}}) - V(\xi_t(w_i), \bar{\rho}_t^m)\| \le \epsilon_m.$$
$$\|D_t(i) - \bar{D}_t(i)\| \le \epsilon_m.$$

Proof. [Proof of Lemma 17] Fix $t \le T$ and $w \in \mathbb{S}^{d-1}$ By Equation (2.1), we have that

$$V(w, \rho_t^{\text{MF}}) - V(w, \bar{\rho}_t^m) := (I - ww^T)\Big(\mathbb{E}_{w' \sim \rho_t^{\text{MF}}} \nabla_w K(w, w') - \mathbb{E}_{w' \sim \bar{\rho}_t^m} \nabla_w K(w, w')\Big)$$

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

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

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

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

For any $w, u \in \mathbb{S}^{d-1}$, for any ρ, by Lemma 14, we have

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

Similarly, by Lemma 14, 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 C, for all $w \in \mathbb{S}^{d-1}$ and $t \le T$, we have

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

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 18 (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 18] 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 14, 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 17, 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 19. *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 $\epsilon_m^{19} = \frac{d \log(md)}{\sqrt{m}}$ we have*

$$\|\mathbb{E}_{w'\sim\rho_0} H_\infty^{\perp}(w, w') v(w') \mathbf{1}(\xi_t(w') \in S) - \mathbb{E}_{w'\sim\rho_0^m} H_\infty^{\perp}(w, w') v(w') \mathbf{1}(\xi_t(w') \in S)\| \le \|v\|_\infty \epsilon_m^{19}$$
$$\left|\mathbb{P}_{w'\sim\rho_0}[\xi_t(w') \in S] - \mathbb{P}_{w'\sim\rho_0^m}[\xi_t(w') \in S]\right| \le \epsilon_m^{19}.$$

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

$$\mathbb{P}\left[\left\|\mathbb{E}_{w\sim\rho_0^m} X(w') - \mathbb{E}_{w\sim\rho_0} X(w')\right\| \ge \frac{\epsilon_m^{19}}{2}\right] \le 2\exp\left(\frac{\Omega(m\epsilon_m^{19\,2})}{4C_{\text{reg}}^2 d}\right).$$

We need to build an ϵ-net of scale $\frac{\epsilon_m^{19}}{4C_{\text{reg}}}$ over $w \in \mathbb{S}^{d-1}$ since by Lemma 14, $X(w)$ is C_{reg}-Lipschitz in w. This net has size $\left(\left(\frac{O(C_{\text{reg}})}{\epsilon_m}\right)^d\right)$. Thus with $\epsilon_m^{19} = \frac{d \log(m)}{\sqrt{m}}$, we have that with high probability, for all $w \in \mathbb{S}^{d-1}$, the desired quantity is uniformly bounded. □

Lemma 20 (Averaging Lemma)**.** *Suppose $\mathcal{Q}$ is C_b-balanced, and the high probability event in Lemma 18 holds for $S = B_\tau$. If Assumption 2 holds, then for any $s \le t$,*

$$\mathbb{E}_i \|J_{t,s}(i) m_s(i)\| \mathbf{1}(\xi_t(w_i) \notin B_\tau) \le (1 + C_b)\big(\epsilon_m^{18} + J_{avg}(\tau)\big)\Phi_{\mathcal{Q}}(t).$$

In particular,

$$\mathbb{E}_i \|m_t(i)\| \mathbf{1}(\xi_t(w_i) \notin B_\tau) \le (1 + C_b)\big(\epsilon_m^{18} + J_{avg}(\tau)\big)\Phi_{\mathcal{Q}}(t).$$

Proof. Recall that

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

Thus

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

Now for any vector $v \in \mathbb{R}^d$, by Lemma 18 and Assumption 2, we have that

$$\mathbb{E}_i \|J_{t,s}(i) H_s^{\perp}(i,j) v\| \le \epsilon_m^{18}\|v\| + J_{\text{avg}}(\tau)\|v\|,$$

and so

$$\mathbb{E}_i \|J_{t,s}(i) m_s(i)\| \mathbf{1}(\xi_t(w_i) \notin B_\tau) \le \big(\epsilon_m^{18} + J_{\text{avg}}(\tau)\big)\mathbb{E}_i\|\Delta_t(i)\| \le \big(\epsilon_m^{18} + J_{\text{avg}}(\tau)\big)\Phi_{\mathcal{Q}}(t).$$

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

Lemma 21. *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*

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

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

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

We expand

$$V_{\hat{\mathcal{D}}}(w,\delta_{w'}) = (I - ww^T)\mathbb{E}_{x\sim\hat{\mathcal{D}}}(y - \sigma(w'^T x))\sigma'(w^T x)x;$$

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

$$\begin{aligned}\left|\mathbb{E}_{x\sim\hat{\mathcal{D}}}\sigma(w'^T x)\sigma'(w^T x)x^T v - \mathbb{E}_{x\sim\mathcal{D}}\sigma(w'^T x)\sigma'(w^T x)x^T v\right| &\le \epsilon_n \\ \left|\mathbb{E}_{x\sim\hat{\mathcal{D}}} y\sigma'(w^T x)x^T v - \mathbb{E}_{x\sim\mathcal{D}} y\sigma'(w^T x)x^T v\right| &\le \epsilon_n\end{aligned}$$

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 14 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

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

□

D Proof of Results Relating to Potential Function Analysis

D.1 Notation

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

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

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

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

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 / quadradic 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 $\mathbb{S}^{d-1}$.

D.2 Proof of Lemmas on the properties of the potential

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

Definition 22. *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 23. *The* automorphism group $\mathcal{G}$ *of a problem $(\rho^*, \mathcal{D}_x, \rho_0)$ 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 24. *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 25. *Conditional on A1 holding up to time t, we have*

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

Proof.

$$\begin{aligned} V(w, \rho_t^{\mathrm{MF}}) &= -(I - ww^T)\nabla_w F_{\mathcal{D}}(w) + (I - ww^T)\nabla_w \mathbb{E}_{w' \sim \rho_t^{\mathrm{MF}}} K_{\mathcal{D}}(w, w') \\ &= -(I - ww^T)\mathbb{E}_{x \sim \mathcal{D}_x} f^*(x)\sigma'(w^T x)x + (I - ww^T)\mathbb{E}_{w' \sim \rho_t^{\mathrm{MF}}}\mathbb{E}_{x \sim \mathcal{D}_x}\sigma(w'^T x)\sigma'(w^T x)x \end{aligned}$$

Now

$$\begin{aligned} &P_w^\perp \mathbb{E}_{w' \sim \rho_t^{\mathrm{MF}}}\mathbb{E}_{x \sim \mathcal{D}_x}\sigma(w'^T x)\sigma'(w^T x)x \\ &\qquad = P_w^\perp \mathbb{E}_{w' \sim \rho_t^{\mathrm{MF}}}\mathbb{E}_{x \sim \mathcal{D}_x}\sigma(g(w')^T g(x))\sigma'(g(w)^T g(x))x \\ &\qquad = P_w^\perp \mathbb{E}_{w' \sim \rho_t^{\mathrm{MF}}}\mathbb{E}_{x \sim \mathcal{D}_x}\sigma(g(w')^T x)\sigma'(g(w)^T x)g^{-1}(x) \\ &\qquad = P_w^\perp \mathbb{E}_{w' \sim \rho_t^{\mathrm{MF}}}\mathbb{E}_{x \sim \mathcal{D}_x}\sigma(w'^T x)\sigma'(g(w)^T x)g^{-1}(x) \\ &\qquad = (g^{-1}(x) - ww^T g^{-1}(x))\mathbb{E}_{w' \sim \rho_t^{\mathrm{MF}}}\mathbb{E}_{x \sim \mathcal{D}_x}\sigma(w'^T x)\sigma'(g(w)^T x) \\ &\qquad = (g^{-1}(x) - wg(w)^T x)\mathbb{E}_{w' \sim \rho_t^{\mathrm{MF}}}\mathbb{E}_{x \sim \mathcal{D}_x}\sigma(w'^T x)\sigma'(g(w)^T x) \\ &\qquad = g^{-1}(x - g(w)g(w)^T x)\mathbb{E}_{w' \sim \rho_t^{\mathrm{MF}}}\mathbb{E}_{x \sim \mathcal{D}_x}\sigma(w'^T x)\sigma'(g(w)^T x) \\ &\qquad = g^{-1}\Big(P_{g(w)}^\perp \nabla_{g(w)}\mathbb{E}_{w' \sim \rho_t^{\mathrm{MF}}} K_{\mathcal{D}}(g(w), w')\Big). \end{aligned}$$

Here (1) is because $z^T y = z^T U^T U y$ 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

$$\begin{aligned} P_w^\perp \mathbb{E}_{x \sim \mathcal{D}_x} f^*(x)\sigma'(w^T x)x &= \mathbb{E}_{x \sim \mathcal{D}_x} f^*(x)\sigma'(w^T x)P_w^\perp x \\ &= \mathbb{E}_{x \sim \mathcal{D}_x} f^*(x)\sigma'(w^T x)g^{-1}(P_{g(w)}^\perp)g(x) \\ &= \mathbb{E}_{x \sim \mathcal{D}_x} f^*(x)\sigma'(g(w)^T g(x))g^{-1}(P_{g(w)}^\perp g(x)) \\ &= \mathbb{E}_{x \sim \mathcal{D}_x}\mathbb{E}_{w^* \sim \rho^*}\sigma(w^{*T} x)\sigma'(g(w)^T g(x))g^{-1}(P_{g(w)}^\perp g(x)) \\ &= \mathbb{E}_{x \sim \mathcal{D}_x}\mathbb{E}_{w^* \sim \rho^*}\sigma(w^{*T} g^{-1}(x))\sigma'(g(w)^T x)g^{-1}(P_{g(w)}^\perp x) \\ &= \mathbb{E}_{x \sim \mathcal{D}_x}\mathbb{E}_{w^* \sim \rho^*}\sigma(g(w^*)^T x)\sigma'(g(w)^T x)g^{-1}(P_{g(w)}^\perp x) \\ &= \mathbb{E}_{x \sim \mathcal{D}_x} f^*(x)\sigma'(g(w)^T x)g^{-1}(P_{g(w)}^\perp x) \\ &= g^{-1}\Big(P_{g(w)}^\perp \mathcal{F}_{\mathcal{D}}(g(w))\Big). \end{aligned}$$

Putting these two computations together yields the desired conclusion,

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

□

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 $\text{supp}(\rho^*)$.

Finally for **A3**,

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

□

Lemma 26. *Suppose $(\rho^*, \mathcal{D}_x, \rho_0)$ is transitive. Then $(H_\infty^\perp, \rho_0)$ has consistent isometry with $B_\tau^t = \xi_t^{-1}(B_\tau)$ for any $t \leq T$, $\tau \geq 0$.*

Proof. We will use a series of small claims.

Claim 27. *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

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

□

Claim 28. *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 $v' : \mathrm{supp}(\rho_\infty^{\mathrm{MF}}) \to \mathbb{S}^{d-1}$.*

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 $v' : \mathrm{supp}(\rho_\infty^{\mathrm{MF}}) \to \mathbb{S}^{d-1}$ 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 27. □

D.2.2 Construction of the potential.

Remark 7. *We can verify that the action $\overline{H_\infty^\perp}$ (from Section 4.1) is well-defined in $\mathcal{Z}$ since $\|\overline{H_\infty^\perp} v\|_{\mathcal{Z}} \le \sup_{w,w'} \|H_\infty^\perp(w, w')\| \|v\|_{\mathcal{Z}}$). We verify that $\overline{H_\infty^\perp}$ is self-adjoint in $\mathcal{Z}$, ie $\langle v, \overline{H_\infty^\perp} v'\rangle_{\mathcal{Z}} = \langle \overline{H_\infty^\perp} v, v'\rangle_{\mathcal{Z}}$. We also verify that the span of $\overline{H_\infty^\perp}$ is finite-dimensional, thanks to the atomic nature of ρ^*. Indeed, for each $w^* \in$ supp(ρ^*) 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 $\overline{H_\infty^\perp} v = 0$.*

The following lemma implies Lemma 9. Recall that $C_{\rho^*} = \min\big(|\mathrm{supp}(\rho^*)|, \dim(\rho^*)^{2\,\mathrm{degree}(\sigma)+1}\big)$.

Lemma 29. *Suppose Assumption **I2** holds. Then for any μ, there exists an balanced spectral distribution $\mathcal{Q}$ of $(H_\infty^\perp, \mu)$ 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 $(H_\infty^\perp, \rho_0)$ which is $2C_{\rho^*}$-balanced.*

Proof. [Proof of Lemma 29] We will show that the linear operator induced by $(H_\infty^\perp, \mu)$ has an WED $\mathcal{Q}$ which is balanced for some constant depending on ρ^*.

Claim 30. *We can write*

$$H_\infty^\perp(w, w') = M_1(\xi^\infty(w), \xi^\infty(w'))UU^T + M_2(\xi^\infty(w), \xi^\infty(w')),$$

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

$$\begin{aligned}
M_1(w^*, w^{*\prime}) &:= \mathbb{E}_{x\sim\mathcal{D}_V} \sigma'(x^T w^*)\sigma'(x^T w^{*\prime}) \\
M_2(w^*, w^{*\prime}) &:= \mathbb{E}_{x\sim\mathcal{D}_V} \sigma'(x^T w^*)\sigma'(x^T w^{*\prime}) P_{w^*}^\perp x x^T P_{w^{*\prime}}^\perp.
\end{aligned}$$

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

Proof. Let V be the orthonormal basis spanning $\mathrm{supp}(\rho^*)$, and let U be any orthonormal basis of $\mathbb{R}^d \setminus \mathrm{span}(V)$. 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 $\mathrm{span}(\mathcal{D}_U) \in \mathrm{span}(U)$, $\mathrm{span}(\mathcal{D}_V) \in \mathrm{span}(V)$, $\mathbb{E}_{x\sim\mathcal{D}_U} xx^T = UU^T$, and $\mathbb{E}_{x\sim\mathcal{D}_U} x = 0$.

Recall that $H^{\perp}_{\infty}(w,w') = \mathbb{E}_{x\sim\mathcal{D}_x}\sigma'(x^T\xi^{\infty}(w))\sigma'(x^T\xi^{\infty}(w'))xx^T$. Observe that for $u, v \in \mathrm{Span}(U)$, we have

$$\begin{aligned} u^T H^{\perp}_{\infty}(w,w')v &= \mathbb{E}_{x\sim\mathcal{D}_x}\sigma'(x^T\xi^{\infty}(w))\sigma'(x^T\xi^{\infty}(w'))(u^Tx)(v^Tx) \\ &= \mathbb{E}_{x\sim\mathcal{D}_V}\sigma'(x^T\xi^{\infty}(w))\sigma'(x^T\xi^{\infty}(w'))\mathbb{E}_{x\sim\mathcal{D}_U} u^Txx^Tv \\ &= \mathbb{E}_{x\sim\mathcal{D}_V}\sigma'(x^T\xi^{\infty}(w))\sigma'(x^T\xi^{\infty}(w'))\mathbb{E}_{x\sim\mathcal{D}_U} u^Tv. \end{aligned}$$

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

$$u^T H^{\perp}_{\infty}(w,w')v = \mathbb{E}_{x_V\sim\mathcal{D}}\sigma'(x_V^T\xi^{\infty}(w))\sigma'(x_V^T\xi^{\infty}(w'))(v^Tx_V)\mathbb{E}_{x_U\sim\mathcal{D}_U}(u^Tx_U) = 0.$$

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

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

such that by the above computations,

$$H^{\perp}_{\infty}(w,w') = M_1(\xi^{\infty}(w),\xi^{\infty}(w'))UU^T + M_2(\xi^{\infty}(w),\xi^{\infty}(w')).$$

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

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

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

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

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 $H^{\perp}_{\infty}$ in Claim 30.

Claim 31. *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*

$$\begin{aligned} P_\lambda(w,w') &:= \sum_{f\in\mathcal{F}} f(\xi^\infty(w))f(\xi^\infty(w')^T\mathbf{1}(\lambda_f=\lambda) + UU^T\sum_{y\in\mathcal{Y}} y(\xi^\infty(w))y(\xi^\infty(w')\mathbf{1}(\lambda_y=\lambda) \\ &= \sum_{v\in\mathcal{B}_\lambda} v(w)v(w')^T, \end{aligned}$$

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

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

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 $\max_{w,f\in\mathcal{F},y\in\mathcal{Y}}(\|f(w)\|, |y(w)|) \le \frac{1}{\sqrt{p}}$, since the eigenfunctions are orthonormal. Fix $\lambda\in\Lambda$. We have

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

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 30, we have that

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

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 23), 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 23. 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 32 ($\mathcal{G}$-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

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

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

$$(I - ww^T)z = z - ww^Tz = z - wg(w)^Tg(z) = g^{-1}\big((I - g(w)g(w)^T)g(z)\big)$$

If the third line, we just used that for $z, z' \in \mathbb{R}^d$, we have $z^Tz' = g(z)^Tg(z')$. 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,

$$g^{-1}(U)g^{-1}(U)^T = UU^T. \tag{D.1}$$

Claim 33. *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',

$$\sum_{f\in\mathcal{F}}\lambda_f g(v^f)(w)g(v^f)(w')^T + \sum_{y\in\mathcal{Y}}\lambda_y g(v^{y,i})(w)g(v^{y,i})(w')^T = H^\perp_\infty(w, w'). \tag{D.2}$$

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

$$\begin{aligned} z^T \sum_{f\in\mathcal{F}} \lambda_f g(v^f)(w) g(v^f)(w')^T z' &= z^T \sum_{f\in\mathcal{F}} \lambda_f g^{-1}(f(\xi^\infty(g(w))))g^{-1}\big(f(\xi^\infty(g(w')))^T\big) z' \\ &= z^T \sum_{f\in\mathcal{F}} \lambda_f g^{-1}(f(g(\xi^\infty(w))))g^{-1}\big(f(g(\xi^\infty(w')))^T\big) z' \\ &= \sum_{f\in\mathcal{F}} \lambda_f g(z)^T f(g(\xi^\infty(w)))f(g(\xi^\infty(w')))^T g(z') \\ &= g(z)^T M_2(g(\xi^\infty(w)), g(\xi^\infty(w')))^T g(z') \\ &= z^T M_2(\xi^\infty(w), \xi^\infty(w'))^T z', \end{aligned} \tag{D.3}$$

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

$$z^T M_2(w^*, w^{*\prime})^T z' = g(z)^T M_2(g(w^*), g(w^{*\prime}))^T g(z')$$

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 (D.1),

$$\begin{aligned} \sum_{y\in\mathcal{Y}} \lambda_y g(v^{y,i})(w) g(v^{y,i})(w')^T &= M_1(\xi^\infty(w), \xi^\infty(w'))g^{-1}(U)g^{-1}(U)^T \\ &= M_1(\xi^\infty(w), \xi^\infty(w'))UU^T. \end{aligned} \tag{D.4}$$

Employing (D.4) and (D.3) yields (D.2) almost surely as desired.

Now, to prove the claim, we use (1) the fact from Claim 32 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 34. *For any $w \in \mathbb{S}^{d-1}$, we have*

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

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

$$z^T P_\lambda(w, w) z = \sum_{v\in\mathcal{B}_\lambda} z^T v(w)v(w)^T z = \frac{\langle \overline{P}_\lambda \pi_z, \pi_z\rangle}{(\mathbb{P}_{w'\sim\rho_0}[\xi^{\infty -1}(w^*)])^2} = |\mathrm{supp}(\rho^*)|^2 \langle \overline{P}_\lambda \pi_z, \pi_z\rangle.$$

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

Now recall that by Claim 33, 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

$$\begin{aligned} z^T P_\lambda(w,w) z &= |\mathrm{supp}(\rho^*)|^2 \langle \overline{P}_\lambda \pi_z, \pi_z \rangle \\ &= |\mathrm{supp}(\rho^*)|^2 z^T \mathbb{E}_{g\sim\mu_\mathcal{G}} \sum_{v \in g(\mathcal{B}_\lambda)} \mathbb{E}_{w',w''\sim\rho_0} v(w) v(w')^T \mathbf{1}(\xi^\infty(w') = \xi^\infty(w'') = w^*) z \\ &= z^T \mathbb{E}_{g\sim\mu_\mathcal{G}} \left(\sum_{f \in \mathcal{F} | \lambda_f = \lambda} g^{-1}(f(g(w^*))) g^{-1}(f(g(w^*)))^T z + \sum_{y \in \mathcal{F} | \lambda_y = \lambda} |y(g(w^*)|^2 g^{-1}(U) g^{-1}(U)^T \right) z. \end{aligned} \tag{D.5}$$

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

$$\begin{aligned} \mathbb{E}_{g\sim\mu_\mathcal{G}} g^{-1}(f(g(w^*)))(g^{-1}(f(g(w^*))))^T &\preceq \mathbb{E}_{g\sim\mu_\mathcal{G}} \|f(g(w^*))\|^2 I \\ &= \mathbb{E}_{w^{*\prime}\sim\rho^*} \|f(w^{*\prime})\|^2 I \\ &= I. \end{aligned} \tag{D.6}$$

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 (D.1),

$$\begin{aligned} \mathbb{E}_{g\sim\mu_\mathcal{G}} |(g(y))(w^*)|^2 g^{-1}(U) g^{-1}(U)^T &= \mathbb{E}_{g\sim\mu_\mathcal{G}} |y(g(w^*))|^2 U U^T \\ &= \mathbb{E}_{w^{*\prime}\sim\rho^*} |y(w^{*\prime})|^2 U U^T \\ &= U U^T. \end{aligned} \tag{D.7}$$

Combining Equations (D.6) and (D.7) with (D.5) 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. □

□

D.2.3 Properties of Potential

To prove our key lemmas 10, 11, 12, we will need several preliminary lemmas.

Lemma 35. *Suppose the high probability event in Lemma 19 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

$$\mathcal{E} \le \epsilon^{19}_m \mathbb{E}_i \|\Delta_t(i)\| + \mathbb{E}_i \|\Delta_t(i)\| \mathbf{1}(\xi_t(w_i) \notin B_\tau).$$

Proof. First observe that

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

and thus

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

Now by the conclusion of the concentration Lemma 19, we have

$$\langle v, \Delta_t \rangle_{H_\infty^\perp}^{B_\tau^t} = \mathbb{E}_i X(i) \Delta_t(i) \mathbf{1}(\xi_t(w_i) \in B_\tau) \pm \|v\|_\infty \epsilon_m^{19} \mathbb{E}_i \|\Delta_t(i)\|.$$

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

$$\langle v, \Delta_t \rangle_{H_\infty^\perp}^{B_\tau^t} = \lambda_v \langle v, \Delta_t \rangle^{B_\tau^t} \mathbb{P}_{\rho_t^{\mathrm{MF}}}[B_\tau] \pm \epsilon_m^{19} \|v\|_\infty \mathbb{E}_i \|\Delta_t(i)\|.$$

Now

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

Plugging this back in yields the lemma. □

Now we prove Lemma 10, which we restate here.

Lemma 36 (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 19 holds for $S = B_\tau$, we have*

$$\langle \nabla \Phi_{\mathcal{Q}}(t), -H_t^\perp \Delta_t \rangle \le (1 + C_b) \mathbb{E}_i \|\mathbb{E}_j H_t^\perp(i,j)\| \mathbf{1}(\xi_t(w_i) \notin B_\tau) + \mathcal{E}_{10},$$

where $\mathcal{E}_{10} = O_{C_{reg}, C_b}(\mathbb{E}_i \|\Delta_t(i)\| \mathbf{1}(\xi_t(w_i) \notin B_\tau) + (\tau + C_b \epsilon_m^{19}) \Omega(t))$.

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

$$\langle \nabla \Phi_{\mathcal{Q}}(t), \Delta_t \rangle_{H_t^\perp} = \langle \nabla \Phi_{\mathcal{Q}}(t), \Delta_t \rangle_{H_t^\perp}^{B_\tau^t, B_\tau^t} + \langle \nabla \Phi_{\mathcal{Q}}(t), \Delta_t \rangle_{H_t^\perp}^{B_\tau^t, \bar{B}_\tau^t} + \langle \nabla \Phi_{\mathcal{Q}}(t), \Delta_t \rangle_{H_t^\perp}^{\bar{B}_\tau^t, \mathbb{S}^{d-1}}. \tag{D.8}$$

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 37.

$$\langle \nabla \Phi_{\mathcal{Q}}(t), \Delta_t \rangle_{H_t^\perp}^{B_\tau^t} \ge -(C_{reg} + 1) \mathbb{E}_i \|\Delta_t(i)\| \mathbf{1}(\xi_t(w_i) \notin B_\tau) - C_b \epsilon_m^{19} \Omega(t) + |\langle \nabla \Phi(t), G \rangle|,$$

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

Proof. We have

$$\langle \nabla \Phi_{\mathcal{Q}}(t), \Delta_t \rangle_{H_t^\perp}^{B_\tau^t} = \langle \nabla \Phi_{\mathcal{Q}}(t), \Delta_t \rangle_{H_\infty^\perp}^{B_\tau^t} + \langle \nabla \Phi_{\mathcal{Q}}(t), G \rangle, \tag{D.9}$$

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

$$\begin{aligned}
\langle \nabla \Psi_{\mathcal{Q}}(t), \Delta_t \rangle_{H_\infty^\perp}^{B_\tau^t} &= \sum_{\lambda \in \Lambda} \eta_\lambda \frac{\sum_{v \in \mathcal{B}_\lambda} \phi_v(t) \langle \nabla \phi_v(t), \Delta_t \rangle_{H_\infty^\perp}^{B_\tau^t}}{\sqrt{\sum_{v \in \mathcal{B}_\lambda} (\phi_v(t))^2}} \\
&= \sum_{\lambda \in \Lambda} \eta_\lambda \frac{\sum_{v \in \mathcal{B}_\lambda} \phi_v(t) \Big(\lambda \phi_v(t) \mathbb{P}_{\rho_t^{\text{MF}}}[B_\tau] + \mathcal{E}_{\mathbf{v}} \Big)}{\sqrt{\sum_{v \in \mathcal{B}_\lambda} (\phi_v(t))^2}} \\
&= \mathbb{P}_{\rho_t^{\text{MF}}}[B_\tau] \sum_{\lambda \in \Lambda} \eta_\lambda \left(\lambda \sqrt{\sum_{v \in \mathcal{B}_\lambda} (\phi_v(t))^2} + \frac{\sum_{v \in \mathcal{B}_\lambda} \phi_v(t) \mathcal{E}_{\mathbf{v}}}{\sqrt{\sum_{v \in \mathcal{B}_\lambda} (\phi_v(t))^2}} \right) \\
&\geq \mathbb{P}_{\rho_t^{\text{MF}}}[B_\tau] \sum_{\lambda \in \Lambda} \eta_\lambda \lambda \sqrt{\sum_{v \in \mathcal{B}_\lambda} (\phi_v(t))^2} - \mathcal{E} \ ,
\end{aligned} \tag{D.10}$$

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 35.

Next consider the $\langle \nabla \Omega(t), \Delta_t \rangle_{H_\infty^\perp}^{B_\tau^t}$ part. Recall from the definition of WED 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

$$\begin{aligned}
\left| \langle \nabla \Omega(t), \Delta_t \rangle_{H_\infty^\perp}^{B_\tau^t, \mathbb{S}^{d-1}} \right| &= \left| \mathbb{E}_{i,j} \sum_{v \in \mathcal{Q}} \lambda_v u_i^T v(w_i) v(w_j)^T \Delta_t(j) \mathbf{1}(w_i \in B_\tau^t) \right| \\
&= \left| \mathbb{E}_i \sum_{v \in \mathcal{Q}} \lambda_v u_i^T v(w_i) \mathbf{1}(i \in B_\tau^t) \big(\mathbb{E}_j v(w_j)^T \Delta_t(j) \big) \right| \\
&\leq \sum_{v \in \mathcal{Q}} \lambda_v \phi_v(t) \mathbb{E}_i |u_i^T v(w_i)| \mathbf{1}(i \in B_\tau^t).
\end{aligned} \tag{D.11}$$

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

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

Here the final inequality follows from the definition of a WED , which states that for any $w \in \mathbb{S}^{d-1}$, $\sum_{v\in\mathcal{B}_\lambda} v(w)v(w)^T \preceq \eta_\lambda^2 I$. Thus plugging this back into to Equation (D.11), we have

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

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

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

and thus

$$\left|\langle \nabla\Omega(t), \Delta_t\rangle_{H_\infty^\perp}^{B_\tau^t}\right| \le \mathbb{P}_i[B_\tau^t]\sum_{\lambda\in\Lambda}\eta_\lambda\lambda\sqrt{\sum_{v\in\mathcal{B}_\lambda}(\phi_v(t))^2} + C_{\text{reg}}\mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i)\notin B_\tau). \tag{D.12}$$

Now recall that $\Phi_{\mathcal{Q}}(t) := \Omega(t) + \Psi_{\mathcal{Q}}(t)$. Thus combining Equations (D.12) and (D.10), and Equation (D.9), and plugging in the bound on $\mathcal{E}$ from Lemma 35, we have

$$\langle \nabla\Phi_{\mathcal{Q}}(t), \Delta_t\rangle_{H_t^\perp}^{B_\tau^t} \ge -(C_{\text{reg}}+1)\mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i)\notin B_\tau) - C_{\text{b}}\epsilon_m^{19}\Omega(t) + |\langle \nabla\Phi_{\mathcal{Q}}(t), G\rangle|,$$

where $\mathbb{E}_i\|G_i\| \le C_{\text{reg}}\tau\Omega(t)$. Here we have also used the fact that for all v in the WED $\mathcal{Q}$, we have that $\|v\|_\infty \le \sqrt{C_{\text{b}}} \le C_{\text{b}}$ (this is evident from the definition of WED). 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 (D.8). We have

$$\left|\langle \nabla\Phi_{\mathcal{Q}}(t), \Delta_t\rangle_{H_t^\perp}^{B_\tau^t,\bar{B}_\tau^t}\right| = \langle \nabla\Phi_{\mathcal{Q}}(t), H\rangle, \tag{D.13}$$

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

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 (D.8), we have just write

$$\langle \nabla\Phi_{\mathcal{Q}}(t), \Delta_t\rangle_{H_t^\perp}^{\bar{B}_\tau^t,\mathbb{S}^{d-1}} = \langle \nabla\Phi_{\mathcal{Q}}(t), m_t\rangle^{\bar{B}_\tau^t}, \tag{D.14}$$

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

Combining Equations (D.13), (D.14) and Claim 37 into Equation D.8, we obtain that

$$\langle \nabla\Phi_{\mathcal{Q}}(t), \Delta_t\rangle_{H_t^\perp} \ge -(C_{\text{reg}}+1)\mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i)\notin B_\tau) - C_{\text{b}}\epsilon_m^{19}\Omega(t) + |\langle \nabla\Phi_{\mathcal{Q}}(t), G+H+m_t\rangle|,$$

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

Now we use Lemma 12 to bound

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

Plugging this back in to the equation above yields

$$\langle \nabla\Phi_{\mathcal{Q}}(t), \Delta_t\rangle_{H_t^\perp} \ge -(1+C_{\text{b}})\mathbb{E}_i\|m_t(i)\| - \mathcal{E}_{10},$$

where

$$\begin{aligned}\mathcal{E}_{10} &= (C_{\text{reg}}(2+C_{\text{b}})+1)\mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i)\notin B_\tau) + (C_{\text{b}}\epsilon_m^{19} + (1+C_{\text{b}})C_{\text{reg}}\tau)\Omega(t)\\ &= O_{C_{\text{reg}},C_{\text{b}}}(\mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i)\notin B_\tau) + (\tau + C_{\text{b}}\epsilon_m^{19})\Omega(t)).\end{aligned}$$

This proves the lemma. □

Now we prove Lemma 11, which we restate here.

Lemma 38 (Descent with Respect to Local Term)**.** *Suppose Assumption 4 holds with (C_{LSC},τ). Let $\mathcal{Q}$ be a C_b-balanced spectral distribution. Then with $C_{11} = O_{C_{reg},C_b}(1)$, we have*

$$\langle\nabla\Phi_{\mathcal{Q}}(t), D_t^\perp\odot\Delta_t\rangle \le -\left(\frac{c\sqrt{L_{\mathcal{D}}(\rho_t^{\text{MF}})}}{2} - C_{11}\tau\right)\Phi_{\mathcal{Q}}(t) + C_{11}\mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i)\notin B_\tau) + C_b\mathbb{E}_i\|\Delta_t(i)\|^2.$$

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

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

and that

$$\begin{aligned}\langle\nabla\Psi_{\mathcal{Q}}(t), D_t^\perp\odot\Delta_t\rangle \le -(C_{\text{LSC}}\delta)\Psi_{\mathcal{Q}}(t) + \frac{C_{\text{LSC}}\delta + 2C_{\text{b}}C_{\text{reg}}\tau}{2}\Omega(t)\\ 2C_{\text{b}}C_{\text{reg}}\mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i)\notin B_\tau) + C_{\text{b}}\mathbb{E}_i\|\Delta_t(i)\|^2.\end{aligned} \tag{D.15}$$

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

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

as desired.

For the second statement, write

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

where

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

By the structured condition in Assumption 4, we can write such a decomposition where $c_1, c_2 \ge C_{\text{LSC}}\delta$, and for any i such that $\xi_t(w_i)\in B_\tau$, we have $\|D_t^{\text{bad}}(i)\| \le \frac{C_{\text{LSC}}\delta}{2\sqrt{C_{\text{b}}}} + C_{\text{reg}}\tau$. 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 39.

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

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

Proof.

Now recall that in the construction for $\mathcal{Q}$ given in Lemma 29, for any $v \in \mathrm{supp}(\mathcal{Q})$, it holds that either $v(w) \in \mathrm{span}(U)$ for all $w \in \mathbb{S}^{d-1}$, or $v(w) \in \mathrm{span}(V)$ for all $w \in \mathbb{S}^{d-1}$. We consider the two cases separately. First suppose $v(w) \in \mathrm{span}(U)$ for all $w \in \mathbb{S}^{d-1}$. Fix w_i with $\xi_t(w_i) \in B_\tau$. For any w, we have

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

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

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

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 (B.2) in the proof of Lemma **??**), we have that

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

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 12 (see Equation (D.16)) we have

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

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 (D.15), which proves the lemma.

□

We now prove Lemma 12, which we restate here.

Lemma 40 (L1 Perturbation Lemma)**.** *Let $\mathcal{Q}$ be a C_b-balanced spectral distribution. Let $G : [m] \to \mathbb{R}^d$. Then $|\langle \nabla \Phi_{\mathcal{Q}}(t), G\rangle| \le (1 + C_b)\mathbb{E}_i\|G(i)\|$.*

Proof. [Proof of Lemma 12] First observe that $\langle \nabla \Phi_{\mathcal{Q}}(t), G\rangle \le \mathbb{E}_i\|G(i)\|$, since $\nabla_i \Omega(t) = \frac{\Delta_t(i)}{\|\Delta_t(i)\|}$, which has norm 1. Now for any $v \in \mathrm{supp}(\mathcal{Q})$, we have

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

and so

$$\begin{aligned}
|\langle \nabla \Psi_{\mathcal{Q}}(t), G\rangle| &\le \sum_{\lambda\in\Lambda} \eta_\lambda \frac{\sum_{v\in\mathcal{B}_\lambda} \phi_v(t)|\langle \nabla\phi_v(t), G\rangle}{\sqrt{\sum_{v\in\mathcal{B}_\lambda}(\phi_v(t))^2}} \qquad \text{(D.16)}\\
&\le \mathbb{E}_i\left[\sum_{\lambda\in\Lambda} \eta_\lambda \frac{\sum_{v\in\mathcal{B}_\lambda} \phi_v(t)|G(i)^T v(w_i)|}{\sqrt{\sum_{v\in\mathcal{B}_\lambda}(\phi_v(t))^2}}\right]\\
&\le \mathbb{E}_i\left[\sum_{\lambda\in\Lambda} \eta_\lambda \frac{\sqrt{\sum_{v\in\mathcal{B}_\lambda}(\phi_v(t))^2}\sqrt{\sum_{v\in\mathcal{B}_\lambda}|G(i)^T v(w_i)|^2}}{\sqrt{\sum_{v\mathcal{B}_\lambda}(\phi_v(t))^2}}\right]\\
&= \mathbb{E}_i\left[\sum_{\lambda\in\Lambda} \eta_\lambda \sqrt{G(i)^T\left(\sum_{v\in\mathcal{B}_\lambda} v(w_i)v(w_i)^T\right)G(i)}\right]\\
&\le \mathbb{E}_i \sum_{\lambda\in\Lambda} \eta_\lambda^2\|G(i)\| = C_{\mathrm{b}}\mathbb{E}_i\|G(i)\|.
\end{aligned}$$

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

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

as desired.

□

D.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 41. *Fix some δ. With high probability as $d, m, n \to \infty$, the events in all concentration lemmas (Lemma 17,Lemma 21, Lemma 18 and Lemma 19) hold, where we apply Lemma 18 and Lemma 19 for $S = B_\tau$ for all*

$$\tau \in \left\{\frac{C_{\mathrm{LSC}} \cdot rd(e)}{8(C_{10} + C_{11})}\right\}_{e\in[\delta,1]},$$

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 19 for all eigenfunctions v in the WED $\mathcal{Q}$.

Proof. The set $\left\{\frac{c\cdot \mathrm{rd}(e)}{8(C_{10}+C_{11})}\right\}_{e\in[\delta,1]}$ has size at most $O_{C_{10},C_{11}}(\log_2(1/\delta))$, so we can take a union bound over the result in Lemma 18 for all B_τ. Similarly, since there are $O(dC_{\rho^*})$ eigenfunctions in $\mathcal{Q}$ (see the proof of Lemma 29), we take a union bound of Lemma 19 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 1,2,4,5. 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 WED , where by Lemma 9, we have that $C_{\mathrm{b}}=C_{\rho^*}$.

Theorem 3 (Main Potential Dynamics Theorem)**.** *Let $\delta:=\sqrt{L_{\mathcal{D}}(\rho_t^{\mathrm{MF}})}$, and let C be a constant depending on $C_{\mathrm{LSC}},\tau,\delta$ and C_b. Then with high probability over the draw ρ_0^m, for all $t\le T$, Condition on the event that the high probability event in Lemma 41 holds for δ. Let $\epsilon_{n,m}:=\epsilon_n+\epsilon_m^{17}+\epsilon_m^{18}+\epsilon_m^{19}$ from the concentration lemmas. Suppose n and m are large enough such that $J_{max}^4T^3(\epsilon_n+\epsilon_m)\le 1/C$. Suppose that*

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

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

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

Corollary 42 (Solution to Potential Dynamics)**.** *Suppose that for some $\tau=\Omega_{C_{reg},C_b}(\delta)$,*

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

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

$$\Phi_{\mathcal{Q}}(T)\le\exp(CTJ_{avg}(\tau)/(C_{\mathrm{LSC}}\delta))CJ_{max}T\epsilon_{n,m}.$$

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

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

Note that is implies the assumption in Equation 3. Clearly this holds for $t=0$. Since $\Phi_{\mathcal{Q}}(s)$ is continuous, if Equation 3 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 3 holds for some s, then for some ι small enough, it holds at $s+\iota$. To show this, first we use Lemma 44 (which bounds the solution of the ODE given in Theorem 3), to show that for all $s'\le s$,

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

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,

$$\begin{aligned}\int_{s'=0}^t(\Phi_{\mathcal{Q}}(s'))^2ds'&\le(CtJ_{\max}\epsilon_{n,m})^2\int_{s'=0}^s\exp(2CsJ_{\mathrm{avg}}(\tau)/(C_{\mathrm{LSC}}\delta))ds'+\int_{s'=s}^t(\Phi_{\mathcal{Q}}(s)+\epsilon_{n,m})^2ds'\\&\le 2(CtJ_{\max}\epsilon_{n,m})^2t\exp(2CJ_{\mathrm{avg}}(\tau)t/(C_{\mathrm{LSC}}\delta))\end{aligned}$$

Now using the assumption in the corollary that

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

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

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 44 to the result of Theorem 3 at time T. □

Proof. [Proof of Theorem 3] 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

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

Now we have

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

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

$$\tau := \frac{C_{\text{LSC}}\cdot \text{rd}(\delta)}{8(C_{10}+C_{11})},$$

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 10, we have

$$-\langle \nabla\Phi_{\mathcal{Q}}(t), H_t^{\perp}\Delta_t\rangle = -\langle \nabla\Phi_{\mathcal{Q}}(t), \Delta_t\rangle_{H_t^{\perp}} \le (1+C_{\text{b}})\mathbb{E}_i\|m_t(i)\|\mathbf{1}(\xi_t(w_i)\notin B_\tau) + \mathcal{E}_{10},$$

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

$$\mathcal{E}_{10} = C_{10}(\mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i)\notin B_\tau) + (\tau + C_{\text{b}}\epsilon_m^{19})\Omega(t)).$$

Next by Lemma 11, we have

$$\langle \nabla\Phi_{\mathcal{Q}}(t), D_t^{\perp}\odot\Delta_t\rangle \le -\left(\frac{C_{\text{LSC}}\delta}{2} - \tau C_{11}\right)\Phi_{\mathcal{Q}}(t) + C_{11}\mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i)\notin B_\tau) + C_{\text{b}}\mathbb{E}_i\|\Delta_t(i)\|^2.$$

Putting these together, and employing Lemma 12, yields

$$\begin{aligned}\frac{d}{dt}\Phi_{\mathcal{Q}}(t) \le &\left(-\frac{C_{\text{LSC}}\delta}{4}\right)\Phi_{\mathcal{Q}}(t) \\ &+ (C_{10}+C_{11})\mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i)\notin B_\tau) \\ &+ (1+C_{\text{b}})\mathbb{E}_i\|m_t(i)\|\mathbf{1}(\xi_t(w_i)\notin B_\tau) \\ &+ (1+2C_{\text{b}})\mathbb{E}_i\|\epsilon_{t,i}\|,\end{aligned} \tag{D.17}$$

where here we used that τ was chosen such that $(C_{11}+C_{10})(\tau + C_{\text{b}}\epsilon_m^{19}) \le \frac{C_{\text{LSC}}\delta}{8}$, and trivially, $\Omega(t)\le\Phi_{\mathcal{Q}}(t)$. 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 20, we have

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

Now let use consider the term $\mathbb{E}_i\|\Delta_t(i)\|\mathbf{1}(\xi_t(w_i) \notin B_\tau)$. Recall from Equation (2.2) 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 20, we have

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

Plugging this back into Equation (D.17) yields

$$\begin{aligned}\frac{d}{dt}\Phi_{\mathcal{Q}}(t) &\le -\frac{C_{\text{LSC}}\delta}{5}\Phi_{\mathcal{Q}}(t) + (C_{10} + 4\sqrt{C_{\mathrm{b}}}C_{\text{reg}})(1 + C_{\mathrm{b}})(\epsilon_{n,m} + J_{\text{avg}}(\tau))\int_{s=0}^{t}\Phi_{\mathcal{Q}}(s)ds \\ &\quad + (1 + C_{\mathrm{b}})\mathbb{E}_i\|\boldsymbol{\epsilon}_{t,i}\| + (1 + C_{\mathrm{b}})(C_{10} + 4\sqrt{C_{\mathrm{b}}}C_{\text{reg}})\int_{s=0}^{t}\mathbb{E}_i\|J_{t,s}(i)\boldsymbol{\epsilon_{s,i}}\|ds \\ &\le -\frac{C_{\text{LSC}}\delta}{5}\Phi_{\mathcal{Q}}(t) + CJ_{\text{avg}}(\tau)\int_{s=0}^{t}\Phi_{\mathcal{Q}}(s)ds \\ &\quad + (1 + C_{\mathrm{b}})\mathbb{E}_i\|\boldsymbol{\epsilon}_{t,i}\| + C\int_{s=0}^{t}\mathbb{E}_i\|J_{t,s}(i)\boldsymbol{\epsilon_{s,i}}\|ds,\end{aligned}$$

where $C = O_{C_{\text{reg}},C_{\mathrm{b}}}(1)$. Let us simplify the error terms. Appealing to Lemma 43, we have for all i, $\|\Delta_t(i)\|^2 \le 4\epsilon_{n,m}$ and $E_{t,i} := \int_{s=0}^{t}\|J_{t,s}(i)\boldsymbol{\epsilon_{s,i}}\|ds \le 8J_{\max}t\epsilon_{n,m}$.

Thus

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

and

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

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

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

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

Lemma 43 (Inductive Squared Error Bound.)**.** *Suppose Assumption 2 hold with value J_{max}. Suppose for all $t' \le t$, we have*

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

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

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

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

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

Since

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

by Equation 2.2, we have

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

Here in the second last inequality, we used the fact that $\|m_s(i)\| \le \Phi_{\mathcal{Q}}(s)$ for any i, and in the last line, we used assumption of the lemma. Note that this same calculation holds for all $s \le t$, so we have

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

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

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

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

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

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 44 (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 $\frac{d}{dr} Z_r \leq -aZ_r + b\iota\big(\frac{\epsilon}{b}\big) - \epsilon \leq -aZ_r$ 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

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

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

E Applications to Learning a Single Index Model

E.1 Setting

We will study the setting of learning a well-specified even single index function $f^*(x) = \sigma(x^T w^*)$, 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 $\frac{1}{C_{\mathrm{SIM}}} \leq c_{k^*} \leq C_{\mathrm{SIM}} \max_k c_k$.
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 2 in the main body.

Theorem 4. *Fix any δ small enough. Consider the setting $(f^*, \rho_0, \mathcal{D}_x)$ described above for d large enough (in terms of δ). Then for some $t \leq O_{K, C_{reg}}(\sqrt{d}^{k^*-2} \delta^{-k^*})$, we have*

$$\mathbb{E}_{x \sim \mathcal{D}_x}(f_{\rho_t^{\mathrm{MF}}}(x) - f^*(x))^2 \leq \delta^2.$$

Further, suppose $n, m \geq d^{13k^*}$. Let $\hat{\mathcal{D}}$ be the empirical distribution of n samples drawn from $\mathcal{D}$. With high probability over $\hat{\mathcal{D}}$ and the initialization ρ_0^m, we have

$$\mathbb{E}_{x \sim \mathcal{D}_x}(f_{\rho_t^m}(x) - f^*(x))^2 \leq 2\delta^2.$$

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

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

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

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 $\mathbb{E}_x \mathrm{He}_k(w^T x) \mathrm{He}_j(v^T x) = k! \delta_{jk} (w^T v)^k$, we have

$$\mathbb{E}_x \sigma(w^T x) \sigma(v^T x) = q_\sigma(w^T v).$$
$$\mathbb{E}_x \sigma'(w^T x) \sigma'(v^T x) = q_{\sigma'}(w^T v).$$

E.2 Bounds on the Velocity and its Derivative

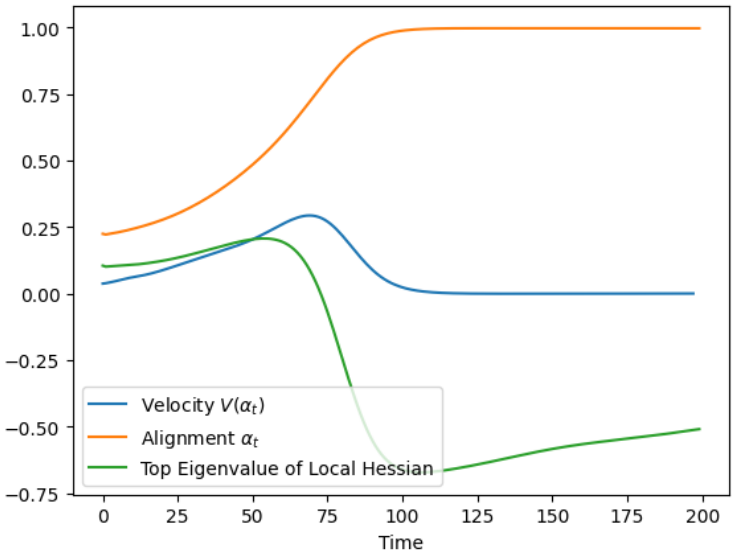

Figure 3: Self-Concordance Property: The top eigenvalue of the Local Hessian is Bounded as by $\frac{k-1}{\alpha_t} V(\alpha_t)$

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

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

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

Lemma 45 (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* $\iota \le \min\Big(\Theta_K(1), \delta^{6K^2}\Big)$. *Then*

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

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* $\alpha \ge \frac{\delta^{3K}}{\sqrt{d}}$, *for* d *large enough (in terms of* δ, K*), we have that*

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

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 $w^T w^* > 0$. 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

$$\begin{aligned} V(w, \rho_t^{\mathrm{MF}})^T w^* &= \mathbb{E}_x(f^*(x) - f_{\rho_t^{\mathrm{MF}}}(x))\sigma'(w^T x)x^T P_w^\perp w^* \\ &= \mathbb{E}_x \sigma(w^{*T} x)\sigma'(w^T x)x^T P_w^\perp w^* - \mathbb{E}_{w'\sim\rho_t^{\mathrm{MF}}}\mathbb{E}_x\sigma(w'^T x)\sigma'(w^T x)x^T P_w^\perp w^* \\ &= q_{\sigma'}(w^T w^*)w^{*T} P_w^\perp w^* - \mathbb{E}_{w'\sim\rho_t^{\mathrm{MF}}} q_{\sigma'}(w^T w')(w')^T P_w^\perp w^* \\ &= q_{\sigma'}(\alpha)(1-\alpha^2) - \mathbb{E}_{w'\sim\rho_t^{\mathrm{MF}}} q_\sigma(w^T w')(w')^T P_w^\perp w^* \\ &= q_{\sigma'}(\alpha)(1-\alpha^2) - \mathbb{E}_{\alpha'\sim\rho_t^{\mathrm{MF}}}\mathbb{E}_{y'\sim\sqrt{1-\alpha'^2}\mathbb{S}^{d-2}} q_\sigma(\alpha\alpha' + y'^T w)\big(\alpha'(1-\alpha^2) - \alpha y'^T w\big). \end{aligned} \tag{E.1}$$

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

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

where

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

Note here that we have used the fact that k is even and $\mathbb{E}_{y'}(y'^T w)^j = O_j(\big((1-\alpha'^2)(1-\alpha^2)d^{-1}\big)^{j/2})$, 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 $\mathbb{E}_{\alpha'\sim\rho_t^{\mathrm{MF}}}(\alpha')^2(1-\alpha'^2) \le 2\iota$, so for all k,

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

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

$$v(\alpha,t) = \sum_{k=k^*-1}^{K-1} q_k \alpha^k (1-\alpha^2)(1-r_{t,k+1}) + (1-\alpha^2)\mathcal{R}_\alpha \tag{E.2}$$
$$\ge q_{\sigma'}(\alpha)(1-\alpha^2)(1-r_t) - (1-\alpha^2)\mathcal{R}_\alpha,$$

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

$$v(\alpha,t) \ge q_{\sigma'}(\alpha)(1-\alpha^2)(1-r_t)$$
$$- O_K\Big(\iota(1-\alpha^2)(\alpha^{k^*-1}\delta^{-3K(k^*-2)} + \alpha^{k^*-1}\delta^{-6K}) + (1-\alpha^2)\alpha^{k^*-1}\delta^{-3K(k^*-2)}/d\Big),$$

Since by Lemma 48, we have $(1-r_t) = \Omega(\delta)$, it follows that

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

□

In the following lemma, we analyze $\frac{d}{d\alpha}v(\alpha,t)$. As will be shown in Section E.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 4).

Lemma 46. *Let* $\delta := \sqrt{\mathbb{E}_x(f_{\rho_t^{\mathrm{MF}}}(x) - f^*(x))^2}$. *Suppose* ($\star$) *holds at time* t *for* $\iota \le \min\Big(\Theta_K(1), \delta^{6K^2}\Big)$. *Then*

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

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

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 (E.1) arising from that order term:

$$q_k k\alpha^{k-1}\Big(1 - \frac{k+2}{k}\alpha^2\Big)$$
$$- q_k \sum_{j=0}^{k}\binom{k}{j} j(\alpha\alpha')^{j-1}\mathbb{E}_{y'\sim\mathbb{S}^{d-2}_{\sqrt{1-\alpha'^2}}}(y'^T w)^{k-j}\Big(\alpha'\Big(1 - \frac{j+2}{j}\alpha^2\Big) + \frac{j+1}{j}\alpha y'^T w\Big)$$
$$= k q_k \alpha^{k-1}\Big(1 - \frac{k+2}{k}\alpha^2\Big)(1 - r_{t,k+1}) + \mathcal{E}_{\alpha,k},$$

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 45. Summing over all odd $k^* - 1 \le k \le K-1$ yields

$$\frac{d}{d\alpha}v(\alpha,t) = \sum_{k=k^*-1}^{K} q_k k\alpha^{k-1}\Big(1 - \frac{k+2}{k}\alpha^2\Big)(1 - r_{t,k+1}) + \mathcal{E}_{\alpha,k} \tag{E.3}$$
$$= (k^*-1)q_{k^*-1}(1-r_t)\alpha^{k^*-2} + \Theta_K\Big(\alpha^{k^*} + \iota(\sqrt{d}^{-(k^*-2)} + \alpha^{k^*-4}\sqrt{d}^{-2}) + \sqrt{d}^{-(k^*-2)}\Big)$$

Combining Lemma 45 with the previous equation, we obtain

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

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

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

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

We will compare the terms with coefficient q_k in the first line of Equation (E.3) and the first line of Equation (E.2). Let

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

such that Equation (E.2) 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 45. Thus the first line of Equation (E.3) gives

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

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. In the last inequality, we have used the bounds on $\mathcal{R}_\alpha$ and $\mathcal{E}_{\alpha,k}$, along with the fact from Lemma 45 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 47. *Suppose that for all $s \leq t$, we have $\sqrt{\mathbb{E}_x(f_{\rho_s^{\mathrm{MF}}}(x) - f^*(x))^2} \geq \delta$. Suppose $\iota \leq \min\Big(\Theta_K(1), \delta^{6K^2}\Big)$, and $t \leq \frac{\sqrt{d}^{k^*-2}}{\iota}$. Finally suppose ($\star$) holds for all $s \leq t$. Then for and $\tau \leq 1/2$ and any w for which $\alpha_t(w) \leq 1-\tau$, 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

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

From Lemma 46, 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

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

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 45, we have $\mathcal{E}_\alpha \leq O_K\Big(\sqrt{\iota}\frac{v(\alpha,t)}{\alpha_t^3}\sqrt{d}^{-2} + v(\alpha,t)\alpha/\delta\Big)$.

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 45 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 $O_K\Big(\frac{\log(1/\tau)}{\delta^2}\Big)$, 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 48. *For d large enough in terms of $\delta = \sqrt{\mathbb{E}_x(f_{\rho_t^{\text{MF}}}(x) - f^*(x))^2}$, we have*

$$1 - \mathbb{E}_{\alpha' \sim \rho_t^{\text{MF}}}(\alpha')^{k^*} \geq \Omega_{K, C_{reg}}(\delta).$$

Proof. Observe that

$$\mathbb{E}_x(f^*(x))^2 = \mathbb{E}_x \sigma(w^{*T} x)\sigma(w^{*T} x) = q_\sigma(1).$$

$$\mathbb{E}_x f_{\rho_t^{\text{MF}}}(x) f^*(x) = \mathbb{E}_x \mathbb{E}_{w' \sim \rho_t^{\text{MF}}} \sigma(w'^T x)\sigma(w^{*T} x) = \mathbb{E}_{\alpha' \sim \rho_t^{\text{MF}}} q_\sigma(\alpha').$$

Further

$$\mathbb{E}_x(f_{\rho_t^{\text{MF}}}(x))^2 = \mathbb{E}_x \mathbb{E}_{w,w' \sim \rho_t^{\text{MF}}} \sigma(w^T x)\sigma(w'^T x) = \mathbb{E}_{w,w' \sim \rho_t^{\text{MF}}} q_\sigma(w^T w').$$

Now for even k, we have

$$\mathbb{E}_{w,w' \sim \rho_t^{\text{MF}}}(w^T w')^k = \mathbb{E}_{\alpha,\alpha' \sim \rho_t^{\text{MF}}} \mathbb{E}_\zeta (\alpha\alpha' + \sqrt{(1-\alpha^2)(1-\alpha')^2}\zeta)^k,$$

where ζ is $\frac{1}{\sqrt{d}}$-subgaussian. Thus by Minowski's inequality, we have

$$\begin{aligned}
\mathbb{E}_{w,w' \sim \rho_t^{\text{MF}}}(w^T w')^k &\leq \left(\left(\mathbb{E}_{\alpha,\alpha' \sim \rho_t^{\text{MF}}}(\alpha\alpha')^k\right)^{1/k} + \frac{O_K(1)}{\sqrt{d}}\right)^k \\
&\leq \mathbb{E}_{\alpha,\alpha' \sim \rho_t^{\text{MF}}}(\alpha\alpha')^k + \frac{O_K(1)}{\sqrt{d}} \\
&= \left(\mathbb{E}_{\alpha \sim \rho_t^{\text{MF}}} \alpha^k\right)^2 + \frac{O_K(1)}{\sqrt{d}}.
\end{aligned}$$

It follows that with $q_\sigma(z) = \sum_k q_k z^k$, we have

$$\begin{aligned}
\mathbb{E}_x(f_{\rho_t^{\text{MF}}}(x) - f^*(x))^2 &= \mathbb{E}_x(f^*(x))^2 + \mathbb{E}_x(f_{\rho_t^{\text{MF}}}(x))^2 - 2\mathbb{E} f^*(x) f_{\rho_t^{\text{MF}}}(x) \\
&= \sum_{k=k^*}^{K} q_k \left(1^k + \left(\mathbb{E}_{\alpha \sim \rho_t^{\text{MF}}} \alpha^k\right)^2 - 2\mathbb{E}_{\alpha \sim \rho_t^{\text{MF}}} \alpha^k\right) + \frac{O_K(1)}{\sqrt{d}} \\
&= \sum_{k=k^*}^{K} q_k \left(1 - \mathbb{E}_{\alpha \sim \rho_t^{\text{MF}}} \alpha^k\right)^2 + \frac{O_K(1)}{\sqrt{d}}
\end{aligned}$$

Now for all $k > k^*$, with $1 - s := r := \mathbb{E}_{\alpha\sim\rho_t^{\text{MF}}}(\alpha)^{k^*}$, using ($\star$), we have

$$r^{\frac{k}{k^*}} \le \mathbb{E}_{\alpha'\sim\rho_t^{\text{MF}}}(\alpha')^k,$$

so

$$\begin{aligned}\left(1 - \mathbb{E}_{\alpha\sim\rho_t^{\text{MF}}}\alpha^k\right)^2 &\le \left(1 - r^{k/k^*}\right)^2 = \left(1 - (1-s)^{k/k^*}\right)^2 \\ &\le (1 - (1 - sk/k^*)) = O_K(s^2).\end{aligned}$$

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.

□

E.3 MF Convergence Analysis

Proposition 49 (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*

$$T(\delta) := \arg\min\{t : \mathbb{E}_x(f_{\rho_t^{\text{MF}}}(x) - f^*(x))^2 \le \delta^2\} = O_K(\sqrt{d}^{k^*-2}\delta^{-(k^*-1)}).$$

We also have the following implication (which we will use for the analysis of J_{max} and J_{avg}) for any $t \le T(\delta)$ and for any $\tau > 0$:

$$\mathbb{E}_{w\sim\rho_t^{\text{MF}}}[(\alpha(w))^{k^*-1}\mathbf{1}(\alpha(w) \le 1 - \tau)] \le \sqrt{d}^{-(k^*-2)}O_{K,\delta}\left(\frac{1}{\tau^{O_K(1)}}\right).$$

Proof. First we need to prove by induction on t that for all $t \le T(\delta)$, the hypothesis ($\star$) holds. First observe that it holds at time 0, because

$$\mathbb{P}_{w\sim\mathbb{S}^{d-1}}[\alpha(w) \ge \iota] \le \exp(\Theta(d/\iota^2)) \le \iota$$

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 45 and Lemma 46 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 50. *Suppose ($\star$) holds up to time t. For any $\tau \le 1/2$ and $\gamma \le \frac{1-\tau}{2}$, we have*

$$\mathbb{P}_{w\sim\rho_t^{\text{MF}}}[\alpha(w) \in [\gamma, 1 - \tau]] \le \frac{2}{\gamma^{k^*-2}}\sqrt{d}^{-(k^*-2)}\exp\left(O_K\left(\frac{\log(1/\tau)}{\delta}\right)\right)$$

Proof. We will show that

$$\mathbb{P}_{w\sim\rho_t^{\mathrm{MF}}}[\alpha(w)\in[\gamma,2\gamma]]\le\frac{1}{\gamma^{k^*-2}}\sqrt{d}^{-(k^*-2)}\exp\left(O_K\left(\frac{\log(1/\tau)}{\delta}\right)\right)$$

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 47 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

$$\alpha_t(w)-\alpha_t(w')\ge\left|\alpha_0(w)-\alpha_0(w')\right|\min_{\tilde{w}:\alpha_0(\tilde{w})\in[\alpha_0(w'),\alpha_0(w')]}\left(\frac{\alpha_t(\tilde{w})}{\alpha_0(\tilde{w})}\right)^{k^*-1}\exp\left(O_K\left(\frac{\log(\tau/(k^*-1))}{\delta}\right)\right)$$

$$\ge\left|\alpha_0(w)-\alpha_0(w')\right|\left(\frac{\gamma}{\alpha_0(w')}\right)^{k^*-1}\exp\left(O_K\left(\frac{\log(\tau)}{\delta}\right)\right),$$

Thus since $|\alpha_t(w)-\alpha_t(w')|\le\gamma$, we have that

$$\left|\alpha_0(w)-\alpha_0(w')\right|\le\frac{1}{\gamma^{k^*-2}}\left(\alpha_0(w')\right)^{k^*-1}\exp\left(O_K\left(\frac{\log(\tau)}{\delta}\right)\right).$$

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

$$I_\lambda:=\left[\frac{\lambda}{\sqrt{d}},\frac{\lambda}{\sqrt{d}}+\frac{1}{\gamma^{k^*-2}}\left(\frac{\lambda}{\sqrt{d}}\right)^{k^*-1}\exp\left(O_K\left(\frac{\log(\tau)}{\delta}\right)\right)\right]$$

for some λ. Recall that the distribution of $\alpha_0(w)$ under $w\sim\rho_0$ is $\frac{1}{\sqrt{d}}$-subgaussian. Thus

$$\begin{aligned}\mathbb{P}_{w\sim\rho_0}[\alpha_0(w)\in I_\lambda]&\le\frac{\lambda^{k^*-1}}{\gamma^{k^*-2}}\sqrt{d}^{-(k^*-2)}\exp\left(O_K\left(\frac{\log(1/\tau)}{\delta^2}\right)\right)\left(\exp(-\lambda^2)\right)\\&\le\frac{1}{\gamma^{k^*-2}}\sqrt{d}^{-(k^*-2)}\exp\left(O_K\left(\frac{\log(1/\tau)}{\delta}\right)\right).\end{aligned}$$

This proves the claim. □

Plugging $\gamma=\iota$ and $\tau=\iota$ into this claim yields that

$$\mathbb{P}_{w\sim\rho_t^{\mathrm{MF}}}[\alpha(w)\in[\iota,1-\iota]]\le\frac{2}{\iota^{k^*-2}}\sqrt{d}^{-(k^*-2)}\exp\left(O_K\left(\frac{\log(1/\tau)}{\delta}\right)\right)\le\iota,$$

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 $\alpha_0(w)\ge\frac{\delta^2}{\sqrt{d}}$, we have that

$$\alpha_t(w)\ge1-\frac{1}{2K}$$

for $t\ge\frac{\Theta(1)}{\delta^2(\alpha_0(w))^{k^*-2}}$. This arises directly from the fact that Lemma 45 guarantees that for $\alpha\ge\frac{\delta^2}{\sqrt{d}}$,

$$v(\alpha,t)\ge\Theta_K(\delta\alpha^{k^*-1}(1-\alpha^2)).$$

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 50 and the fact that $(1-\tau)^{\frac{1}{k^*-1}} \geq 1 - \frac{\tau}{k^*-1}$. This proves the additional clause.

□

E.4 Proving Assumptions of Theorem 1 for the Single Index Model.

We need to check that the problem $(f^*, \mathcal{D}_x, \rho_0)$ introduced in Section E.1 satisfies the Assumptions of Theorem 1.

Fix a desired loss δ, and let $T(\delta)$ be as in Proposition 49.

Local Strong Convexity.

Lemma 51 (Local Strong Convexity for SIM)**.** *If d is large enough, then for any $t \leq T(\delta)$, we have for any w with $|\xi_t(w) - w^* \operatorname{sign}(\xi_t(w)^T w^*)| \leq \frac{1}{5K}$,*

$$D_t^{\perp}(w) \preceq -\Omega_{K,C_{reg}}\left(\sqrt{L(\rho_t^{\mathrm{MF}})}\right).$$

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

$$D_t^{\perp}(w) = \frac{d}{dw}V(w_t, \rho_t^{\mathrm{MF}})$$

It is evident that $V(w_t, \rho_t^{\text{MF}})$ 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

$$V(w_t, \rho_t^{\text{MF}}) = v(\alpha, t)\frac{\tilde{w}}{\sqrt{1-\alpha^2}}.$$

We will consider the quadratic form $y^T D_t^\perp(w) y$ 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 $y^T D_t^\perp(w) y \le -\Omega_{K,C_{\text{reg}}}\left(\sqrt{L(\rho_t^{\text{MF}})}\right)\|y\|^2$.

Lets start with the first, letting $y = \tilde{w}$. We have

$$\begin{aligned} D_t^\perp(w)y &= \frac{dV(w, \rho_t^{\text{MF}})}{d(y^T w)} \\ &= \frac{v(\alpha,t)}{\sqrt{1-\alpha^2}}\frac{d\tilde{w}}{d(y^T w_t)} + v(\alpha,t)\tilde{w}\frac{d(1-\alpha^2)^{-1/2}}{d(y^T w_t)} + \left(\frac{\tilde{w}}{\sqrt{1-\alpha^2}}\right)\frac{dv(\alpha,t)}{d(y^T w_t)} \end{aligned}$$

Now

$$\left(\frac{\tilde{w}}{\sqrt{1-\alpha^2}}\right)\frac{dv(\alpha,t)}{d(y^T w_t)} = \left(\frac{\tilde{w}}{\sqrt{1-\alpha^2}}\right)\frac{dv(\alpha,t)}{d\alpha}\frac{d\alpha}{d(y^T w_t)} = \tilde{w}\frac{dv(\alpha,t)}{d\alpha}.$$

Next,

$$\begin{aligned} \frac{d(1-\alpha^2)^{-1/2}}{d(y^T w_t)} &= \frac{d(1-\alpha^2)^{-1/2}}{d\alpha}\frac{d\alpha}{d(y^T w_t)} \\ &= \frac{-\alpha}{(1-\alpha^2)^{3/2}}\frac{1}{\sqrt{1-\alpha^2}} \\ &= \frac{\alpha}{(1-\alpha^2)}. \end{aligned}$$

Finally,

$$\frac{d\tilde{w}}{d(y^T w_t)} = 0$$

Thus in summary, putting these three terms together we have

$$y^T D_t^\perp(w) y = v(\alpha,t)\frac{\alpha}{(1-\alpha^2)} + \frac{dv(\alpha,t)}{d\alpha}.$$

By Lemma 46, we have for $y = \tilde{w}$,

$$y^T D_t^\perp(w) y \le -\Omega_{K,C_{\text{reg}}}\left(\sqrt{L(\rho_t^{\text{MF}})}\right).$$

Now we consider $y \perp \tilde{w}, w_t$. We have

$$\begin{aligned} y^T \frac{dV(w_t, \rho_t^{\text{MF}})}{d(y^T w_t)} &= y^T \tilde{w} \frac{d\left(\frac{v(\alpha,t)}{\sqrt{1-\alpha^2}}\right)}{dy^T w} + \frac{v(\alpha,t)}{\sqrt{1-\alpha^2}} y^T \frac{d\tilde{w}}{d(y^T w_t)} \\ &= 0 + \frac{v(\alpha,t)}{\sqrt{1-\alpha^2}} y^T \frac{d\tilde{w}}{d(y^T w_t)} \\ &= -\alpha \frac{v(\alpha,t)}{\sqrt{1-\alpha^2}} y^T \frac{dw_\perp}{d(y^T w_t)} \\ &= -\alpha \frac{v(\alpha,t)}{\sqrt{1-\alpha^2}} y^T \frac{y}{\sqrt{1-\alpha^2}} \\ &= -\frac{\alpha v(\alpha,t)}{1-\alpha^2} \|y\| \\ &\le -\Omega_{K, C_{\text{reg}}}\left(\sqrt{L(\rho_t^{\text{MF}})}\right). \end{aligned}$$

Here the final inequality follows from Lemma 45. □

Proving Assumption 2 for SIM. First we will need the following lemma.

Lemma 52. *For any w and $s \le t \le T(\delta)$, we have*

$$\left\| \frac{d\xi_{t,s}(w_s)}{dw}|_{w_s=\xi_s(w)} \right\| \le O_K\left(\left(\frac{\alpha_t(w)}{\alpha_s(w)}\right)^{k^*-1}\right) \exp\left(O_K\left(\frac{1}{\delta}\right)\right).$$

Proof. It suffices to check that this holds for times where $\alpha_t(w) \le \frac{1}{5K}$, because after that, by Lemma 46, $D_t^\perp(w)$ is negative definite, and so $\left\| \frac{d\xi_{t,s}(w)}{dw}|_{w=\xi_s(w)} \right\|$ can only decrease.

Claim 53. *In the setting of of the lemma, for any w with $\alpha_t(w) \le \frac{1}{5K}$, we have*

$$\left\| \frac{d\xi_{t,s}(w)}{dw}|_{w=\xi_s(w)} \right\| \le O_K\left(\left.\frac{d\alpha_{t,s}(z)}{dz}\right|_{z=\alpha_s(w)}\right) + 1.$$

Proof. Let $w_s = \xi_s(w)$. Without loss of generality assume $w_s^T w^* > 0$ such that $\alpha_s(w) = \xi_s(w)^T w^*$. 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

$$\frac{d\xi_{t,s}(w_s)}{dw_s} = \frac{d\alpha_{t,s}(w_s)}{dw_s} w^* + \frac{-\alpha_{t,s}(w_s)}{\sqrt{1-\alpha_{t,s}(w_s)^2}} \frac{d\alpha_{t,s}(w_s)}{dw_s} w_\perp + \frac{\sqrt{1-\alpha_{t,s}(w_s)^2}}{\sqrt{1-\alpha_s(w_s)^2}} P_{w^*}^\perp,$$

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 45) and $\alpha_t(w) \le 1 - \frac{1}{5K}$, we have

$$\left\| \frac{d\xi_{t,s}(w_s)}{dw} \right\| \le O_K\left(\frac{d\alpha_{t,s}(w_s)}{dw_s}\right) + 1.$$

□

The conclusion now follows from combining this claim and 47. □

We are now ready to bound $J_{\max}$ and J_{avg}.

Lemma 54. *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 52, for all w, we have

$$\|J_{t,s}^{\perp}(w)\| = O_{K,\delta}\left(\left(\frac{\alpha_t(w)}{\alpha_s(w)}\right)^{k^*-1}\right) \tag{E.4}$$

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 (E.2) 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 55) 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 (E.4) and Lemma 55.

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 49 bounds this second term, yielding

$$\begin{aligned}
\mathbb{E}_{w\sim\rho_0}\|J_{t,s}(w)H_s^{\perp}(w,w')v\|\mathbf{1}(\xi_t(w) \notin B_\tau) &\le O_{K,\delta}\left(\sqrt{d}^{(k^*-1)}\right) + \sqrt{d}^{-(k^*-2)} O_{K,\delta}\left(\frac{1}{\tau^{O_K(1)}}\right) \\
&= O_{K,\delta,\tau}(1/T(\delta)).
\end{aligned}$$

This proves the lemma. □

Lemma 55. *For any distribution μ over w, for and $w', v \in \mathbb{S}^{d-1}$, with $w_s := \xi_s(w)$, we have*

$$\sup_{w',v} \mathbb{E}_{w\sim\mu}\|H_s^{\perp}(w,w')v\| \lessapprox \sup_{\|u\|=1} \sqrt{\mathbb{E}_{w\sim\mu}({w_s}^T u)^{2(k^*-1)}}\|v\| + \sup_{\|u\|=1} \sqrt{\mathbb{E}_{w\sim\mu}({w_s}^T u)^{2(k^*-2)}({w_s}^T v)^2}.$$

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

$$\sup_{w',v} \mathbb{E}_{w\sim\mu}\|H_s^{\perp}(w,w')v\| \le O_K\Big(\alpha^{k-1} + \sqrt{d}^{-(k^*-1)}\Big).$$

Proof. [Proof of Lemma 55] By Cauchy-Schwartz,

$$\mathbb{E}_{w\sim\mu}\|H_s^{\perp}(w,w')v\| \le \sqrt{\mathbb{E}_{w\sim\mu} v(H_s^{\perp}(w,w'))^T H_s^{\perp}(w,w')v}.$$

Let us expand $H_s^{\perp}(w,w')$. With $w_s := \xi_s(w)$ and $:= \xi_s(w')$, we have

$$H_s^{\perp}(w,w') = \sum_{k=k^*-1}^{K-1} P_{w_s}^{\perp}\Big(c(w,w')({w_s}^T u)^k I + c'(w,w')({w_s}^T u)^{k-1} u {w_s}^T\Big)P_u^{\perp},$$

where $c(w,w'), c'(w,w') \le C_{\text{reg}}$. Thus we have

$$\begin{aligned} &H_s^{\perp}(w,w')^T H_s^{\perp}(w,w') \\ &\quad\preceq \sum_k 2C_{\text{reg}}({w_s}^T u)^{2k} P_u^{\perp} \\ &\qquad + 2C_{\text{reg}}({w_s}^T u)^{2(k-1)} P_u^{\perp} w_s u^T P_{w_s}^{\perp} u {w_s}^T P_u^{\perp} \\ &\quad\preceq 2C_{\text{reg}}({w_s}^T u)^{2(k^*-1)} I \\ &\qquad + 2C_{\text{reg}}({w_s}^T u)^{2(k^*-2)} w_s {w_s}^T, \end{aligned}$$

and thus

$$\mathbb{E}_{w\sim\mu} v(H_s^{\perp}(w,w'))^T H_s^{\perp}(w,w')v \le 2C_{\text{reg}}({w_s}^T u)^{2(k^*-1)}\|v\|^2 + 2C_{\text{reg}}({w_s}^T u)^{2(k^*-1)}(v^T w_s)^2.$$

Taking a square root yields the desired result. The second statement follows observing that $\mathbb{E}_w[(u^T w_s)^k] = O_k(\sqrt{d}^{-k})$ if u is in the span of the rotationally invariant directions, because $u^T w_s$ $\frac{1}{\sqrt{d}}$- subgaussian. □

Proof. [Proof of Theorem 2] Fix a desired loss δ, and let $T(\delta) = O_K(\sqrt{d}^{k^*-2}\delta^{k^*-1})$ be as in Proposition 49, such that

$$\mathbb{E}_x(f_{\rho_t^{\text{MF}}}(x) - f^*(x))^2 \le \delta.$$

Let us check the conditions of Theorem 1. First, the regularity conditions in Assumption 1 trivially hold for $C_{\text{reg}} = O_{C_{\text{SIM}}}(1)$ by our choice of Gaussian data and σ.

By Lemma 54, up to time $T(\delta)$, $(f^*, \rho_0, \mathcal{D}_x)$ satisfies Assumption 2 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 51, $(f^*, \rho_0, \mathcal{D}_x)$ 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 $\nabla_w V(w, \rho_t^{\text{MF}}) P_w^\perp$ in w, and the fact that at $\nabla_{\xi^\infty(w_i)} V(\xi^\infty(w_i), \rho_t^{\text{MF}}) P^\perp_{\xi^\infty(w_i)}$ (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 4 holds.

Finally, the symmetry conditions in Assumption 5 trivially hold because the data is Guassian, and there is a reflection symmetry between w^* and $-w^*$.

Now suppose $n \geq d^{11k^*} \geq J_{\text{max}}^8 (T(\delta))^6 d^4$ and $m \geq d^{13k^*} \geq J_{\text{max}}^{10} (T(\delta))^6 d^4$ such that

$$\epsilon_n + \epsilon_m = \frac{\log(n) d^{3/2}}{\sqrt{n}} + \frac{\log(mT) \max(d^{1/2} J_{\text{max}}, d^{3/2})}{\sqrt{m}} \leq \frac{1}{d J_{\text{max}}^4 T^3}.$$

Thus for d large enough, the condition on $\epsilon_n + \epsilon_m$ in Theorem 1 holds. Thus all the assumptions of Theorem 1 hold, and the result guarantees that for $t \leq T(\delta)$, with high probability over the draw of the data and of the neural network initialization, we have with $\lambda = \min(\tau, \delta)$,

$$\begin{aligned}
\mathbb{E}_x (f_{\rho_t^{\text{MF}}}(x) - f_{\rho_t^m}(x))^2 &\leq t J_{\text{max}} (\epsilon_m + \epsilon_n) \exp\left(\frac{O(t J_{\text{avg}}(\lambda))}{c\lambda - \Omega(J_{\text{avg}}/\lambda)} \right) \\
&\leq t d^{2(k^*-1)} (\epsilon_n + \epsilon_m) O_{K,\delta}(1).
\end{aligned}$$

Combining this with Equation E.4, we have that

$$\mathbb{E}_x (f_{\rho_t^{\text{MF}}}(x) - f_{\rho_t^m}(x))^2 \leq 2\delta^2 + 2t d^{2(k^*-1)} (\epsilon_n + \epsilon_m) O_{K,\delta}(1) \leq 3\delta^2.$$

This proves the theorem. □

F Discussion for Future Work