Appendix

We present the detailed proofs of the main results of the paper below. The appendix is organized as follows. We provide proofs to the simple propositions regarding the NTK presented in the paper in Appendix A, and prove the main results for $\mathbf{V}$-dominated and $\mathbf{G}$-dominated convergence in the settings of gradient flow and gradient descent in Appendices B and C. The proofs for gradient flow and gradient descent share the same main idea, yet the proof for gradient descent has a considerate number of additional technicalities. In Appendices D and E we prove the lemmas used in the analysis of Appendices B and C respectively. Before we move forward we highlight some of the challenges of the WN proof.

Distinctive aspects of the WN analysis The main idea of our proof are familiar and structured similarly to the work by Du et al. (2019b) on the un-normalized setting. However, the majority of the proofs are modified significantly to account for WN. To the best of our knowledge, the finite-step analysis that we present in Appendix C is entirely new, incorporating updates of both $\mathbf{v}$ and g. The proof of Theorem C.1 is crucially dependent on the geometry of WN gradient descent and the orthogonality property, in particular (2.3). Updates of the weights in both the numerator and denominator require additional analysis that is presented in Lemma B.10. In Appendix E we prove Theorems 4.1, 4.2 based on the general Theorem C.1 and Property 1 which is based on new detailed decomposition of the finite-step difference between iterations. In contrast to the un-normalized setting, the auxiliary matrices $\mathbf{V}^{\infty}, \mathbf{G}^{\infty}$ that we have in the WN analysis are not piece-wise constant in $\mathbf{v}$. To prove they are positive definite, we prove Lemma 4.1 based on two new constructive arguments. We develop the technical Lemma D.1 and utilize Bernstein's inequality to reduce the amount of required over-parametrization in our final bounds on the width m. The amount of over-parameterization in relation to the sample size n is reduced (from n^6 to n^4) through more careful arguments in Lemmas B.3 and B.4, which introduce an intermediate matrix $\hat{\mathbf{V}}(t)$ and follow additional geometrical identities. Lemma B.9 reduces the polynomial dependence on the failure probability δ to logarithmic dependence based on sub-Gaussian concentration. The denominator in the WN architecture necessities worst bound analysis which we handle in Lemma B.10 that is used throughout the proofs.

A. Weight Normalization Dynamics Proofs

In this section we provide proofs for Proposition 1, which describes the relation between vanilla and WeightNorm NTKs and Observation 1 of the paper.

Proof of Proposition 1:
We would like to show that $\mathbf{V}(0) + \mathbf{G}(0) = \mathbf{H}(0)$. For each entry, consider

$$(\mathbf{V}(0) + \mathbf{G}(0))_{ij} = \frac{1}{m}\sum_{k=1}^{m} \langle \mathbf{x}_i^{\mathbf{v}_k(0)^{\perp}},\ \mathbf{x}_j{}^{\mathbf{v}_k(0)^{\perp}} \rangle \mathbb{1}_{ik}(0)\mathbb{1}_{jk}(0) + \frac{1}{m}\sum_{k=1}^{m} \langle \mathbf{x}_i^{\mathbf{v}_k(0)},\ \mathbf{x}_j{}^{\mathbf{v}_k(0)} \rangle \mathbb{1}_{ik}(0)\mathbb{1}_{jk}(0).$$

Note that

$$\langle \mathbf{x}_i,\ \mathbf{x}_j \rangle = \langle \mathbf{x}_i^{\mathbf{v}_k(0)} + \mathbf{x}_i^{\mathbf{v}_k(0)^{\perp}},\ \mathbf{x}_j^{\mathbf{v}_k(0)} + \mathbf{x}_j^{\mathbf{v}_k(0)^{\perp}} \rangle = \langle \mathbf{x}_i^{\mathbf{v}_k(0)^{\perp}},\ \mathbf{x}_j{}^{\mathbf{v}_k(0)^{\perp}} \rangle + \langle \mathbf{x}_i^{\mathbf{v}_k(0)},\ \mathbf{x}_j{}^{\mathbf{v}_k(0)} \rangle.$$

This gives

$$(\mathbf{V}(0) + \mathbf{G}(0))_{ij} = \frac{1}{m}\sum_{k=1}^{m} \langle \mathbf{x}_i,\ \mathbf{x}_j \rangle \mathbb{1}_{ik}(0)\mathbb{1}_{jk}(0) = \mathbf{H}_{ij}(0)$$

which proves the claim. □

Proof of Observation 1:
We show that the initialization of the network is independent of α. Take $\alpha, \beta > 0$, and for each k, initialize $\mathbf{v}_k^{\alpha}, \mathbf{v}_k^{\beta}$ as

$$\mathbf{v}_k^{\alpha}(0) \sim N(0, \alpha^2\mathbf{I}), \quad \mathbf{v}_k^{\beta}(0) \sim N(0, \beta^2\mathbf{I}).$$

Then

$$\frac{\mathbf{v}_k^{\alpha}(0)}{\|\mathbf{v}_k^{\alpha}(0)\|_2} \sim \frac{\mathbf{v}_k^{\beta}(0)}{\|\mathbf{v}_k^{\beta}(0)\|_2} \sim \text{ Unif}(\mathcal{S}^{d-1}) \quad \text{(in distribution)}.$$

Hence the distribution of each neuron $\sigma\big(\frac{\mathbf{v}_k(0)}{\|\mathbf{v}_k(0)\|_2}\big)$ at initialization is independent of α. Next for $g_k(0)$, we note that

$$\|\mathbf{v}_k^\alpha(0)\|_2 \sim \frac{\alpha}{\beta}\|\mathbf{v}_k^\beta(0)\|_2.$$

Initializing $g_k^\alpha(0), g_k^\beta(0)$ as in (2.4),

$$g_k^\alpha(0) = \frac{\|\mathbf{v}_k(0)\|_2}{\alpha}, \quad g_k^\beta(0) = \frac{\|\mathbf{v}_k(0)\|_2}{\beta},$$

gives

$$g_k^\alpha(0), \quad g_k^\beta(0) \sim \chi_d, \quad \text{and} \quad \frac{g_k^\alpha(0)\mathbf{v}_k^\alpha(0)}{\|\mathbf{v}_k^\alpha(0)\|_2} \sim \frac{g_k^\beta(0)\mathbf{v}_k^\beta(0)}{\|\mathbf{v}_k^\beta(0)\|_2} \sim N(0, \mathbf{I}),$$

for all α, β. This shows that the network initialization is independent of α and is equivalent to the initialization of the un-normalized setting. Similarly, inspecting the terms in the summands of $\mathbf{V}(0), \mathbf{G}(0)$ shows that they are also independent of α. For

$$\mathbf{V}_{ij}(0) = \frac{1}{m}\sum_{k=1}^m \mathbb{1}_{ik}(0)\mathbb{1}_{jk}(0)\left(\frac{\alpha c_k \cdot g_k(0)}{\|\mathbf{v}_k(0)\|_2}\right)^2 \langle \mathbf{x}_i^{\mathbf{v}_k(0)^\perp}, \mathbf{x}_j^{\mathbf{v}_k(0)^\perp}\rangle$$

the terms $\mathbb{1}_{ik}(0)$, $\mathbf{x}_i^{\mathbf{v}_k(0)^\perp}$ are independent of scale, and the fraction in the summand is identically 1. $\mathbf{G}(0)$ defined as

$$\mathbf{G}_{ij}(0) = \frac{1}{m}\sum_{k=1}^m \mathbb{1}_{ik}(0)\mathbb{1}_{jk}(0)\langle \mathbf{x}_i^{\mathbf{v}_k(0)}, \mathbf{x}_j^{\mathbf{v}_k(0)}\rangle$$

is also invariant of scale since the projection onto a vector direction $\mathbf{v}_k(0)$ is independent of scale. □

B. Convergence Proof for Gradient Flow

In this section we derive the convergence results for gradient flow.

The main results are analogous to Theorems 4.1, 4.2 but by considering gradient flow instead of gradient descent the proofs are simplified. In Appendix C we prove the main results from Section 4 (Theorem 4.1, 4.2) for finite step gradient descent.

We state our convergence results for gradient flow.

Theorem B.1 ($\mathbf{V}$-dominated convergence)**.** *Suppose a network from the class (1.2) is initialized as in (2.4) with $\alpha < 1$ and that assumptions 1,2 hold. In addition, suppose the neural network is trained via the regression loss (2.5) with target $\mathbf{y}$ satisfying $\|\mathbf{y}\|_\infty = O(1)$. Then if $m = \Omega\big(n^4\log(n/\delta)/\lambda_0^4\big)$, WeightNorm training with gradient flow converges at a linear rate, with probability $1-\delta$, as*

$$\|\mathbf{f}(t) - \mathbf{y}\|_2^2 \le \exp(-\lambda_0 t/\alpha^2)\|\mathbf{f}(0) - \mathbf{y}\|_2^2.$$

This theorem is analogous to Theorem 4.1 but since here, the settings are of gradient flow there is no mention of the step-size. It is worth noting that smaller α leads to faster convergence and appears to not affect the other hypotheses of the flow theorem. This "un-interuptted" fast convergence behavior does not extend to finite-step gradient descent where the increased convergence rate is balanced by decreasing the allowed step-size.

The second main result for gradient flow is for $\mathbf{G}$-dominated convergence.

Theorem B.2 ($\mathbf{G}$-dominated convergence)**.** *Suppose a network from the class (1.2) is initialized as in (2.4) with $\alpha > 1$ and that assumptions 1, 2 hold. In addition, suppose the neural network is trained on the regression loss (2.5) with target $\mathbf{y}$ satisfying $\|\mathbf{y}\|_\infty = O(1)$. Then if $m = \Omega\big(\max\big\{n^4\log(n/\delta)/\alpha^4\mu_0^4, n^2\log(n/\delta)/\mu_0^2\big\}\big)$, WeightNorm training with gradient flow converges at a linear rate, with probability $1-\delta$, as*

$$\|\mathbf{f}(t) - \mathbf{y}\|_2^2 \le \exp(-\mu_0 t)\|\mathbf{f}(0) - \mathbf{y}\|_2^2.$$

B.1. Proof Sketch

To prove the results above we follow the steps introduced in the proof sketch of Section 4. The main idea of the proofs for $\mathbf{V}$ and $\mathbf{G}$ dominated convergence are analogous and a lot of the proofs are based of Du et al. (2019b). We show that in each regime, we attain linear convergence by proving that the least eigenvalue of the evolution matrix $\mathbf{\Lambda}(t)$ is strictly positive. For the $\mathbf{V}$-dominated regime we lower bound the least eigenvalue of $\mathbf{\Lambda}(t)$ as $\lambda_{\min}(\mathbf{\Lambda}(t)) \geq \lambda_{\min}(\mathbf{V}(t))/\alpha^2$ and in the $\mathbf{G}$-dominated regime we lower bound the least eigenvalue as $\lambda_{\min}(\mathbf{\Lambda}(t)) \geq \lambda_{\min}(\mathbf{G}(t))$.

The main part of the proof is showing that $\lambda_{\min}(\mathbf{V}(t)), \lambda_{\min}(\mathbf{G}(t))$ stay uniformly positive. We use several lemmas to show this claim.

In each regime, we first show that at initialization the kernel under consideration, $\mathbf{V}(0)$ or $\mathbf{G}(0)$, has a positive least eigenvalue. This is shown via concentration to an an auxiliary kernel (Lemmas B.1, B.2), and showing that the auxiliary kernel is also strictly positive definite (Lemma 4.1).

Lemma B.1. *Let $\mathbf{V}(0)$ and $\mathbf{V}^\infty$ be defined as in (3.3) and (4.2), assume the network width m satisfies $m = \Omega\big(\frac{n^2\log(n/\delta)}{\lambda_0^2}\big)$. Then with probability $1-\delta$,*

$$\|\mathbf{V}(0) - \mathbf{V}^\infty\|_2 \leq \frac{\lambda_0}{4}.$$

Lemma B.2. *Let $\mathbf{G}(0)$ and $\mathbf{G}^\infty$ be defined as in (3.4) and (4.3), assume m satisfies $m = \Omega\big(\frac{n^2\log(n/\delta)}{\mu_0^2}\big)$. Then with probability $1-\delta$,*

$$\|\mathbf{G}(0) - \mathbf{G}^\infty\|_2 \leq \frac{\mu_0}{4}.$$

After showing that $\mathbf{V}(0), \mathbf{G}(0)$ have a positive least-eigenvalue we show that $\mathbf{V}(t), \mathbf{G}(t)$ maintain this positive least eigenvalue during training. This part of the proof depends on the over-parametrization of the networks. The main idea is showing that if the individual parameters $\mathbf{v}_k(t), g_k(t)$ do not change too much during training, then $\mathbf{V}(t), \mathbf{G}(t)$ remain close enough to $\mathbf{V}(0), \mathbf{G}(0)$ so that they are still uniformly strictly positive definite. We prove the results for $\mathbf{V}(t)$ and $\mathbf{G}(t)$ separately since each regime imposes different restrictions on the trajectory of the parameters.

For now, in Lemmas B.3, B.4, B.5, we make assumptions on the parameters of the network not changing "too much"; later we show that this holds and is the result of over-parametrization. Specifically, over-parametrization ensures that the parameters stay at a small maximum distance from their initialization.

$\mathbf{V}$-dominated convergence To prove the least eigenvalue condition on $\mathbf{V}(t)$, we introduce the surrogate Gram matrix $\hat{\mathbf{V}}(t)$ defined entry-wise as

$$\hat{\mathbf{V}}_{ij}(t) = \frac{1}{m}\sum_{k=1}^{m} \langle \mathbf{x}_i^{\mathbf{v}_k(t)^\perp},\ \mathbf{x}_j^{\mathbf{v}_k(t)^\perp}\rangle \mathbb{1}_{ik}(t)\mathbb{1}_{jk}(t). \tag{B.1}$$

This definition aligns with $\mathbf{V}(t)$ if we replace the scaling term $\big(\frac{\alpha c_k g_k(t)}{\|\mathbf{v}_k(t)\|_2}\big)^2$ in each term in the sum $\mathbf{V}_{ij}(t)$ by 1.

To monitor $\mathbf{V}(t) - \mathbf{V}(0)$ we consider $\hat{\mathbf{V}}(t) - \mathbf{V}(0)$ and $\mathbf{V}(t) - \hat{\mathbf{V}}(t)$ in Lemmas B.3 and B.4 respectively:

Lemma B.3 (Rectifier sign-changes)**.** *Suppose $\mathbf{v}_1(0), \ldots, \mathbf{v}_k(0)$ are sampled i.i.d. as (2.4). In addition assume we have $m = \Omega\big(\frac{(m/\delta)^{1/d} n\log(n/\delta)}{\lambda_0}\big)$ and $\|\mathbf{v}_k(t) - \mathbf{v}_k(0)\|_2 \leq \frac{\alpha\lambda_0}{96n(m/\delta)^{1/d}} =: R_v$. Then with probability $1-\delta$,*

$$\|\hat{\mathbf{V}}(t) - \mathbf{V}(0)\|_2 \leq \frac{\lambda_0}{8}.$$

Lemma B.4. *Define*

$$R_g = \frac{\lambda_0}{48n(m/\delta)^{1/d}}, \quad R_v = \frac{\alpha\lambda_0}{96n(m/\delta)^{1/d}}. \tag{B.2}$$

Suppose the conditions of Lemma B.3 hold, and that $\|\mathbf{v}_k(t) - \mathbf{v}_k(0)\|_2 \le R_v$, $\|g_k(t) - g_k(0)\|_2 \le R_g$ *for all* $1 \le k \le m$. *Then with probability* $1-\delta$,

$$\|\mathbf{V}(t) - \mathbf{V}(0)\|_2 \le \frac{\lambda_0}{4}.$$

G-dominated convergence We ensure that $\mathbf{G}(t)$ stays uniformly positive definite if the following hold.

Lemma B.5. *Given* $\mathbf{v}_1(0), \dots, \mathbf{v}_k(0)$ *generated i.i.d. as in (2.4), suppose that for each* k, $\|\mathbf{v}_k(t) - \mathbf{v}_k(0)\|_2 \le \frac{\sqrt{2\pi}\alpha\mu_0}{8n(m/\delta)^{1/d}} =: \tilde{R}_v$, *then with probability* $1-\delta$,

$$\|\mathbf{G}(t) - \mathbf{G}(0)\|_2 \le \frac{\mu_0}{4}.$$

After deriving sufficient conditions to maintain a positive least eigenvalue at training, we restate the discussion of linear convergence from Section 4 formally.

Lemma B.6. *Consider the linear evolution* $\frac{d\mathbf{f}}{dt} = -\big(\mathbf{G}(t) + \frac{\mathbf{V}(t)}{\alpha^2}\big)(\mathbf{f}(t) - \mathbf{y})$ *from (3.5). Suppose that* $\lambda_{\min}\big(\mathbf{G}(t) + \frac{\mathbf{V}(t)}{\alpha^2}\big) \ge \frac{\omega}{2}$ *for all times* $0 \le t \le T$. *Then*

$$\|\mathbf{f}(t) - \mathbf{y}\|_2^2 \le \exp(-\omega t)\|\mathbf{f}(0) - \mathbf{y}\|_2^2$$

for all times $0 \le t \le T$.

Using the linear convergence result of Lemma B.6, we can now bound the trajectory of the parameters from their initialization.

Lemma B.7. *Suppose that for all* $0 \le t \le T$, $\lambda_{\min}\left(\mathbf{G}(t) + \frac{1}{\alpha^2}\mathbf{V}(t)\right) \ge \frac{\omega}{2}$ *and* $|g_k(t) - g_k(0)| \le R_g \le 1/(m/\delta)^{1/d}$. *Then with probability* $1-\delta$ *over the initialization*

$$\|\mathbf{v}_k(t) - \mathbf{v}_k(0)\|_2 \le \frac{4\sqrt{n}\|\mathbf{f}(0) - \mathbf{y}\|_2}{\alpha\omega\sqrt{m}} =: R_v' \tag{B.3}$$

for each k *and all times* $0 \le t \le T$.

Lemma B.8. *Suppose that for all* $0 \le t \le T$, $\lambda_{\min}\left(\mathbf{G}(t) + \frac{1}{\alpha^2}\mathbf{V}(t)\right) \ge \frac{\omega}{2}$. *Then with probability* $1-\delta$ *over the initialization*

$$|g_k(t) - g_k(0)| \le \frac{4\sqrt{n}\|\mathbf{f}(0) - \mathbf{y}\|_2}{\sqrt{m}\omega} =: R_g'$$

for each k *and all times* $0 \le t \le T$.

The distance of the parameters from initialization depends on the convergence rate (which depends on $\lambda_{\min}(\mathbf{\Lambda}(t))$) and the width of the network m. We therefore are able to find sufficiently large m for which the maximum parameter trajectories are not too large so that we have that the least eigenvalue of $\mathbf{\Lambda}(t)$ is bounded from 0; this proves the main claim.

Before proving the main results in the case of gradient flow, we use two more technical lemmas.

Lemma B.9. *Suppose that the network is initialized as (2.4) and that* $\mathbf{y} \in \mathbb{R}^n$ *has bounded entries* $|y_i| \le M$. *Then* $\|\mathbf{f}(0) - \mathbf{y}\|_2 \le C\sqrt{n\log(n/\delta)}$ *for some absolute constant* $C > 0$.

Lemma B.10 (Failure over initialization)**.** *Suppose* $\mathbf{v}_1(0), \dots, \mathbf{v}_k(0)$ *are initialized i.i.d. as in (2.4) with input dimension* d. *Then with probability* $1-\delta$,

$$\max_{k\in[m]} \frac{1}{\|\mathbf{v}_k(0)\|_2} \le \frac{(m/\delta)^{1/d}}{\alpha}.$$

In addition by (2.3), for all $t \ge 0$, *with probability* $1-\delta$,

$$\max_{k\in[m]} \frac{1}{\|\mathbf{v}_k(t)\|_2} \le \frac{(m/\delta)^{1/d}}{\alpha}.$$

Remark (Assumption 2). *Predominately, machine learning applications reside in the high dimensional regime with $d \geq 50$. Typically $d \gg 50$. This therefore leads to an expression $(m/\delta)^{1/d}$ that is essentially constant. For example, if $d = 50$, for $\max_{k\in[m]} \frac{1}{\|\mathbf{v}_k(0)\|_2} \geq 10$, one would need $m/\delta \geq 10^{80}$ (the tail of χ^2_d also has a factor of $(d/2)! \cdot 2^{d/2}$ which makes the assumption even milder). The term $(m/\delta)^{1/d}$ therefore may be taken as a constant for practicality,*

$$\max_{k\in[m]} \frac{1}{\|\mathbf{v}_k(0)\|_2} \leq \frac{C}{\alpha}.$$

While we make Assumption 2 when presenting our final bounds, for transparency we do not use Assumption 2 during our analysis and apply it only when we present the final over-parametrization results to avoid the overly messy bound. Without the assumption the theory still holds yet the over-parametrization bound worsens by a power $1 + 1/(d-1)$. This is since the existing bounds can be modified, replacing m with $m^{1-\frac{1}{d}}$.

Proof of Theorem B.1:
By substituting $m = \Omega\big(n^4 \log(n/\delta)/\lambda_0^4\big)$ and using the bound on $\|\mathbf{f}(0) - \mathbf{y}\|_2$ of Lemma B.9, a direct calculation shows that

$$\|\mathbf{v}_k(t) - \mathbf{v}_k(0)\|_2 \overset{B.7}{\leq} \frac{\alpha\sqrt{n}\|\mathbf{f}(0) - \mathbf{y}\|_2}{\sqrt{m}\lambda_0} \leq R_v.$$

Similarly m ensures that

$$|g_k(t) - g_k(0)| \overset{B.8}{\leq} \frac{\alpha^2\sqrt{n}\|\mathbf{f}(0) - \mathbf{y}\|_2}{\sqrt{m}\lambda_0} \leq R_g.$$

The over-parametrization of m implies that the parameter trajectories stay close enough to initialization to satisfy the hypotheses of Lemmas B.3, B.4 and that $\lambda_{\min}(\mathbf{\Lambda}(t)) \geq \lambda_{\min}(\mathbf{V}(t))/\alpha^2 \geq \frac{\lambda_0}{2\alpha^2}$. To prove that $\lambda_{\min}(\mathbf{\Lambda}(t)) \geq \frac{\lambda_0}{2\alpha^2}$ holds for all $0 \leq t \leq T$, we proceed by contradiction and suppose that one of Lemmas B.7, B.8 does not hold. Take T_0 to be the first failure time. Clearly $T_0 > 0$ and for $0 < t < T_0$ the above conditions hold, which implies that $\lambda_{\min}(\mathbf{V}(t)) \geq \frac{\lambda_0}{2}$ for $0 \leq t \leq T_0$; this contradicts one of Lemmas B.7, B.8 at time T_0. Therefore we conclude that Lemmas B.7, B.8 hold for $t > 0$ and we can apply B.6 to guarantee linear convergence. □

Here we consider the case where the convergence is dominated by $\mathbf{G}$. This occurs when $\alpha > 1$.
Proof of Theorem B.2:
By substituting $m = \Omega\big(n^4 \log(n/\delta)/\alpha^4\mu_0^4\big)$ and using the bound on $\|\mathbf{f}(0) - \mathbf{y}\|_2$ of Lemma B.9 we have that

$$\|\mathbf{v}_k(t) - \mathbf{v}_k(0)\|_2 \overset{B.7}{\leq} \frac{4\sqrt{n}\|\mathbf{f}(0) - \mathbf{y}\|_2}{\alpha\mu_0\sqrt{m}} \overset{B.9}{\leq} \frac{Cn\sqrt{\log(n/\delta)}}{\alpha\mu_0\sqrt{m}} \leq \tilde{R}_v.$$

Where the inequality is shown by a direct calculation substituting m.

This means that the parameter trajectories stay close enough to satisfy the hypotheses of Lemma B.5 if $m = \Omega\big(n^4 \log(n/\delta)/\alpha^4\mu_0^4\big)$. Using the same argument as Theorem B.1, we show that this holds for all $t > 0$. We proceed by contradiction, supposing that one of Lemmas B.7, B.8 do not hold. Take T_0 to be the first time one of the conditions of Lemmas B.7, B.8 fail. Clearly $T_0 > 0$ and for $0 < t < T_0$ the above derivation holds, which implies that $\lambda_{\min}(\mathbf{G}(t)) \geq \frac{\mu_0}{2}$. This contradicts Lemmas B.7 B.8 at time T_0, therefore we conclude that Lemma B.6 holds for all $t > 0$ and guarantees linear convergence. □

Note that if α is large, the required complexity on m is reduced. Taking $\alpha = \Omega(\sqrt{n/\mu_0})$ gives the improved bound

$$m = \Omega\left(\frac{n^2 \log(n/\delta)}{\mu_0^2}\right).$$

C. Finite Step-size Training

The general technique of proof for gradient flow extends to finite-step gradient descent. Nonethless, proving convergence for WeightNorm gradient descent exhibits additional complexities arising from the discrete updates and joint training with the new parametrization (1.2). We first introduce some needed notation.

Define $S_i(R)$ as the set of indices $k \in [m]$ corresponding to neurons that are close to the activity boundary of ReLU at initialization for a data point $\mathbf{x}_i$,

$$S_i(R) := \{k \in [m] : \exists\, \mathbf{v} \text{ with } \|\mathbf{v} - \mathbf{v}_k(0)\|_2 \leq R \text{ and } \mathbb{1}_{ik}(0) \neq \mathbb{1}\{\mathbf{v}^\top \mathbf{x}_i \geq 0\}\}.$$

We upper bound the cardinality of $|S_i(R)|$ with high probability.

Lemma C.1. *With probability $1-\delta$, we have that for all i*

$$|S_i(R)| \leq \frac{\sqrt{2}mR}{\sqrt{\pi}\alpha} + \frac{16\log(n/\delta)}{3}.$$

Next we review some additional lemmas needed for the proof of Theorems 4.1, 4.2. Analogous to Lemmas B.7, B.8, we bound the finite-step parameter trajectories in Lemmas C.2, C.3.

Lemma C.2. *Suppose the norm of $\|\mathbf{f}(s) - \mathbf{y}\|_2^2$ decreases linearly for some convergence rate ω during gradient descent training for all iteration steps $s = 0, 1, \ldots, K$ with step-size η as $\|\mathbf{f}(s) - \mathbf{y}\|_2^2 \leq (1 - \frac{\eta\omega}{2})^s \|\mathbf{f}(0) - \mathbf{y}\|_2^2$. Then for each k we have*

$$|g_k(s) - g_k(0)| \leq \frac{4\sqrt{n}\|\mathbf{f}(0) - \mathbf{y}\|_2}{\sqrt{m}\omega}$$

for iterations $s = 0, 1, \ldots, K+1$.

Lemma C.3. *Under the assumptions of Lemma C.2, suppose in addition that $|g_k(s) - g_k(0)| \leq 1/(m/\delta)^{1/d}$ for all iterations steps $s = 0, 1, \ldots K$. Then for each k,*

$$\|\mathbf{v}_k(s) - \mathbf{v}_k(0)\|_2 \leq \frac{8\sqrt{n}\|\mathbf{f}(0) - \mathbf{y}\|_2}{\alpha\sqrt{m}\omega}$$

for $s = 0, 1, \ldots, K+1$.

To prove linear rate of convergence we analyze the $s+1$ iterate error $\|\mathbf{f}(s+1) - \mathbf{y}\|_2$ relative to that of the s iterate, $\|\mathbf{f}(s) - \mathbf{y}\|_2$. Consider the network's coordinate-wise difference in output between iterations, $f_i(s+1) - f_i(s)$, writing this explicitly based on gradient descent updates yields

$$f_i(s+1) - f_i(s) = \frac{1}{\sqrt{m}} \sum_{k=1}^{m} \frac{c_k g_k(s+1)}{\|\mathbf{v}_k(s+1)\|_2} \sigma(\mathbf{v}_k(s+1)^\top \mathbf{x}_i) - \frac{c_k g_k(s)}{\|\mathbf{v}_k(s)\|_2} \sigma(\mathbf{v}_k(s)^\top \mathbf{x}_i). \tag{C.1}$$

We now decompose the summand in (C.1) looking at the updates in each layer, $f_i(s+1) - f_i(s) = a_i(s) + b_i(s)$ with

$$a_i(s) = \frac{1}{\sqrt{m}} \sum_{k=1}^{m} \frac{c_k g_k(s+1)}{\|\mathbf{v}_k(s+1)\|_2} \sigma(\mathbf{v}_k(s)^\top \mathbf{x}_i) - \frac{c_k g_k(s)}{\|\mathbf{v}_k(s)\|_2} \sigma(\mathbf{v}_k(s)^\top \mathbf{x}_i),$$

$$b_i(s) = \frac{1}{\sqrt{m}} \sum_{k=1}^{m} \frac{c_k g_k(s+1)}{\|\mathbf{v}_k(s+1)\|_2} \big(\sigma(\mathbf{v}_k(s+1)^\top \mathbf{x}_i) - \sigma(\mathbf{v}_k(s)^\top \mathbf{x}_i)\big).$$

Further, each layer summand is then subdivided into a primary term and a residual. $a_i(s)$, corresponding to the difference in the first layer $\left(\frac{c_k g_k(s+1)}{\|\mathbf{v}_k(s+1)\|_2} - \frac{c_k g_k(s)}{\|\mathbf{v}_k(s)\|_2}\right)$, is subdivided into $a_i^I(s)$ and $a_i^{II}(s)$ as follows:

$$a_i^I(s) = \frac{1}{\sqrt{m}} \sum_{k=1}^{m} \left(\frac{c_k g_k(s+1)}{\|\mathbf{v}_k(s)\|_2} - \frac{c_k g_k(s)}{\|\mathbf{v}_k(s)\|_2}\right) \sigma(\mathbf{v}_k(s)^\top \mathbf{x}_i), \tag{C.2}$$

$$a_i^{II}(s) = \frac{1}{\sqrt{m}} \sum_{k=1}^{m} \left(\frac{c_k g_k(s+1)}{\|\mathbf{v}_k(s+1)\|_2} - \frac{c_k g_k(s+1)}{\|\mathbf{v}_k(s)\|_2}\right) \sigma(\mathbf{v}_k(s)^\top \mathbf{x}_i). \tag{C.3}$$

$b_i(s)$ is sub-divided based on the indices in the set S_i that monitor the changes of the rectifiers. For now, $S_i = S_i(R)$ with R to be set later in the proof. $b_i(s)$ is partitioned to summands in the set S_i and the complement set,

$$b_i^I(s) = \frac{1}{\sqrt{m}} \sum_{k \notin S_i} \frac{c_k g_k(s+1)}{\|\mathbf{v}_k(s+1)\|_2} \big(\sigma(\mathbf{v}_k(s+1)^\top \mathbf{x}_i) - \sigma(\mathbf{v}_k(s)^\top \mathbf{x}_i)\big),$$
$$b_i^{II}(s) = \frac{1}{\sqrt{m}} \sum_{k \in S_i} \frac{c_k g_k(s+1)}{\|\mathbf{v}_k(s+1)\|_2} \big(\sigma(\mathbf{v}_k(s+1)^\top \mathbf{x}_i) - \sigma(\mathbf{v}_k(s)^\top \mathbf{x}_i)\big).$$

With this sub-division in mind, the terms corresponding to convergence are $\mathbf{a}^I(s), \mathbf{b}^I(s)$ whereas $\mathbf{a}^{II}(s), \mathbf{b}^{II}(s)$ are residuals that are the result of discretization. We define the primary and residual vectors $\mathbf{p}(s), \mathbf{r}(s)$ as

$$\mathbf{p}(s) = \frac{\mathbf{a}_I(s) + \mathbf{b}_I(s)}{\eta}, \quad \mathbf{r}(s) = \frac{\mathbf{a}_{II} + \mathbf{b}_{II}(s)}{\eta}. \tag{C.4}$$

If the residual $\mathbf{r}(s)$ is sufficiently small and $\mathbf{p}(s)$ may be written as $\mathbf{p}(s) = -\mathbf{\Lambda}(s)(\mathbf{f}(s) - \mathbf{y})$ for some iteration dependent evolution matrix $\mathbf{\Lambda}(s)$ that has

$$\lambda_{\min}(\mathbf{\Lambda}(s)) = \omega/2 \tag{C.5}$$

for $\omega > 0$ then the neural network (1.2) converges linearly when trained with WeightNorm gradient descent of step size η. We formalize the condition on $\mathbf{r}(s)$ below and later derive the conditions on the over-parametrization (m) ensuring that $\mathbf{r}(s)$ is sufficiently small.

Property 1. *Given a network from the class (1.2) initialized as in (2.4) and trained with gradient descent of step-size η, define the residual $\mathbf{r}(s)$ as in (C.4) and take ω as in (C.5). We specify the "residual condition" at iteration s as*

$$\|\mathbf{r}(s)\|_2 \leq c\omega \|\mathbf{f}(s) - \mathbf{y}\|_2$$

for a sufficiently small constant $c > 0$ independent of the data or initialization.

Here we present Theorem C.1 which is the backbone of Theorems 4.1 and 4.2.

Theorem C.1. *Suppose a network from the class (1.2) is trained via WeightNorm gradient descent with an evolution matrix $\mathbf{\Lambda}(s)$ as in (C.5) satisfying $\lambda_{\min}(\mathbf{\Lambda}(s)) \geq \omega/2$ for $s = 0, 1, \dots K$. In addition if the data meets assumptions 1, 2, the step-size η of gradient descent satisfies $\eta \leq \frac{1}{3\|\mathbf{\Lambda}(s)\|_2}$ and that the residual $\mathbf{r}(s)$ defined in (C.4) satisfies Property 1 for $s = 0, 1, \dots, K$ then we have that*

$$\|\mathbf{f}(s) - \mathbf{y}\|_2^2 \leq \left(1 - \frac{\eta\omega}{2}\right)^s \|\mathbf{f}(0) - \mathbf{y}\|_2^2$$

for $s = 0, 1, \dots, K$.

Proof of Theorem C.1:
This proof provides the foundation for the main theorems. In the proof we also derive key bounds to be used in Theorems 4.1, 4.2. We use the decomposition we described above and consider again the difference between consecutive terms $\mathbf{f}(s+1) - \mathbf{f}(s)$,

$$f_i(s+1) - f_i(s) = \frac{1}{\sqrt{m}} \sum_{k=1}^{m} \frac{c_k g_k(s+1)}{\|\mathbf{v}_k(s+1)\|_2} \sigma(\mathbf{v}_k(s+1)^\top \mathbf{x}_i) - \frac{c_k g_k(s)}{\|\mathbf{v}_k(s)\|_2} \sigma(\mathbf{v}_k(s)^\top \mathbf{x}_i). \tag{C.6}$$

Following the decomposition introduced in (C.2), $a_i^I(s)$ is re-written in terms of $\mathbf{G}(s)$,

$$\begin{aligned}
a_i^I(s) &= \frac{1}{\sqrt{m}}\sum_{k=1}^m \frac{c_k}{\|\mathbf{v}_k(s)\|_2}\bigg(-\eta\frac{\partial L(s)}{\partial g_k}\bigg)\sigma(\mathbf{v}_k(s)^\top \mathbf{x}_i) \\
&= -\frac{\eta}{m}\sum_{k=1}^m \frac{c_k}{\|\mathbf{v}_k(s)\|_2}\sum_{j=1}^n (f_j(s)-y_j)\frac{c_k}{\|\mathbf{v}_k(s)\|_2}\sigma(\mathbf{v}_k^\top(s)\mathbf{x}_j)\sigma(\mathbf{v}_k^\top(s)\mathbf{x}_i) \\
&= -\eta\sum_{j=1}^n (f_j(s)-y_j)\frac{1}{m}\sum_{k=1}^m (c_k)^2\sigma\bigg(\frac{\mathbf{v}_k(s)^\top\mathbf{x}_i}{\|\mathbf{v}_k(s)\|_2}\bigg)\sigma\bigg(\frac{\mathbf{v}_k(s)^\top\mathbf{x}_j}{\|\mathbf{v}_k(s)\|_2}\bigg) \\
&= -\eta\sum_{j=1}^n (f_j(s)-y_j)\mathbf{G}_{ij}(s),
\end{aligned}$$

where the first equality holds by the gradient update rule $g_k(s+1) = g_k(s) - \eta\nabla_{g_k}L(s)$. In this proof we also derive bounds on the residual terms of the decomposition which we will aid us in the proofs of Theorems 4.1, 4.2. $a_i^I(s)$ is the primary term of $a_i(s)$, now we bound the residual term $a_i^{II}(s)$. Recall $a_i^{II}(s)$ is written as

$$a_i^{II}(s) = \frac{1}{\sqrt{m}}\sum_{k=1}^m\bigg(\frac{c_k g_k(s+1)}{\|\mathbf{v}_k(s+1)\|_2} - \frac{c_k g_k(s+1)}{\|\mathbf{v}_k(s)\|_2}\bigg)\sigma(\mathbf{v}_k(s)^\top\mathbf{x}_i),$$

which corresponds to the difference in the normalization in the second layer. Since $\nabla_{\mathbf{v}_k}L(s)$ is orthogonal to $\mathbf{v}_k(s)$ we have that

$$\begin{aligned}
&c_k g_k(s+1)\bigg(\frac{1}{\|\mathbf{v}_k(s+1)\|_2} - \frac{1}{\|\mathbf{v}_k(s)\|_2}\bigg)\sigma(\mathbf{v}_k(s)^\top\mathbf{x}_i) \\
&= c_k g_k(s+1)\bigg(\frac{1}{\sqrt{\|\mathbf{v}_k(s)\|_2^2 + \eta^2\|\nabla_{\mathbf{v}_k}L(s)\|_2^2}} - \frac{1}{\|\mathbf{v}_k(s)\|_2}\bigg)\sigma(\mathbf{v}_k(s)^\top\mathbf{x}_i) \\
&= \frac{-c_k g_k(s+1)\eta^2\|\nabla_{\mathbf{v}_k}L(s)\|_2^2}{\|\mathbf{v}_k(s+1)\|_2\|\mathbf{v}_k(s)\|_2(\|\mathbf{v}_k(s)\|_2 + \|\mathbf{v}_k(s+1)\|_2)}\sigma(\mathbf{v}_k(s)^\top\mathbf{x}_i) \\
&\le \frac{-c_k g_k(s+1)\eta^2\|\nabla_{\mathbf{v}_k}L(s)\|_2^2}{2\|\mathbf{v}_k(s)\|_2\|\mathbf{v}_k(s+1)\|_2}\sigma\bigg(\frac{\mathbf{v}_k(s)^\top\mathbf{x}_i}{\|\mathbf{v}_k(s)\|_2}\bigg),
\end{aligned}$$

where the first equality above is by completing the square, and the inequality is due to the increasing magnitudes of $\|\mathbf{v}_k(s)\|_2$.

Since $0 \le \sigma\left(\frac{\mathbf{v}_k(s)^\top\mathbf{x}_i}{\|\mathbf{v}_k(s)\|_2}\right) \le 1$, the above can be bounded as

$$\begin{aligned}
|a_i^{II}(s)| &\le \frac{1}{\sqrt{m}}\sum_{k=1}^m\bigg|\frac{g_k(s+1)\eta^2\|\nabla_{\mathbf{v}_k}L(s)\|_2^2}{2\|\mathbf{v}_k(s)\|_2\|\mathbf{v}_k(s+1)\|_2}\bigg| \\
&\le \frac{1}{\sqrt{m}}\sum_{k=1}^m\frac{\eta^2\big(1+R_g(m/\delta)^{1/d}\big)^3 n\|\mathbf{f}(s)-\mathbf{y}\|_2^2(m/\delta)^{1/d}}{\alpha^4 m} \\
&= \frac{\eta^2 n\big(1+R_g(m/\delta)^{1/d}\big)^3\|\mathbf{f}(s)-\mathbf{y}\|_2^2(m/\delta)^{1/d}}{\alpha^4\sqrt{m}}.
\end{aligned} \tag{C.7}$$

The second inequality is the result of applying the bound in equation (E.1) on the gradient norm $\|\nabla_{\mathbf{v}_k}L(s)\|_2$ and using Lemma B.10.

Next we analyze $b_i(s)$ and sub-divide it based on the sign changes of the rectifiers. Define the set $S_i \coloneqq S_i(R)$ as in Lemma C.1 with R taken to be such that $\|\mathbf{v}_k(s+1)-\mathbf{v}_k(0)\|_2 \le R$ for all k. Take $b_i^{II}(s)$ as the sub-sum of $b_i(s)$ with indices k from the set S_i.

$b_i^I(s)$ corresponds to the sub-sum with the remaining indices. By the definition of S_i, for $k \notin S_i$ we have that $\mathbb{1}_{ik}(s+1) = \mathbb{1}_{ik}(s)$. This enables us to factor $\mathbb{1}_{ik}(s)$ and represent $b_i^I(s)$ as a Gram matrix similar to $\mathbf{V}(s)$ with a correction term from

the missing indices in S_i.

$$b_i^I(s) = -\frac{1}{\sqrt{m}} \sum_{k \notin S_i} \left(\frac{c_k g_k(s+1)}{\|\mathbf{v}_k(s+1)\|_2} \right) \left(\eta \langle \nabla_{\mathbf{v}_k} L(s),\ \mathbf{x}_i \rangle \right) \mathbb{1}_{ik}(s)$$
$$= -\frac{\eta}{m} \sum_{k \notin S_i} \left(\frac{c_k g_k(s+1)}{\|\mathbf{v}_k(s+1)\|_2} \right) \left(\frac{c_k g_k(s)}{\|\mathbf{v}_k(s)\|_2} \right) \sum_{j=1}^{n} (f_j(s) - y_j) \mathbb{1}_{ik}(s) \mathbb{1}_{jk}(s) \langle \mathbf{x}_j^{\mathbf{v}_k(s)^\perp},\ \mathbf{x}_i \rangle.$$

Note that $\langle \mathbf{x}_j^{\mathbf{v}_k(s)^\perp},\ \mathbf{x}_i \rangle = \langle \mathbf{x}_j^{\mathbf{v}_k(s)^\perp},\ \mathbf{x}_i^{\mathbf{v}_k(s)^\perp} \rangle$ therefore,

$$b_i^I(s) = -\frac{\eta}{m} \sum_{k \notin S_i} \left(\frac{c_k g_k(s+1)}{\|\mathbf{v}_k(s+1)\|_2} \right) \left(\frac{c_k g_k(s)}{\|\mathbf{v}_k(s)\|_2} \right) \sum_{j=1}^{n} (f_j(s) - y_j) \mathbb{1}_{ik}(s) \mathbb{1}_{jk}(s) \langle \mathbf{x}_j^{\mathbf{v}_k(s)^\perp},\ \mathbf{x}_i^{\mathbf{v}_k(s)^\perp} \rangle.$$

Define $\tilde{\mathbf{V}}(s)$ as

$$\tilde{\mathbf{V}}_{ij}(s) = \frac{1}{m} \sum_{k=1}^{m} \left(\frac{\alpha c_k g_k(s+1)}{\|\mathbf{v}_k(s+1)\|_2} \right) \left(\frac{\alpha c_k g_k(s)}{\|\mathbf{v}_k(s)\|_2} \right) \mathbb{1}_{jk}(s) \mathbb{1}_{ik}(s) \langle \mathbf{x}_i^{\mathbf{v}_k(s)^\perp},\ \mathbf{x}_j^{\mathbf{v}_k(s)^\perp} \rangle.$$

This matrix is identical to $\mathbf{V}(s)$ except for a modified scaling term $\left(\frac{c_k^2 g_k(s+1) g_k(s)}{\|\mathbf{v}_k(s)\|_2 \|\mathbf{v}_k(s+1)\|_2}\right)$. We note however that

$$\min \left(\left(\frac{c_k g_k(s)}{\|\mathbf{v}_k(s)\|_2} \right)^2, \left(\frac{c_k g_k(s+1)}{\|\mathbf{v}_k(s+1)\|_2} \right)^2 \right) \leq \left(\frac{c_k g_k(s)}{\|\mathbf{v}_k(s)\|_2} \right) \left(\frac{c_k g_k(s+1)}{\|\mathbf{v}_k(s+1)\|_2} \right)$$
$$\leq \max \left(\left(\frac{c_k g_k(s)}{\|\mathbf{v}_k(s)\|_2} \right)^2, \left(\frac{c_k g_k(s+1)}{\|\mathbf{v}_k(s+1)\|_2} \right)^2 \right)$$

because $g_k(s), c_k^2$ are positive. Hence the matrix $\tilde{\mathbf{V}}(s)$ satisfies the hypothesis of Lemma B.4 entirely. We write $b_i^I(s)$ as

$$b_i^I(s) = -\eta/\alpha^2 \sum_{j=1}^{n} (f_j(s) - y_j)(\tilde{\mathbf{V}}_{ij}(s) - \tilde{\mathbf{V}}_{ij}^{\perp}(s)),$$

where we have defined

$$\tilde{\mathbf{V}}_{ij}^{\perp}(s) = \frac{1}{m} \sum_{k \in S_i} \left(\frac{\alpha c_k g_k(s)}{\|\mathbf{v}_k(s)\|_2} \right) \left(\frac{\alpha c_k g_k(s+1)}{\|\mathbf{v}_k(s+1)\|_2} \right) \mathbb{1}_{ik}(s) \mathbb{1}_{jk}(s) \langle \mathbf{x}_i^{\mathbf{v}_k(s)^\perp},\ \mathbf{x}_j^{\mathbf{v}_k(s)^\perp} \rangle. \tag{C.8}$$

We then bound the magnitude of each entry $\tilde{\mathbf{V}}_{ij}^{\perp}(s)$:

$$\tilde{\mathbf{V}}_{ij}^{\perp}(s) = \frac{1}{m} \sum_{k \in S_i} \left(\frac{\alpha c_k g_k(s)}{\|\mathbf{v}_k(s)\|_2} \right) \left(\frac{\alpha c_k g_k(s+1)}{\|\mathbf{v}_k(s+1)\|_2} \right) \mathbb{1}_{ik}(s) \mathbb{1}_{jk}(s) \langle \mathbf{x}_i^{\mathbf{v}_k(s)^\perp},\ \mathbf{x}_j^{\mathbf{v}_k(s)^\perp} \rangle$$
$$\leq \frac{(1 + R_g(m/\delta)^{1/d})^2 |S_i|}{m}. \tag{C.9}$$

Lastly we bound the size of the residual term $b_i^{II}(s)$,

$$|b_i^{II}(s)| = \left| -\frac{1}{\sqrt{m}} \sum_{k \in S_i} \frac{c_k g_k(s+1)}{\|\mathbf{v}_k(s+1)\|_2} \left(\sigma(\mathbf{v}_k(s+1)^\top \mathbf{x}_i) - \sigma(\mathbf{v}_k(s)^\top \mathbf{x}_i) \right) \right|$$
$$\leq \frac{g_k(s+1) \eta |S_i| \cdot \|\nabla_{\mathbf{v}_k} L(s)\|_2}{\sqrt{m} \|\mathbf{v}_k(s+1)\|_2}$$
$$\leq \frac{\eta |S_i| (1 + (m/\delta)^{1/d} R_g) \|\nabla_{\mathbf{v}_k} L(s)\|_2}{\alpha \sqrt{m}}.$$

Where we used the Lipschitz continuity of σ in the first bound, and took $R_g > 0$ that satisfies $|g_k(s+1) - g_k(0)| \leq R_g$ in the second inequality. Applying the bound (E.1),

$$|b_i^{II}(s)| \leq \frac{\eta |S_i| \sqrt{n}(1 + R_g(m/\delta)^{1/d})^2 \|\mathbf{f}(s) - \mathbf{y}\|_2}{\alpha^2 m}. \tag{C.10}$$

The sum $\mathbf{f}(s+1) - \mathbf{f}(s) = \mathbf{a}^I(s) + \mathbf{a}^{II}(s) + \mathbf{b}^I(s) + \mathbf{b}^{II}(s)$ is separated into the primary term $\eta \mathbf{p}(s) = \mathbf{a}_I(s) + \mathbf{b}_I(s)$ and the residual term $\eta \mathbf{r}(s) = \mathbf{a}_{II}(s) + \mathbf{b}_{II}(s)$ which is a result of the discretization. With this, the evolution matrix $\mathbf{\Lambda}(s)$ in (C.5) is re-defined as

$$\mathbf{\Lambda}(s) \coloneqq \mathbf{G}(s) + \frac{\tilde{\mathbf{V}}(s) - \tilde{\mathbf{V}}^{\perp}(s)}{\alpha^2}$$

and

$$\mathbf{f}(s+1) - \mathbf{f}(s) = -\eta \mathbf{\Lambda}(s)(\mathbf{f}(s) - \mathbf{y}) + \eta \mathbf{r}(s).$$

Now we compare $\|\mathbf{f}(s+1) - \mathbf{y}\|_2^2$ with $\|\mathbf{f}(s) - \mathbf{y}\|_2^2$,

$$\begin{aligned}
\|\mathbf{f}(s+1) - \mathbf{y}\|_2^2 =& \|\mathbf{f}(s+1) - \mathbf{f}(s) + \mathbf{f}(s) - \mathbf{y}\|_2^2 \\
=& \|\mathbf{f}(s) - \mathbf{y}\|_2^2 + 2\big\langle \mathbf{f}(s+1) - \mathbf{f}(s),\, \mathbf{f}(s) - \mathbf{y} \big\rangle \\
&+ \big\langle \mathbf{f}(s+1) - \mathbf{f}(s),\, \mathbf{f}(s+1) - \mathbf{f}(s) \big\rangle.
\end{aligned}$$

Substituting

$$\mathbf{f}(s+1) - \mathbf{f}(s) = \mathbf{a}^I(s) + \mathbf{b}^I(s) + \mathbf{a}^{II}(s) + \mathbf{b}^{II}(s) = -\eta \mathbf{\Lambda}(s)(\mathbf{f}(s) - \mathbf{y}) + \eta \mathbf{r}(s)$$

we obtain

$$\begin{aligned}
\|\mathbf{f}(s+1) - \mathbf{y}\|_2^2 =& \|\mathbf{f}(s) - \mathbf{y}\|_2^2 + 2(-\eta \mathbf{\Lambda}(s)(\mathbf{f}(s) - \mathbf{y}) + \eta \mathbf{r}(s))^\top (\mathbf{f}(s) - \mathbf{y}) \\
&+ \eta^2 (\mathbf{\Lambda}(s)(\mathbf{f}(s) - \mathbf{y}) - \mathbf{r}(s))^\top (\mathbf{\Lambda}(s)(\mathbf{f}(s) - \mathbf{y}) - \mathbf{r}(s)) \\
\leq& \|\mathbf{f}(s) - \mathbf{y}\|_2^2 + (\mathbf{f}(s) - \mathbf{y})^\top (-\eta \mathbf{\Lambda}(s) + \eta^2 \mathbf{\Lambda}^2(s))(\mathbf{f}(s) - \mathbf{y}) \\
&+ \eta \mathbf{r}(s)^\top (\mathbf{I} - \eta \mathbf{\Lambda}(s))(\mathbf{f}(s) - \mathbf{y}) + \eta^2 \|\mathbf{r}(s)\|_2^2.
\end{aligned}$$

Now as $\lambda_{\min}(\mathbf{\Lambda}(s)) \geq \omega/2$ and $\eta = \frac{1}{3\|\mathbf{\Lambda}(s)\|_2}$, we have that

$$(\mathbf{f}(s) - \mathbf{y})^\top (-\eta \mathbf{\Lambda}(s) + \eta^2 \mathbf{\Lambda}^2(s))(\mathbf{f}(s) - \mathbf{y}) = -\eta (\mathbf{f}(s) - \mathbf{y})^\top (\mathbf{I} - \eta \mathbf{\Lambda}(s)) \mathbf{\Lambda}(s)(\mathbf{f}(s) - \mathbf{y}) \leq -\frac{3\eta\omega}{8} \|\mathbf{f}(s) - \mathbf{y}\|_2^2.$$

Next we analyze the rest of the terms and group them as $\mathbf{q}(s)$,

$$\begin{aligned}
\mathbf{q}(s) &\coloneqq \eta \mathbf{r}(s)^\top (\mathbf{I} - \eta \mathbf{\Lambda}(s))(\mathbf{f}(s) - \mathbf{y}) + \eta^2 \|\mathbf{r}(s)\|_2^2 \\
&\leq \eta \|\mathbf{r}(s)\|_2 \|\mathbf{f}(s) - \mathbf{y}\|_2 + \eta^2 \|\mathbf{r}(s)\|_2^2.
\end{aligned}$$

By Property 1 we have

$$\mathbf{q}(s) \leq \eta c\omega \|\mathbf{f}(s) - \mathbf{y}\|_2^2 (1 + \eta c\omega) \leq 2c\eta\omega \|\mathbf{f}(s) - \mathbf{y}\|_2^2,$$

so that

$$\mathbf{q}(s) \leq c' \eta\omega \|\mathbf{f}(s) - \mathbf{y}\|_2^2,$$

for c' sufficiently small. Substituting, the difference $\mathbf{f}(s+1) - \mathbf{y}$ is bounded as

$$\begin{aligned}
\|\mathbf{f}(s+1) - \mathbf{y}\|_2^2 &\leq \|\mathbf{f}(s) - \mathbf{y}\|_2^2 - \eta\omega(1 - \eta \|\mathbf{\Lambda}(s)\|_2) \|\mathbf{f}(s) - \mathbf{y}\|_2^2 + c'\eta\omega \|\mathbf{f}(s) - \mathbf{y}\|_2^2 \\
&\leq (1 - \eta\omega(1 - \eta \|\mathbf{\Lambda}(s)\|_2) + c'\eta\omega) \|\mathbf{f}(s) - \mathbf{y}\|_2^2 \\
&\leq (1 - \eta\omega/2) \|\mathbf{f}(s) - \mathbf{y}\|_2^2,
\end{aligned}$$

for well chosen absolute constant c. Hence for each $s = 0, 1, \ldots, K$,

$$\|\mathbf{f}(s+1) - \mathbf{y}\|_2^2 \leq (1 - \eta\omega/2)\|\mathbf{f}(s) - \mathbf{y}\|_2^2,$$

so the prediction error converges linearly. □

In what comes next we prove the necessary conditions for Property 1, and define the appropriate ω for the $\mathbf{V}$ and $\mathbf{G}$ dominated regimes, in order to show $\lambda_{\min}(\mathbf{\Lambda}(s)) \geq \omega/2$.

Proof of Theorem 4.1:
To prove convergence we would like to apply Theorem C.1 with $\omega/2 = \frac{\lambda_0}{2\alpha^2}$. To do so we need to show that $m = \Omega\big(n^4 \log(n/\delta)/\lambda_0^4\big)$ guarantees that Property 1 holds and that $\lambda_{\min}(\mathbf{\Lambda}(s)) \geq \lambda_0/2\alpha^2$. For finite-step gradient training, take

$$R_v = \frac{\alpha\lambda_0}{192n(m/\delta)^{1/d}}, \quad R_g = \frac{\lambda_0}{96n(m/\delta)^{1/d}}. \tag{C.11}$$

Note the residual $\mathbf{r}(s)$ and the other terms $\mathbf{b}_I(s), \mathbf{b}_{II}(s)$ depend on the sets S_i that we define here using R_v. We make the assumption that $\|\mathbf{v}_k(s) - \mathbf{v}_k(0)\|_2 \leq R_v$ and $|g_k(s) - g_k(0)| \leq R_g$ for all k and that $s = 0, 1, \ldots K+1$, this guarantees that $\mathbf{b}_I(s)$ and $\mathbf{\Lambda}(s)$ are well defined. Applying Lemmas B.1, B.4 with R_v, R_g defined above, we have that $\lambda_{\min}(\tilde{\mathbf{V}}(s)) \geq \frac{5\lambda_0}{8}$. Then the least eigenvalue of the evolution matrix $\mathbf{\Lambda}(s)$ is bounded below

$$\begin{aligned}
\lambda_{\min}(\mathbf{\Lambda}(s)) &= \lambda_{\min}\left(\mathbf{G}(s) + \frac{\tilde{\mathbf{V}}(s) - \tilde{\mathbf{V}}^{\perp}(s)}{\alpha^2}\right) \\
&\geq \lambda_{\min}\left(\frac{\tilde{\mathbf{V}}(s) - \tilde{\mathbf{V}}^{\perp}(s)}{\alpha^2}\right) \\
&= \frac{\lambda_{\min}(\tilde{\mathbf{V}}(s) - \tilde{\mathbf{V}}^{\perp}(s))}{\alpha^2} \\
&\geq \frac{5\lambda_0}{8\alpha^2} - \frac{\|\tilde{\mathbf{V}}^{\perp}(s)\|_2}{\alpha^2}.
\end{aligned}$$

The first inequality holds since $\mathbf{G}(s) \succ 0$ and the last inequality is since $\lambda_{\min}(\tilde{\mathbf{V}}(s)) \geq \frac{5\lambda_0}{8}$.

To show $\lambda_{\min}(\mathbf{\Lambda}(s)) \geq \frac{\lambda_0}{2\alpha^2}$ we bound $\|\tilde{\mathbf{V}}^{\perp}(s)\|_2 \leq \frac{\lambda_0}{8}$. By (C.9), we have

$$|\tilde{\mathbf{V}}_{ij}^{\perp}(s)| \leq \frac{(1 + R_g(m/\delta)^{1/d})|S_i|}{m} \leq (1 + R_g(m/\delta)^{1/d})\left(\frac{\sqrt{2}\tilde{R}_v}{\sqrt{\pi}\alpha} + \frac{16\log(n/\delta)}{3m}\right).$$

Substituting R_v, R_g and m, a direct calculation shows that

$$|\tilde{\mathbf{V}}_{ij}^{\perp}(s)| \leq \frac{\lambda_0}{8n},$$

which yields

$$\|\tilde{\mathbf{V}}^{\perp}(s)\|_2 \leq \|\tilde{\mathbf{V}}^{\perp}(s)\|_F \leq \frac{\lambda_0}{8}.$$

Hence $\lambda_{\min}(\mathbf{\Lambda}(s)) \geq \frac{\lambda_0}{2\alpha^2}$ for iterations $s = 0, 1, \ldots K$.

We proceed by showing the residual $\mathbf{r}(s)$ satisfies property 1. Recall $\mathbf{r}(s)$ is written as

$$\mathbf{r}(s) = \frac{\mathbf{a}^{II}(s)}{\eta} + \frac{\mathbf{b}^{II}(s)}{\eta}.$$

and Property 1 states that $\|\mathbf{r}(s)\|_2 \leq \frac{c\eta\lambda_0}{\alpha^2}\|\mathbf{f}(s) - \mathbf{y}\|_2$ for sufficiently small absolute constant $c < 1$. This is equivalent to showing that both $\mathbf{a}^{II}(s)$, $\mathbf{b}^{II}(s)$ satisfy

$$\|\mathbf{a}^{II}(s)\|_2 \leq \frac{c\eta\lambda_0}{\alpha^2}\|\mathbf{f}(s) - \mathbf{y}\|_2, \tag{C.12}$$

$$\|\mathbf{b}^{II}(s)\|_2 \leq \frac{c\eta\lambda_0}{\alpha^2}\|\mathbf{f}(s) - \mathbf{y}\|_2. \tag{C.13}$$

We consider each term at turn. By (C.10),

$$\begin{aligned}\|\mathbf{b}^{II}(s)\|_2 &\leq \sqrt{n}\max_i b_i^{II}(s)\\ &\leq \max_i \frac{\eta n(1+R_g(m/\delta)^{1/d})^2|S_i|\|\mathbf{f}(s)-\mathbf{y}\|_2}{\alpha^2 m}\\ &\leq \frac{CmR_v\eta n\|\mathbf{f}(s)-\mathbf{y}\|_2}{\alpha^2 m}\\ &\leq \frac{\lambda_0\eta\|\mathbf{f}(s)-\mathbf{y}\|_2}{\alpha^2}\cdot nCR_v.\end{aligned}$$

In the above we used the values of R_v, R_g defined in (C.11) and applied Lemma C.1 in the third inequality. Taking $m=\Omega\big(n^4\log(n/\delta)/\lambda_0^4\big)$ with large enough constant yields

$$\|\mathbf{b}^{II}(s)\|_2 \leq \frac{c\lambda_0\eta\|\mathbf{f}(s)-\mathbf{y}\|_2}{\alpha^2}.$$

Next we analogously bound $\|\mathbf{a}^{II}(s)\|$ via the bound (C.7),

$$\begin{aligned}\|\mathbf{a}^{II}(s)\|_2 &\leq \sqrt{n}\max_i a_i^{II}(s)\\ &\leq \frac{\eta^2 n^{3/2}\big(1+R_g(m/\delta)^{1/d}\big)^3\|\mathbf{f}(s)-\mathbf{y}\|_2^2(m/\delta)^{1/d}}{\alpha^4\sqrt{m}}\\ &\leq \frac{\eta\lambda_0\|\mathbf{f}(s)-\mathbf{y}\|_2}{\alpha^2}\cdot\frac{\eta\big(1+R_g(m/\delta)^{1/d}\big)^3 n^{3/2}\|\mathbf{f}(s)-\mathbf{y}\|_2(m/\delta)^{1/d}}{\lambda_0\alpha^2\sqrt{m}}\\ &\leq \frac{\eta\lambda_0\|\mathbf{f}(s)-\mathbf{y}\|_2}{\alpha^2}\cdot\frac{\eta}{\alpha^2}\cdot\frac{Cn^2\sqrt{\log(n/\delta)}}{\lambda_0\sqrt{m}}\\ &\leq c\eta\omega\|\mathbf{f}(s)-\mathbf{y}\|_2.\end{aligned}$$

In the above we applied Lemma B.9 once again. The last inequality holds since $m=\Omega(n^4\log(n/\delta)/\lambda_0^4)$ and $\eta = O\Big(\frac{\alpha^2}{\|\mathbf{V}(s)\|_2}\Big)$, hence $\mathbf{r}(s)$ satisfies Property 1. Now since Theorem C.1 holds with $\omega=\lambda_0/\alpha^2$ we have that the maximum parameter trajectories are bounded as $\|\mathbf{v}_k(s)-\mathbf{v}_k(0)\|_2\leq R_v$ and $\|g_k(s)-g_k(0)\|\leq R_g$ for all k and every iteration $s=0,1,\dots,K+1$ via Lemmas C.2, C.3.

To finish the proof, we apply the same contradiction argument as in Theorems B.1, B.2, taking the first iteration $s=K_0$ where one of Lemmas C.2, C.3 does not hold. We note that $K_0>0$ and by the definition of K_0, for $s=0,1,\dots,K_0-1$ the Lemmas C.2, C.3 hold which implies that by the argument above we reach linear convergence in iteration $s=K_0$. This contradicts one of Lemmas C.2, C.3 which gives the desired contradiction, so we conclude that we have linear convergence with rate $\lambda_0/2\alpha^2$ for all iterations. □

Proof of Theorem 4.2:
For $\mathbf{G}$-dominated convergence, we follow the same steps as in the proof of Theorem 4.1. We redefine the trajectory constants for the finite step case

$$\tilde{R}_v \coloneqq \frac{\sqrt{2\pi}\alpha\mu_0}{64n(m/\delta)^{1/d}},\quad R_g \coloneqq \frac{\mu_0}{48n(m/\delta)^{1/d}}.$$

To use Theorem C.1 we need to show that $m=\Omega\big(n^4\log(n/\delta)/\alpha^4\mu_0^4\big)$ guarantees Property 1, and that $\lambda_{\min}(\mathbf{\Lambda}(s))\geq\mu_0/2$. We again note that the residual $\mathbf{r}(s)$ and $\mathbf{b}_I(s),\mathbf{b}_{II}(s)$ depend on the sets S_i that we define here using $\tilde{R}_v$ above as $S_i\coloneqq S_i(\tilde{R}_v)$.

We start by showing the property on the least eigenvalue. We make the assumption that we have linear convergence with $\omega/2=\mu_0/2$ and step-size η for iterations $s=0,\dots K$ so that Lemmas C.2, C.3 hold. Via an analogous analysis of the

continous case we reach that $m = \Omega\big(n^4 \log(n/\delta)/\mu_0^4\alpha^4\big)$ implies

$$\|\mathbf{v}_k(s) - \mathbf{v}_k(0)\|_2 \leq \frac{16\alpha\sqrt{n}\|\mathbf{f}(0) - \mathbf{y}\|_2}{\alpha\sqrt{m}\mu_0} \leq \tilde{R}_v, \quad |g_k(s) - g_k(0)| \leq \frac{8\sqrt{n}\|\mathbf{f}(0) - \mathbf{y}\|_2}{\sqrt{m}\mu_0} \leq R_g.$$

for $s = 0, \dots K+1$ by Lemmas C.2, C.3 and that $\mathbf{\Lambda}(s), \mathbf{b}_I(s)$ are well defined. Using the bounds on the parameter trajectories, Lemma B.5 and $\tilde{R}_v$ defined above yield $\lambda_{\min}(\mathbf{G}(s)) \geq \frac{5\mu_0}{8}$. The least eigenvalue of the evolution matrix $\mathbf{\Lambda}(s)$ is bounded below as

$$\begin{aligned}\lambda_{\min}(\mathbf{\Lambda}(s)) &= \lambda_{\min}\left(\mathbf{G}(s) + \frac{\tilde{\mathbf{V}}(s) - \tilde{\mathbf{V}}^{\perp}(s)}{\alpha^2}\right) \\ &\geq \lambda_{\min}(\mathbf{G}(s)) - \|\tilde{\mathbf{V}}^{\perp}(s)\|_2\end{aligned}$$

since $\tilde{\mathbf{V}}(s) \succ 0$ and $\alpha \geq 1$. We bound the spectral norm of $\|\tilde{\mathbf{V}}^{\perp}(s)\|_2$, for each entry i, j we have by (C.9) that

$$\begin{aligned}|\tilde{\mathbf{V}}_{ij}^{\perp}(s)| &\leq \frac{(1 + R_g(m/\delta)^{1/d})|S_i|}{m} \\ &\leq (1 + R_g(m/\delta)^{1/d})\left(\frac{\sqrt{2}\tilde{R}_v}{\sqrt{\pi}\alpha} + \frac{16\log(n/\delta)}{3m}\right) \\ &\leq \frac{8\tilde{R}_v}{\sqrt{2\pi}\alpha} \\ &\leq \frac{\mu_0}{8n}.\end{aligned}$$

where in the above inequalities we used our bounds on $\tilde{R}_v, R_g$ and m. Then the spectral norm is bounded as

$$\|\tilde{\mathbf{V}}^{\perp}(s)\|_2 \leq \|\tilde{\mathbf{V}}^{\perp}(s)\|_F \leq \mu_0/8.$$

Hence we have that $\lambda_{\min}(\mathbf{\Lambda}(s)) \geq \mu_0/2$ for $s = 0, 1, \dots K$.

Next we show the residual $\mathbf{r}(s)$ satisfies Property 1. Recall $\mathbf{r}(s)$ is written as

$$\mathbf{r}(s) = \frac{\mathbf{a}^{II}(s)}{\eta} + \frac{\mathbf{b}^{II}(s)}{\eta}.$$

Property 1 states the condition $\|\mathbf{r}(s)\|_2 \leq c\omega\eta\|\mathbf{f}(s) - \mathbf{y}\|_2$ for sufficiently small $c < 1$ with $\omega = \mu_0$. This is equivalent to showing that both $\mathbf{a}^{II}(s)$, $\mathbf{b}^{II}(s)$ satisfy that

$$\|\mathbf{a}^{II}(s)\|_2 \leq c\eta\mu_0\|\mathbf{f}(s) - \mathbf{y}\|_2, \tag{C.14}$$

$$\|\mathbf{b}^{II}(s)\|_2 \leq c\eta\mu_0\|\mathbf{f}(s) - \mathbf{y}\|_2, \tag{C.15}$$

for sufficiently small absolute constant c. For $\mathbf{b}_{II}(s)$ we have that (C.10) gives

$$\begin{aligned}\|\mathbf{b}^{II}(s)\|_2 &\leq \sqrt{n}\max_i b_i^{II}(s) \\ &\leq \max_i \frac{\eta(1 + R_g(m/\delta)^{1/d})^2|S_i|n\|\mathbf{f}(s) - \mathbf{y}\|_2}{\alpha^2 m}.\end{aligned}$$

Next applying Lemmas C.1 and B.9 in turn yields

$$\begin{aligned}&\leq \frac{Cm\tilde{R}_v\eta n\|\mathbf{f}(s) - \mathbf{y}\|_2}{\alpha^2 m} \\ &\leq \eta\mu_0\|\mathbf{f}(s) - \mathbf{y}\|_2\frac{\tilde{R}_v}{n\alpha^2}.\end{aligned}$$

Substituting $m = \Omega\big(n^4\log(n/\delta)/\mu_0^4\alpha^4\big)$ for a large enough constant and R_v we get

$$\|\mathbf{b}^{II}(s)\|_2 \leq c\eta\mu_0\|\mathbf{f}(s) - \mathbf{y}\|_2.$$

Analogously we bound $\|\mathbf{a}^{II}(s)\|_2$ using (C.7),

$$\begin{aligned}
\|\mathbf{a}^{II}(s)\|_2 &\leq \sqrt{n}\max_i a_i^{II}(s) \\
&\leq \frac{\eta^2 n^{3/2}\big(1+R_g(m/\delta)^{1/d}\big)^3\|\mathbf{f}(s)-\mathbf{y}\|_2^2(m/\delta)^{1/d}}{\alpha^4\sqrt{m}} \\
&\leq \eta\mu_0\|\mathbf{f}(s)-\mathbf{y}\|_2\cdot\frac{\eta\big(1+R_g(m/\delta)^{1/d}\big)^3 n^{3/2}\|\mathbf{f}(s)-\mathbf{y}\|_2(m/\delta)^{1/d}}{\mu_0\alpha^4\sqrt{m}} \\
&\leq \eta\mu_0\|\mathbf{f}(s)-\mathbf{y}\|_2\cdot\frac{\eta}{\alpha^2}\cdot\frac{Cn^2\sqrt{\log(n/\delta)}}{\alpha^2\mu_0^2\sqrt{m}} \\
&\leq c\eta\mu_0\|\mathbf{f}(s)-\mathbf{y}\|_2.
\end{aligned}$$

Where we have used Lemma B.9 in the third inequality and substituted $m = \Omega(n^4\log(n/\delta)/\alpha^4\mu_0^4)$ noting that $\eta = O\big(\frac{1}{\|\mathbf{\Lambda}(s)\|_2}\big)$ and that $\alpha \geq 1$ in the last inequality. Therefore we have that $\mathbf{r}(s)$ satisfies Property 1 so that Theorem C.1 holds. By the same contradiction argument as in Theorem 4.1 we have that this holds for all iterations. □

D. Additional Technical Lemmas and Proofs of the Lemmas from Appendix B

Proof of Lemma 4.1:
We prove Lemma 4.1 for $\mathbf{V}^\infty$, $\mathbf{G}^\infty$ separately. $\mathbf{V}^\infty$ can be viewed as the covariance matrix of the functionals ϕ_i defined as

$$\phi_i(\mathbf{v}) = \mathbf{x}_i\left(\mathbf{I}-\frac{\mathbf{v}\mathbf{v}^\top}{\|\mathbf{v}\|_2^2}\right)\mathbb{1}\{\mathbf{v}^\top\mathbf{x}_i\geq 0\} \tag{D.1}$$

over the Hilbert space $\mathcal{V}$ of $L^2(N(0,\alpha^2\mathbf{I}))$ of functionals. Under this formulation, if $\phi_1,\phi_2,\ldots,\phi_n$ are linearly independent, then $\mathbf{V}^\infty$ is strictly positive definite. Thus, to show that $\mathbf{V}^\infty$ is strictly positive definite is equivalent to showing that

$$c_1\phi_1 + c_2\phi_2 + \cdots + c_n\phi_n = 0 \ \text{ in } \mathcal{V} \tag{D.2}$$

implies $c_i = 0$ for each i. The ϕ_is are piece-wise continuous functionals, and equality in $\mathcal{V}$ is equivalent to

$$c_1\phi_1 + c_2\phi_2 + \cdots + c_n\phi_n = 0 \ \text{ almost everywhere.}$$

For the sake of contradiction, assume that there exist $c_1,\ldots,c_n$ that are not identically 0, satisfying (D.2). As c_i are not identically 0, there exists an i such that $c_i \neq 0$. We show this leads to a contradiction by constructing a non-zero measure region such that the linear combination $\sum_i c_i\phi_i$ is non-zero.

Denote the orthogonal subspace to $\mathbf{x}_i$ as $D_i \coloneqq \{\mathbf{v}\in\mathbb{R}^d : \mathbf{v}^\top\mathbf{x}_i = 0\}$. By Assumption 1,

$$D_i \not\subseteq \bigcup_{j\neq i} D_j$$

This holds since D_i is a $(d-1)$-dimensional space which may not be written as the finite union of sub-spaces $D_i\cap D_j$ of dimension $d-2$ (since $\mathbf{x}_i$ and $\mathbf{x}_j$ are not parallel). Thus, take $\mathbf{z}\in D_i\backslash\bigcup_{j\neq i}D_j$. Since $\bigcup_{j\neq i}D_j$ is closed in $\mathbb{R}^d$, there exists an $R>0$ such that

$$\mathcal{B}(\mathbf{z},4R)\cap\bigcup_{j\neq i}D_j = \emptyset.$$

Next take $\mathbf{y}\in\partial\mathcal{B}(\mathbf{z},3R)\cap D_i$ (where ∂ denotes the boundary) on the smaller disk of radius $3R$ so that it satisfies $\|\mathbf{y}\|_2 = \max_{\mathbf{y}'\in\partial\mathcal{B}(\mathbf{z},3R)\cap D_i}\|\mathbf{y}'\|_2$. Now for any $r\leq R$, the ball $\mathcal{B}(\mathbf{y},r)$ is such that for all points $\mathbf{v}\in\mathcal{B}(\mathbf{y},r)$ we have $\|\mathbf{v}^{\mathbf{x}_i^\perp}\|_2 \geq 2R$ and $\|\mathbf{v}^{\mathbf{x}_i}\|_2\leq R$. Then for any $r\leq R$, the points $\mathbf{v}\in\mathcal{B}(\mathbf{y},r)\subset\mathcal{B}(\mathbf{z},4R)$ satisfy that

$$\|\mathbf{x}_i^{\mathbf{v}^\perp}\|_2 \geq \|\mathbf{x}_i\|_2 - \frac{\mathbf{x}_i\cdot\mathbf{v}}{\|\mathbf{v}\|_2} \geq \|\mathbf{x}_i\|_2\left(1-\frac{R}{2R}\right) \geq \frac{\|\mathbf{x}_i\|_2}{2}.$$

Next we decompose the chosen ball $\mathcal{B}(\mathbf{y}, r) = B^+(r) \vee B^-(r)$ to the areas where the ReLU function at the point $\mathbf{x}_i$ is active and inactive

$$B^+(r) = \mathcal{B}(\mathbf{y}, r) \cap \{\mathbf{x}_i^\top \mathbf{v} \geq 0\}, \quad B^-(r) = \mathcal{B}(\mathbf{y}, r) \cap \{\mathbf{x}_i^\top \mathbf{v} < 0\}.$$

Note that ϕ_i has a discontinuity on D_i and is continuous within each region $B^+(r)$ and $B^-(r)$. Moreover, for $j \neq i$, ϕ_j is continuous on the entire region of $\mathcal{B}(\mathbf{y}, r)$ since $\mathcal{B}(\mathbf{y}, r) \subset \mathcal{B}(\mathbf{z}, 4R) \subset D_j^c$. Since we have that ϕ_j is continuous in the region, the Lebesgue differentiation theorem implies that for $r \to 0$, ϕ_i satisfies on $B^+(r), B^-(r)$:

$$\lim_{r\to 0} \frac{1}{\mu(B^+(r))} \int_{B^+(r)} \phi_i = \mathbf{x}_i^{\mathbf{y}^\perp} \neq 0, \quad \lim_{r\to 0} \frac{1}{\mu(B^-(r))} \int_{B^-(r)} \phi_i = 0.$$

For $j \neq i$ ϕ_j is continuous on the entire ball $\mathcal{B}(\mathbf{y}, r)$ hence the Lebesgue differentiation theorem also gives

$$\lim_{r\to 0} \frac{1}{\mu(B^+(r))} \int_{B^+(r)} \phi_i = \phi_j(\mathbf{y}), \quad \lim_{r\to 0} \frac{1}{\mu(B^-(r))} \int_{B^-(r)} \phi_i = \phi_j(\mathbf{y}).$$

We integrate $c_1\phi_1 + \dots c_n\phi_n$ over $B^-(r)$ and $B^+(r)$ separately and subtract the integrals. By the assumption, $c_1\phi_1 + \cdots + c_n\phi_n = 0$ almost everywhere so each integral evaluates to 0 and the difference is also 0,

$$0 = \frac{1}{\mu(B^+(r))} \int_{B^+(r)} c_1\phi_1 + \cdots + c_n\phi_n - \frac{1}{\mu(B^-(r))} \int_{B^-(r)} c_1\phi_1 + \cdots + c_n\phi_n. \tag{D.3}$$

By the continuity of ϕ_j for $j \neq i$ taking $r \to 0$ we have that

$$\frac{1}{\mu(B^+(r))} \lim_{r\to 0} \int_{B^+(r)} \phi_j - \frac{1}{\mu(B^-(r))} \int_{B^-(r)} \phi_j = \phi_j(\mathbf{y}) - \phi_j(\mathbf{y}) = 0.$$

For ϕ_i the functionals evaluate differently. For $B^-(r)$ we have that

$$\frac{1}{\mu(B^-(r))} \lim_{r\to 0} \int_{B^-(r)} \phi_i = \frac{1}{\mu(B^-(r))} \lim_{r\to 0} \int_{B^-(r)} 0 = 0,$$

while the integral over the positive side, $B^+(r)$ is equal to

$$\frac{1}{\mu(B^+(r))} \int_{B^+(r)} \phi_i(\mathbf{z}) d\mathbf{z} = \frac{1}{\mu(B^+(r))} \int_{B^+(r)} \mathbf{x}_i^{\mathbf{z}^\perp} d\mathbf{z} = \mathbf{x}_i^{\mathbf{y}^\perp}.$$

By construction, $\|\mathbf{x}_i^{\mathbf{y}^\perp}\|_2 > R$ and is non-zero so we conclude that for (D.3) to hold we must have $c_i = 0$. This gives the desired contradiction and implies that $\phi_1, \dots \phi_n$ are independent and $\mathbf{V}^\infty$ is positive definite with $\lambda_{\min}(\mathbf{V}^\infty) = \lambda_0$.

Next we consider $\mathbf{G}^\infty$ and again frame the problem in the context of the covariance matrix of functionals. Define

$$\theta_i(\mathbf{v}) \coloneqq \sigma\left(\frac{\mathbf{v}^\top \mathbf{x}_i}{\|\mathbf{v}\|_2}\right)$$

for $\mathbf{v} \neq 0$.

Now the statement of the theorem is equivalent to showing that the covariance matrix of $\{\theta_i\}$ does not have zero-eigenvalues, that is, the functionals θ_is are linearly independent. For the sake of contradiction assume $\exists\, c_1, \dots, c_n$ such that

$$c_1\theta_1 + c_2\theta_2 + \cdots + c_n\theta_n = 0 \ \text{ in } \mathcal{V} \ \text{ (equivalent to a.e).}$$

Via the same contradiction argument we show that $c_i = 0$ for all i. Unlike ϕ_i defined in (D.1), each θ_i is continuous and non-negative so equality "a.e" is strengthened to "for all $\mathbf{v}$",

$$c_1\theta_1 + c_2\theta_2 + \cdots + c_n\theta_n = 0.$$

Equality everywhere requires that the derivatives of the function are equal to 0 almost everywhere. Computing derivatives with respect to $\mathbf{v}$ yields

$$c_1 \mathbf{x}_1^{\mathbf{v}^\perp} \mathbb{1}\{\mathbf{v}^\top \mathbf{x}_1 \geq 0\} + c_2 \mathbf{x}_2^{\mathbf{v}^\perp} \mathbb{1}\{\mathbf{v}^\top \mathbf{x}_2 \geq 0\} + \cdots + c_n \mathbf{x}_n^{\mathbf{v}^\perp} \mathbb{1}\{\mathbf{v}^\top \mathbf{x}_n \geq 0\} = 0.$$

Which coincide with

$$c_1 \phi_1 + \cdots + c_n \phi_n$$

By the first part of the proof, the linear combination $c_1\phi_1 + \cdots + c_n\phi_n$ is non-zero around a ball of positive measure unless $c_i = 0$ for all i. This contradicts the assumption that the derivative is 0 almost everywhere; therefore $\mathbf{G}^\infty$ is strictly positive definite with $\lambda_{\min}(G^\infty) =: \mu_0 > 0$. □

We briefly derive an inequality for the sum of indicator functions for events that are bounded by the sum of indicator functions of *independent* events. This enables us to develop more refined concentration than in Du et al. (2019b) for monitoring the orthogonal and aligned Gram matrices during training.

Lemma D.1. *Let $A_1, \ldots, A_m$ be a sequence of events and suppose that $A_k \subseteq B_k$ with $B_1, \ldots, B_m$ mutually independent. Further assume that for each k, $\mathbb{P}(B_k) \leq p$, and define $S = \frac{1}{m}\sum_{k=1}^m \mathbb{1}_{A_k}$. Then with probability $1-\delta$, S satisfies*

$$S \leq p\left(2 + \frac{8\log(1/\delta)}{3mp}\right).$$

Proof of Lemma D.1:
Bound S as

$$S = \frac{1}{m}\sum_{k=1}^m \mathbb{1}_{A_k} \leq \frac{1}{m}\sum_{k=1}^m \mathbb{1}_{B_k}.$$

We apply Bernstein's concentration inequality to reach the bound. Denote $X_k = \frac{\mathbb{1}_{B_k}}{m}$ and $\tilde{S} = \sum_{k=1}^m X_k$. Then

$$\mathrm{Var}(X_k) \leq \mathbb{E}X_k^2 = (1/m)^2 \mathbb{P}(X_k) + 0 \leq \frac{p}{m^2}, \quad \mathbb{E}\tilde{S} = \mathbb{E}\sum_{k=1}^m X_k \leq p.$$

Applying Bernstein's inequality yields

$$\mathbb{P}(\tilde{S} - \mathbb{E}\tilde{S} \geq t) \leq \exp\left(\frac{-t^2/2}{\sum_{k=1}^m \mathbb{E}X_k^2 + \frac{t}{3m}}\right).$$

Fix δ and take the smallest t such that $\mathbb{P}(\tilde{S} - \mathbb{E}\tilde{S} \geq t) \leq \delta$. Denote $t = r \cdot \mathbb{E}\tilde{S}$, either $\mathbb{P}(\tilde{S} - \mathbb{E}\tilde{S} \geq \mathbb{E}\tilde{S}) \leq \delta$, or $t = r\mathbb{E}\tilde{S}$ corresponds to $r \geq 1$. Note that $t = r\mathbb{E}\tilde{S} \leq rp$. In the latter case, the bound is written as

$$\mathbb{P}(\tilde{S} - \mathbb{E}\tilde{S} \geq rp) \leq \exp\left(\frac{-(pr)^2/2}{p/m + \frac{pr}{3m}}\right) \leq \exp\left(\frac{-(pr)^2/2}{\frac{p}{m}(1+\frac{r}{3})}\right) \leq \exp\left(\frac{-(pr)^2/2}{\frac{p}{m}(\frac{4r}{3})}\right) = \exp\left(\frac{-3prm}{8}\right).$$

Solving for δ gives

$$rp \leq \frac{8\log(1/\delta)}{3m}.$$

Hence with probability $1-\delta$,

$$S \leq \tilde{S} \leq \max\left\{p\left(1 + \frac{8\log(1/\delta)}{3mp}\right), 2p\right\} \leq p\left(2 + \frac{8\log(1/\delta)}{3mp}\right).$$

□

Proof of Lemma B.1:
We prove the claim by applying concentration on each entry of the difference matrix. Each entry $\mathbf{V}_{ij}(0)$ is written as

$$\mathbf{V}_{ij}(0) = \frac{1}{m}\sum_{k=1}^{m} \langle \mathbf{x}_i^{\mathbf{v}_k(0)^\perp},\, \mathbf{x}_j^{\mathbf{v}_k(0)^\perp} \rangle \left(\frac{\alpha c_k \cdot g_k}{\|\mathbf{v}_k\|_2}\right)^2 \mathbb{1}_{ik}(0)\mathbb{1}_{jk}(0).$$

At initialization $g_k(0) = \|\mathbf{v}_k(0)\|_2/\alpha$, $c_k^2 = 1$ so $\mathbf{V}_{ij}(0)$ simplifies to

$$\mathbf{V}_{ij}(0) = \frac{1}{m}\sum_{k=1}^{m} \langle \mathbf{x}_i^{\mathbf{v}_k(0)^\perp},\, \mathbf{x}_j^{\mathbf{v}_k(0)^\perp} \rangle \mathbb{1}_{ik}(0)\mathbb{1}_{jk}(0).$$

Since the weights $\mathbf{v}_k(0)$ are initialized independently for each entry we have $\mathbb{E}_{\mathbf{v}}\mathbf{V}_{ij}(0) = \mathbf{V}_{ij}^{\infty}$. We measure the deviation $\mathbf{V}(0) - \mathbf{V}^{\infty}$ via concentration. Each term in the sum $\frac{1}{m}\sum_{j=1}^{m} \langle \mathbf{x}_i^{\mathbf{v}_k(0)^\perp},\, \mathbf{x}_j^{\mathbf{v}_k(0)^\perp} \rangle \mathbb{1}_{ik}(0)\mathbb{1}_{jk}(0)$ is independent and bounded,

$$-1 \le \langle \mathbf{x}_i^{\mathbf{v}_k(0)^\perp},\, \mathbf{x}_j^{\mathbf{v}_k(0)^\perp} \rangle \mathbb{1}_{ik}(0)\mathbb{1}_{jk}(0) \le 1.$$

Applying Hoeffding's inequality to each entry yields that with probability $1 - \delta/n^2$, for all i, j,

$$|\mathbf{V}_{ij}(0) - \mathbf{V}_{ij}^{\infty}| \le \frac{2\sqrt{\log(n^2/\delta)}}{\sqrt{m}}.$$

Taking a union bound over all entries, with probability $1 - \delta$,

$$|\mathbf{V}_{ij}(0) - \mathbf{V}_{ij}^{\infty}| \le \frac{4\sqrt{\log(n/\delta)}}{\sqrt{m}}.$$

Bounding the spectral norm, with probability $1 - \delta$,

$$\|\mathbf{V}(0) - \mathbf{V}^{\infty}\|_2^2 \le \|\mathbf{V}(0) - \mathbf{V}^{\infty}\|_F^2 \le \sum_{i,j} |\mathbf{V}_{ij}(0) - \mathbf{V}_{ij}^{\infty}|^2$$
$$\le \frac{16n^2 \log(n/\delta)}{m}.$$

Taking $m = \Omega\big(\frac{n^2 \log(n/\delta)}{\lambda_0^2}\big)$ therefore guarantees

$$\|\mathbf{V}(0) - \mathbf{V}^{\infty}\|_2 \le \frac{\lambda_0}{4}.$$

□

Proof of Lemma B.2:
This is completely analogous to B.1. Recall $\mathbf{G}(0)$ is defined as,

$$\mathbf{G}_{ij}(0) = \frac{1}{m}\sum_{k=1}^{m} \langle \mathbf{x}_i^{\mathbf{v}_k(0)},\, \mathbf{x}_j^{\mathbf{v}_k(0)} \rangle c_k^2 \mathbb{1}_{ik}(0)\mathbb{1}_{jk}(0)$$

with $c_k^2 = 1$ and $\mathbf{v}_k(0) \sim N(0, \alpha^2\mathbf{I})$ are initialized i.i.d. Since each term is bounded like B.1. The same analysis gives

$$\|\mathbf{G}_{ij}(0) - \mathbf{G}_{ij}^{\infty}\|_2^2 \le \frac{16n^2 \log(n/\delta)}{m}.$$

Taking $m = \Omega\Big(\frac{n^2 \log(n/\delta)}{\mu_0^2}\Big)$ therefore guarantees,

$$\|\mathbf{G}(0) - \mathbf{G}^{\infty}\|_2 \leq \frac{\mu_0}{4}.$$

□

Proof of Lemma B.3:
For a given R, define the event of a possible sign change of neuron k at point $\mathbf{x}_i$ as

$$A_{i,k}(R) = \{\exists \mathbf{v} : \|\mathbf{v} - \mathbf{v}_k(0)\|_2 \leq R, \text{ and } \mathbb{1}\{\mathbf{v}_k(0)^\top \mathbf{x}_i \geq 0\} \neq \mathbb{1}\{\mathbf{v}^\top \mathbf{x}_i \geq 0\}\}$$

$A_{i,k}(R)$ occurs exactly when $|\mathbf{v}_k(0)^\top \mathbf{x}_i| \leq R$, since $\|\mathbf{x}_i\|_2 = 1$ and the perturbation may be taken in the direction of $-\mathbf{x}_i$. To bound the probability $A_{i,k}(R)$ we consider the probability of the event

$$\mathbb{P}(A_{i,k}(R)) = \mathbb{P}(|\mathbf{v}_k(0)^\top \mathbf{x}_i| < R) = \mathbb{P}(|z| < R).$$

Here, $z \sim N(0, \alpha^2)$ since the product $\mathbf{v}_k(0)^\top \mathbf{x}_i$ follows a centered normal distribution. The norm of $\|\mathbf{x}_i\|_2 = 1$ which implies that z computes to a standard deviation α. Via estimates on the normal distribution, the probability on the event is bounded like

$$\mathbb{P}(A_{i,k}(R)) \leq \frac{2R}{\alpha\sqrt{2\pi}}.$$

We use the estimate for $\mathbb{P}(A_{i,k}(R))$ to bound the difference between the surrogate Gram matrix and the Gram matrix at initialization $\mathbf{V}(0)$.
Recall the surrogate $\hat{\mathbf{V}}(t)$ is defined as

$$\hat{\mathbf{V}}_{ij}(t) = \frac{1}{m}\sum_{k=1}^{m} \langle \mathbf{x}_i^{\mathbf{v}_k(t)^\perp}, \mathbf{x}_k^{\mathbf{v}_k(t)^\perp} \rangle \mathbb{1}_{ik}(t)\mathbb{1}_{jk}(t).$$

Thus for entry i, j we have

$$|\hat{\mathbf{V}}_{ij}(t) - \mathbf{V}_{ij}(0)| = \left| \frac{1}{m}\sum_{k=1}^{m} \langle \mathbf{x}_i^{\mathbf{v}_k(t)^\perp}, \mathbf{x}_j^{\mathbf{v}_k(t)^\perp} \rangle \mathbb{1}_{ik}(t)\mathbb{1}_{jk}(t) - \langle \mathbf{x}_i^{\mathbf{v}_k(0)^\perp}, \mathbf{x}_j^{\mathbf{v}_k(0)^\perp} \rangle \mathbb{1}_{ik}(0)\mathbb{1}_{jk}(0) \right|$$

This sum is decomposed into the difference between the inner product and the difference in the rectifier patterns terms respectively:

$$\Big(\langle \mathbf{x}_i^{\mathbf{v}_k(t)^\perp}, \mathbf{x}_j^{\mathbf{v}_k(t)^\perp} \rangle - \langle \mathbf{x}_i^{\mathbf{v}_k(0)^\perp}, \mathbf{x}_j^{\mathbf{v}_k(0)^\perp} \rangle \Big), \qquad \Big(\mathbb{1}_{ik}(t)\mathbb{1}_{jk}(t) - \mathbb{1}_{ik}(0)\mathbb{1}_{jk}(0) \Big).$$

Define

$$Y_{ij}^k = \Big(\langle \mathbf{x}_i^{\mathbf{v}_k(t)^\perp}, \mathbf{x}_j^{\mathbf{v}_k(t)^\perp} \rangle - \langle \mathbf{x}_i^{\mathbf{v}_k(0)^\perp}, \mathbf{x}_j^{\mathbf{v}_k(0)^\perp} \rangle \Big) \big(\mathbb{1}_{ik}(t)\mathbb{1}_{jk}(t) \big),$$
$$Z_{ij}^k = \Big(\langle \mathbf{x}_i^{\mathbf{v}_k(0)^\perp}, \mathbf{x}_j^{\mathbf{v}_k(0)^\perp} \rangle \Big) \Big(\mathbb{1}_{ik}(t)\mathbb{1}_{jk}(t) - \mathbb{1}_{ik}(0)\mathbb{1}_{jk}(0) \Big).$$

Then

$$|\hat{\mathbf{V}}_{ij}(t) - \mathbf{V}_{ij}(0)| = \left| \frac{1}{m}\sum_{k=1}^{m} Y_{ij}^k + Z_{ij}^k \right| \leq \left| \frac{1}{m}\sum_{k=1}^{m} Y_{ij}^k \right| + \left| \frac{1}{m}\sum_{k=1}^{m} Z_{ij}^k \right|.$$

To bound $|\frac{1}{m}\sum_{k=1}^{m} Y_{ij}^k|$ we bound each $|Y_{ij}^k|$ as follows.

$$
\begin{aligned}
|Y_{ij}^k| &= \left| \left(\langle \mathbf{x}_i^{\mathbf{v}_k(t)^\perp}, \mathbf{x}_j^{\mathbf{v}_k(t)^\perp} \rangle - \langle \mathbf{x}_i^{\mathbf{v}_k(0)^\perp}, \mathbf{x}_j^{\mathbf{v}_k(0)^\perp} \rangle \right) (\mathbb{1}_{ik}(t)\mathbb{1}_{jk}(t)) \right| \\
&\leq \left| \langle \mathbf{x}_i^{\mathbf{v}_k(t)^\perp}, \mathbf{x}_j^{\mathbf{v}_k(t)^\perp} \rangle - \langle \mathbf{x}_i^{\mathbf{v}_k(0)^\perp}, \mathbf{x}_j^{\mathbf{v}_k(0)^\perp} \rangle \right| \\
&= \left| \langle \mathbf{x}_i, \mathbf{x}_j \rangle - \langle \mathbf{x}_i^{\mathbf{v}_k(t)}, \mathbf{x}_j^{\mathbf{v}_k(t)} \rangle + \langle \mathbf{x}_i^{\mathbf{v}_k(0)}, \mathbf{x}_j^{\mathbf{v}_k(0)} \rangle - \langle \mathbf{x}_i, \mathbf{x}_j \rangle \right| \\
&= \left| \left\langle \frac{\mathbf{x}_i^\top \mathbf{v}_k(t)}{\|\mathbf{v}_k(t)\|_2} \cdot \frac{\mathbf{v}_k(t)}{\|\mathbf{v}_k(t)\|_2}, \frac{\mathbf{x}_j^\top \mathbf{v}_k(t)}{\|\mathbf{v}_k(t)\|_2} \cdot \frac{\mathbf{v}_k(t)}{\|\mathbf{v}_k(t)\|_2} \right\rangle - \langle \mathbf{x}_i^{\mathbf{v}_k(0)}, \mathbf{x}_j^{\mathbf{v}_k(0)} \rangle \right| \\
&= \left| \frac{\mathbf{x}_i^\top \mathbf{v}_k(t)}{\|\mathbf{v}_k(t)\|_2} \cdot \frac{\mathbf{x}_j^\top \mathbf{v}_k(t)}{\|\mathbf{v}_k(t)\|_2} - \langle \mathbf{x}_i^{\mathbf{v}_k(0)}, \mathbf{x}_j^{\mathbf{v}_k(0)} \rangle \right| \\
&= \left| \frac{\mathbf{x}_i^\top \mathbf{v}_k(0)}{\|\mathbf{v}_k(0)\|_2} \cdot \frac{\mathbf{x}_j^\top \mathbf{v}_k(0)}{\|\mathbf{v}_k(0)\|_2} + \mathbf{x}_i^\top \left(\frac{\mathbf{v}_k(t)}{\|\mathbf{v}_k(t)\|_2} - \frac{\mathbf{v}_k(0)}{\|\mathbf{v}_k(0)\|_2} \right) \cdot \frac{\mathbf{x}_j^\top \mathbf{v}_k(t)}{\|\mathbf{v}_k(t)\|_2} \right. \\
&\qquad \left. + \mathbf{x}_j^\top \left(\frac{\mathbf{v}_k(t)}{\|\mathbf{v}_k(t)\|_2} - \frac{\mathbf{v}_k(0)}{\|\mathbf{v}_k(0)\|_2} \right) \cdot \frac{\mathbf{x}_i^\top \mathbf{v}_k(0)}{\|\mathbf{v}_k(0)\|_2} - \langle \mathbf{x}_i^{\mathbf{v}_k(0)}, \mathbf{x}_j^{\mathbf{v}_k(0)} \rangle \right| \\
&\leq \left| \mathbf{x}_i^\top \left(\frac{\mathbf{v}_k(t)}{\|\mathbf{v}_k(t)\|_2} - \frac{\mathbf{v}_k(0)}{\|\mathbf{v}_k(0)\|_2} \right) \cdot \frac{\mathbf{x}_j^\top \mathbf{v}_k(t)}{\|\mathbf{v}_k(t)\|_2} \right| + \left| \mathbf{x}_i^\top \left(\frac{\mathbf{v}_k(t)}{\|\mathbf{v}_k(t)\|_2} - \frac{\mathbf{v}_k(0)}{\|\mathbf{v}_k(0)\|_2} \right) \cdot \frac{\mathbf{x}_j^\top \mathbf{v}_k(t)}{\|\mathbf{v}_k(t)\|_2} \right| \\
&\leq 2 \left\| \frac{\mathbf{v}_k(t)}{\|\mathbf{v}_k(t)\|_2} - \frac{\mathbf{v}_k(0)}{\|\mathbf{v}_k(0)\|_2} \right\|_2 .
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
\left| \frac{1}{m} \sum_{k=1}^m Y_{ij}^k \right| &\leq \frac{2}{m} \sum_{k=1}^m \left\| \frac{\mathbf{v}_k(t)}{\|\mathbf{v}_k(t)\|_2} - \frac{\mathbf{v}_k(0)}{\|\mathbf{v}_k(0)\|_2} \right\|_2 \\
&\leq \frac{4R_{\mathbf{v}}(2m/\delta)^{1/d}}{\alpha} \\
&\leq \frac{8R_{\mathbf{v}}(m/\delta)^{1/d}}{\alpha},
\end{aligned}
$$

where the first inequality follows from Lemma B.10. Note that the inequality holds with high probability $1 - \delta/2$ for all i, j. For the second sum, $|\frac{1}{m}\sum_{k=1}^m Z_{ij}^k| \leq \frac{1}{m}\sum_{k=1}^m \mathbb{1}_{A_{ik}(R)} + \frac{1}{m}\sum_{k=1}^m \mathbb{1}_{A_{jk}(R)}$ so we apply Lemma D.1 to get, with probability $1 - \delta/2n^2$

$$
\begin{aligned}
\left| \frac{1}{m} \sum_{k=1}^m Z_{ij}^k \right| &\leq \frac{2R_v}{\alpha\sqrt{2\pi}} \left(2 + \frac{2\sqrt{2\pi}\alpha \log(2n^2/\delta)}{3mR_v} \right) \\
&\leq \frac{8R_v}{\alpha\sqrt{2\pi}},
\end{aligned}
$$

since m satisfies $m = \Omega\big(\frac{(m/\delta)^{1/d} n^2 \log(n/\delta)}{\lambda_0}\big)$. Combining the two sums for Y_{ij}^k and Z_{ij}^k, with probability $1 - \frac{\delta}{2n^2}$,

$$
|\hat{\mathbf{V}}_{ij}(t) - \mathbf{V}_{ij}(0)| \leq \frac{8R_v}{\alpha\sqrt{2\pi}} + \frac{8R_v(m/\delta)^{1/d}}{\alpha} \leq \frac{12R_v(m/\delta)^{1/d}}{\alpha}.
$$

Taking a union bound, with probability $1 - \delta/2$,

$$
\|\hat{\mathbf{V}}(t) - \mathbf{V}(0)\|_F = \sqrt{\sum_{i,j} |\hat{\mathbf{V}}_{ij}(t) - \mathbf{V}_{ij}(0)|^2} \leq \frac{12nR_v(m/\delta)^{1/d}}{\alpha}.
$$

Bounding the spectral norm by the Frobenous norm,

$$\|\hat{\mathbf{V}}(t) - \mathbf{V}(0)\|_2 \leq \frac{12nR_v(m/\delta)^{1/d}}{\alpha}.$$

Taking $R_v = \frac{\alpha\lambda_0}{96n(m/\delta)^{1/d}}$ gives the desired bound.

$$\|\hat{\mathbf{V}}(t) - \mathbf{V}(0)\|_2 \leq \frac{\lambda_0}{8}.$$

□

Proof of Lemma B.4:
To bound $\|\mathbf{V}(t) - \mathbf{V}(0)\|_2$ we now consider $\|\mathbf{V}(t) - \hat{\mathbf{V}}(t)\|_2$. The entries of $\mathbf{V}_{ij}(t)$ are given as

$$\mathbf{V}_{ij}(t) = \frac{1}{m}\sum_{k=1}^{m} \langle \mathbf{x}_i^{\mathbf{v}_k(t)^\perp},\, x_j^{\mathbf{v}_k(t)^\perp}\rangle \mathbb{1}_{ik}(t)\mathbb{1}_{jk}(t)\left(\frac{\alpha c_k \cdot g_k}{\|\mathbf{v}_k(0)\|_2}\right)^2.$$

The surrogate $\hat{\mathbf{V}}(t)$ is defined as

$$\hat{\mathbf{V}}_{ij}(t) = \frac{1}{m}\sum_{k=1}^{m} \langle \mathbf{x}_i^{\mathbf{v}_k(t)^\perp},\, x_j^{\mathbf{v}_k(t)^\perp}\rangle \mathbb{1}_{ik}(t)\mathbb{1}_{jk}(t).$$

The only difference is in the second layer terms. The difference between each entry is written as

$$\begin{aligned}
|\mathbf{V}_{ij}(t) - \hat{\mathbf{V}}_{ij}(t)| &= \left|\frac{1}{m}\sum_{k=1}^{m} \langle \mathbf{x}_i^{\mathbf{v}_k(t)^\perp},\, x_j^{\mathbf{v}_k(t)^\perp}\rangle \mathbb{1}_{ik}(t)\mathbb{1}_{jk}(t)\left(\left(\frac{\alpha c_k \cdot g_k}{\|\mathbf{v}_k(t)\|_2}\right)^2 - 1\right)\right| \\
&\leq \max_{1\leq k\leq m}\left(\frac{\alpha^2 g_k(t)^2}{\|\mathbf{v}_k(t)\|_2^2} - 1\right).
\end{aligned}$$

Write $1 = \frac{\alpha^2 g_k^2(0)}{\|\mathbf{v}_k(0)\|_2^2}$, since $\|\mathbf{v}_k(t)\|_2$ is increasing in t according to (2.3)

$$\frac{\alpha^2 g_k(t)^2}{\|\mathbf{v}_k(t)\|_2^2} - 1 = \frac{\alpha^2 g_k(t)^2}{\|\mathbf{v}_k(t)\|_2^2} - \frac{\alpha^2 g_k(0)^2}{\|\mathbf{v}_k(0)\|_2^2} \leq 3R_g(m/\delta)^{1/d} + 3R_v(m/\delta)^{1/d}/\alpha.$$

The above inequality is shown by considering different cases for the sign of the difference $g_k(t) - g_k(0)$. Now

$$\begin{aligned}
\left|\frac{\alpha^2 g_k(t)^2}{\|\mathbf{v}_k(t)\|_2^2} - \frac{\alpha^2 g_k(0)^2}{\|\mathbf{v}_k(0)\|_2^2}\right| &= \left|\left(\frac{\alpha g_k(t)}{\|\mathbf{v}_k(t)\|_2} + \frac{\alpha g_k(0)}{\|\mathbf{v}_k(0)\|_2}\right)\left(\frac{\alpha g_k(t)}{\|\mathbf{v}_k(t)\|_2} - \frac{\alpha g_k(0)}{\|\mathbf{v}_k(0)\|_2}\right)\right| \\
&\leq \left|\left(\frac{\alpha g_k(0) + \alpha R_g}{\|\mathbf{v}_k(0)\|_2} + \frac{\alpha g_k(0)}{\|\mathbf{v}_k(0)\|_2}\right)\left(\frac{\alpha g_k(t)}{\|\mathbf{v}_k(t)\|_2} - \frac{\alpha g_k(0)}{\|\mathbf{v}_k(0)\|_2}\right)\right| \\
&\leq (2 + R_g(m/\delta)^{1/d})\left|\left(\frac{\alpha g_k(t)}{\|\mathbf{v}_k(t)\|_2} - \frac{\alpha g_k(0)}{\|\mathbf{v}_k(0)\|_2}\right)\right| \\
&\leq (2 + R_g(m/\delta)^{1/d}) \max\left(\left|\frac{\alpha(g_k(0) + R_g)}{\|\mathbf{v}_k(0)\|_2} - \frac{\alpha g_k(0)}{\|\mathbf{v}_k(0)\|_2}\right|, \left|\frac{\alpha(g_k(0) - R_g)}{\|\mathbf{v}_k(0)\|_2 + R_v} - \frac{\alpha g_k(0)}{\|\mathbf{v}_k(0)\|_2}\right|\right) \\
&\leq (2 + R_g(m/\delta)^{1/d}) \max\left(R_g(m/\delta)^{1/d}, R_g(m/\delta)^{1/d} + R_v(m/\delta)^{1/d}/\alpha\right) \\
&\leq 3R_g(m/\delta)^{1/d} + 3R_v(m/\delta)^{1/d}/\alpha,
\end{aligned}$$

where the second inequality holds due to Lemma B.10 with probability $1 - \delta$ over the initialization.

Hence:

$$\|\hat{\mathbf{V}}(t)-\mathbf{V}(t)\|_2 \le \|\hat{\mathbf{V}}(t)-\mathbf{V}(t)\|_F = \sqrt{\sum_{i,j}|\hat{\mathbf{V}}_{ij}(t)-\mathbf{V}_{ij}(t)|^2} \le 3nR_g(m/\delta)^{1/d} + 3nR_v(m/\delta)^{1/d}/\alpha.$$

Substituting R_v, R_g gives

$$\|\hat{\mathbf{V}}(t)-\mathbf{V}(t)\|_2 \le \frac{\lambda_0}{8}.$$

Now we use Lemma B.3 to get that with probability $1-\delta$

$$\|\hat{\mathbf{V}}(t)-\mathbf{V}(0)\|_2 \le \frac{\lambda_0}{8}.$$

Combining, we get with probability $1-\delta$

$$\|\mathbf{V}(t)-\mathbf{V}(0)\|_2 \le \frac{\lambda_0}{4}.$$

We note that the source for all the high probability uncertainty $1-\delta$ all arise from initialization and the application of Lemma B.10. □

Proof of Lemma B.5:
To prove the claim we consider each entry i,j of $\mathbf{G}(t)-\mathbf{G}(0)$. We have,

$$\begin{aligned}
|\mathbf{G}_{ij}(t)-\mathbf{G}_{ij}(0)| &= \left|\frac{1}{m}\sum_{k=1}^m \sigma\left(\frac{\mathbf{v}_k(t)^\top \mathbf{x}_i}{\|\mathbf{v}_k(t)\|_2}\right)\sigma\left(\frac{\mathbf{v}_k(t)^\top \mathbf{x}_j}{\|\mathbf{v}_k(t)\|_2}\right) - \sigma\left(\frac{\mathbf{v}_k(0)^\top \mathbf{x}_i}{\|\mathbf{v}_k(0)\|_2}\right)\sigma\left(\frac{\mathbf{v}_k(0)^\top \mathbf{x}_j}{\|\mathbf{v}_k(0)\|_2}\right)\right| \\
&\le \frac{1}{m}\left|\sum_{k=1}^m \sigma\left(\frac{\mathbf{v}_k(t)^\top \mathbf{x}_i}{\|\mathbf{v}_k(t)\|_2}\right)\sigma\left(\frac{\mathbf{v}_k(t)^\top \mathbf{x}_j}{\|\mathbf{v}_k(t)\|_2}\right) - \sigma\left(\frac{\mathbf{v}_k(t)^\top \mathbf{x}_i}{\|\mathbf{v}_k(t)\|_2}\right)\sigma\left(\frac{\mathbf{v}_k(0)^\top \mathbf{x}_j}{\|\mathbf{v}_k(0)\|_2}\right)\right| \\
&\quad + \frac{1}{m}\left|\sum_{k=1}^m \sigma\left(\frac{\mathbf{v}_k(t)^\top \mathbf{x}_i}{\|\mathbf{v}_k(t)\|_2}\right)\sigma\left(\frac{\mathbf{v}_k(0)^\top \mathbf{x}_j}{\|\mathbf{v}_k(0)\|_2}\right) - \sigma\left(\frac{\mathbf{v}_k(0)^\top \mathbf{x}_i}{\|\mathbf{v}_k(0)\|_2}\right)\sigma\left(\frac{\mathbf{v}_k(0)^\top \mathbf{x}_j}{\|\mathbf{v}_k(0)\|_2}\right)\right| \\
&\le 2\left\|\frac{\mathbf{v}_k(t)}{\|\mathbf{v}_k(t)\|_2} - \frac{\mathbf{v}_k(0)}{\|\mathbf{v}_k(0)\|_2}\right\|_2 \le \frac{2\tilde{R}_v(m/\delta)^{1/d}}{\alpha}.
\end{aligned}$$

In the last inequality we used the fact that

$$\left\|\frac{\mathbf{v}_k(0)}{\|\mathbf{v}_k(0)\|_2} - \frac{\mathbf{v}_k(t)}{\|\mathbf{v}_k(t)\|_2}\right\|_2 \le \frac{\|\mathbf{v}_k(t)-\mathbf{v}_k(0)\|_2}{\|\mathbf{v}_k(0)\|_2} \le \frac{(m/\delta)^{1/d}}{\alpha}\|\mathbf{v}_k(t)-\mathbf{v}_k(0)\|_2,$$

where the first inequality uses that $\|\mathbf{v}_k(0)\|_2 \le \|\mathbf{v}_k(t)\|_2$ and is intuitive from a geometrical standpoint. Algebraically given vectors $\mathbf{a},\mathbf{b}$, then for any $c \ge 1$

$$\begin{aligned}
\left\|\frac{\mathbf{a}c}{\|\mathbf{a}\|_2} - \frac{\mathbf{b}}{\|\mathbf{b}\|_2}\right\|_2^2 &= \left\|\frac{\mathbf{a}}{\|\mathbf{a}\|_2} - \frac{\mathbf{b}}{\|\mathbf{b}\|_2} + (c-1)\frac{\mathbf{a}}{\|\mathbf{a}\|_2}\right\|_2^2 \\
&= \left\|\frac{\mathbf{a}}{\|\mathbf{a}\|_2} - \frac{\mathbf{b}}{\|\mathbf{b}\|_2}\right\|_2^2 + (c-1)^2 + 2(c-1)\left\langle \frac{\mathbf{a}}{\|\mathbf{a}\|_2} - \frac{\mathbf{b}}{\|\mathbf{b}\|_2}, \frac{\mathbf{a}}{\|\mathbf{a}\|_2}\right\rangle \\
&\ge \left\|\frac{\mathbf{a}}{\|\mathbf{a}\|_2} - \frac{\mathbf{b}}{\|\mathbf{b}\|_2}\right\|_2^2 + (c-1)^2 \ge \left\|\frac{\mathbf{a}}{\|\mathbf{a}\|_2} - \frac{\mathbf{b}}{\|\mathbf{b}\|_2}\right\|_2^2.
\end{aligned}$$

The first inequality in the line above is since $\frac{\langle \mathbf{a},\mathbf{b}\rangle}{\|\mathbf{a}\|_2,\|\mathbf{b}\|_2} \le 1$.

Hence,

$$\|\mathbf{G}(t)-\mathbf{G}(0)\|_2 \le \|\mathbf{G}(t)-\mathbf{G}(0)\|_F = \sqrt{\sum_{i,j}|\mathbf{G}_{ij}(t)-\mathbf{G}_{ij}(0)|^2} \le \frac{2n\tilde{R}_v(m/\delta)^{1/d}}{\alpha\sqrt{2\pi}}.$$

Taking $\tilde{R}_v = \frac{\sqrt{2\pi}\alpha\mu_0}{8n(m/\delta)^{1/d}}$ gives the desired bound. Therefore, with probability $1-\delta$,

$$\|\mathbf{G}(t)-\mathbf{G}(0)\|_2 \le \frac{\mu_0}{4}.$$

□

Now that we have established bounds on $\mathbf{V}(t), \mathbf{G}(t)$ given that the parameters stay near initialization, we show that the evolution converges in that case:

Proof of Lemma B.6:
Consider the squared norm of the predictions $\|\mathbf{f}(t)-\mathbf{y}\|_2^2$. Taking the derivative of the loss with respect to time,

$$\frac{d}{dt}\|\mathbf{f}(t)-\mathbf{y}\|_2^2 = -2(\mathbf{f}(t)-\mathbf{y})^\top\left(\mathbf{G}(t)+\frac{\mathbf{V}(t)}{\alpha^2}\right)(\mathbf{f}(t)-\mathbf{y}).$$

Since we assume that $\lambda_{\min}\left(\mathbf{G}(t)+\frac{\mathbf{V}(t)}{\alpha^2}\right) \ge \frac{\omega}{2}$, the derivative of the squared norm is bounded as

$$\frac{d}{dt}\|\mathbf{f}(t)-\mathbf{y}\|_2^2 \le -\omega\|\mathbf{f}(t)-\mathbf{y}\|_2^2.$$

Applying an integrating factor yields

$$\|\mathbf{f}(t)-\mathbf{y}\|_2^2\exp(\omega t) \le C.$$

Substituting the initial conditions, we get

$$\|\mathbf{f}(t)-\mathbf{y}\|_2^2 \le \exp(-\omega t)\|\mathbf{f}(0)-\mathbf{y}\|_2^2.$$

□

For now, assuming the linear convergence derived in Lemma B.6, we bound the distance of the parameters from initialization. Later we combine the bound on the parameters and Lemmas B.4, B.5 bounding the least eigenvalue of $\mathbf{\Lambda}(t)$, to derive a condition on the over-parametrization m and ensure convergence from random initialization.

Proof of Lemma B.7:
Denote $f(\mathbf{x}_i)$ at time t as $f_i(t)$. Since $\|\mathbf{x}_i^{\mathbf{v}_k(t)^\perp}\|_2 \le \|\mathbf{x}_i\|_2 = 1$, we have that

$$\begin{aligned}\left\|\frac{d\mathbf{v}_k(t)}{dt}\right\|_2 &= \left\|\sum_{i=1}^n (y_i-f_i(t))\frac{1}{\sqrt{m}}c_k g_k(t)\frac{1}{\|\mathbf{v}_k(t)\|_2}\mathbf{x}_i^{\mathbf{v}^\perp}\mathbb{1}_{ik}(t)\right\|_2 \\ &\le \frac{1}{\sqrt{m}}\sum_{i=1}^n |y_i-f_i(t)|\frac{c_k g_k(t)}{\|\mathbf{v}_k(t)\|_2}.\end{aligned}$$

Now using (2.3) and the initialization $\|\mathbf{v}_k(0)\| = \alpha g_k(0)$, we bound $\left|\frac{c_k g_k(t)}{\|\mathbf{v}_k(t)\|_2}\right|$,

$$\left|\frac{c_k g_k(t)}{\|\mathbf{v}_k(t)\|_2}\right| \le \left|c_k\left(\frac{g_k(0)+R_g}{\|\mathbf{v}_k(0)\|_2}\right)\right| \le \frac{1}{\alpha}\left(1+\alpha R_g/\|\mathbf{v}_k(0)\|_2\right).$$

By Lemma B.10, we have that with probability $1-\delta$ over the initialization,

$$\alpha/\|\mathbf{v}_k(0)\|_2 \leq C(m/\delta)^{1/d}.$$

Hence $\alpha R_g/\|\mathbf{v}_k(0)\|_2 \leq 1$. This fact bounds $\left|\frac{c_k g_k(t)}{\|\mathbf{v}_k(t)\|_2}\right|$ with probability $1-\delta$ for each k,

$$\left|\frac{c_k g_k(t)}{\|\mathbf{v}_k(t)\|_2}\right| \leq 2/\alpha.$$

Substituting the bound,

$$\begin{aligned}\left\|\frac{d}{dt}\mathbf{v}_k(t)\right\|_2 &\leq \frac{2}{\alpha\sqrt{m}}\sum_{i=1}^{n}|f_i(t)-y_i| \\ &\leq \frac{2\sqrt{n}}{\alpha\sqrt{m}}\|\mathbf{f}(t)-\mathbf{y}\|_2 \\ &\leq \frac{2\sqrt{n}}{\alpha\sqrt{m}}\exp(-\omega t/2)\|\mathbf{f}(0)-\mathbf{y}\|_2.\end{aligned}$$

Thus, integrating and applying Jensen's inequality,

$$\|\mathbf{v}_k(t)-\mathbf{v}_k(0)\|_2 \leq \int_0^s \left\|\frac{d\mathbf{v}_k(s)}{dt}\right\|_2 ds \leq \frac{4\sqrt{n}\|\mathbf{f}(0)-\mathbf{y}\|_2}{\alpha\omega\sqrt{m}}.$$

Note that the condition $|g_k(t)-g_k(0)| \leq R_g$ is stronger than needed and merely assuring that $|g_k(t)-g_k(0)| \leq 1/(m/\delta)^{1/d}$ suffices. □

Analogously we derive bounds for the distance of g_k from initialization.

Proof of Lemma B.8:
Consider the magnitude of the derivative $\frac{dg_k}{dt}$,

$$\left|\frac{dg_k}{dt}\right| = \left|\frac{1}{\sqrt{m}}\sum_{j=1}^{n}(f_j-y_j)\frac{c_k}{\|\mathbf{v}_k\|_2}\sigma(\mathbf{v}_k^\top\mathbf{x}_j)\right|.$$

Note

$$\left|\frac{c_k}{\|\mathbf{v}_k\|_2}\sigma(\mathbf{v}_k^\top\mathbf{x}_j)\right| = \left|\sigma\left(\frac{\mathbf{v}_k^\top\mathbf{x}_j}{\|\mathbf{v}_k\|_2}\right)\right| \leq 1$$

Thus applying Cauchy Schwartz

$$\left|\frac{dg_k(t)}{dt}\right| \leq \frac{2\sqrt{n}}{\sqrt{m}}\|\mathbf{f}(t)-\mathbf{y}\|_2 \leq \frac{2\sqrt{n}}{\sqrt{m}}\exp(-\omega t/2)\|\mathbf{f}(0)-\mathbf{y}\|_2,$$

and integrating from 0 to t yields

$$|g_k(t)-g_k(0)| \leq \int_0^t \left|\frac{dg_k}{dt}(s)\right| ds \leq \int_0^t \frac{2\sqrt{n}}{\sqrt{m}}\exp(-\omega s/2)\|\mathbf{f}(0)-\mathbf{y}\|_2 ds \leq \frac{4\sqrt{n}\|\mathbf{y}-\mathbf{f}(0)\|_2}{\sqrt{m}\omega}.$$

□

Proof of Lemma B.9:
Consider the ith entry of the network at initialization,

$$f_i(0) = \frac{1}{\sqrt{m}}\sum_{k=1}^{m} c_k\sigma\left(\frac{g_k\mathbf{v}_k^\top\mathbf{x}_i}{\|\mathbf{v}_k\|_2}\right).$$

Since the network is initialized randomly and m is taken to be large we apply concentration to bound $f_i(0)$ for each i. Define $z_k = c_k \sigma\left(\frac{g_k(0)\mathbf{v}_k(0)^\top \mathbf{x}_i}{\|\mathbf{v}_k(0)\|_2}\right)$. Note that z_k are independent sub-Gaussian random variables with

$$\|\mathbf{z}_k\|_\psi \leq \|N(0,1)\|_\psi = C.$$

Here $\|\cdot\|_\psi$ denotes the 2-sub-Gaussian norm, (see (Vershynin, 2018) for example). Applying Hoeffding's inequality bounds $f_i(0)$ as

$$\begin{aligned}\mathbb{P}(|\sqrt{m} f_i(0)| > t) &\leq 2\exp\left(-\frac{t^2/2}{\sum_{k=1}^m \|\mathbf{z}_k\|_{\psi_2}}\right) \\ &= 2\exp\left(\frac{-t^2}{2mC}\right).\end{aligned}$$

Which gives with probability $1-\delta/n$ that

$$|f_i(0)| \leq \tilde{C}\sqrt{\log(n/\delta)}.$$

Now with probability $1-\delta$ we have that, for each i,

$$|f_i(0) - y_i| \leq |y_i| + \tilde{C}\sqrt{\log(n/\delta)} \leq C_2\sqrt{\log(n/\delta)}.$$

Since $y_i = O(1)$. Hence, with probability $1-\delta$,

$$\|\mathbf{f}(0) - \mathbf{y}\|_2 \leq C\sqrt{n\log(n/\delta)}.$$

□

Proof of Lemma B.10:
At initialization $\mathbf{v}_k \sim N(0, \alpha^2\mathbf{I})$ so the norm behaves like $\|\mathbf{v}_k(0)\|_2^2 \sim \alpha^2\chi_d$. The cumulative density of a chi-squared distribution with d degrees of freedom behaves like $F(x) = \Theta(x^{d/2})$ for small x so we have that with probability $1-\frac{\delta}{m}$, that $\|\mathbf{v}_k(0)\|_2 \geq \alpha(m/\delta)^{\frac{1}{d}}$ where d is the input dimension. Applying a union bound, with probability $1-\delta$, for all $1 \leq k \leq m$,

$$\frac{1}{\|\mathbf{v}_k(0)\|_2} \leq \frac{\left(m/\delta\right)^{1/d}}{\alpha}.$$

Now by (2.3) for $t \geq 0$, $\|\mathbf{v}_k(t)\|_2 \geq \|\mathbf{v}_k(0)\|_2$ so

$$\frac{1}{\|\mathbf{v}_k(t)\|_2} \leq \frac{1}{\|\mathbf{v}_k(0)\|_2} \leq \frac{\left(m/\delta\right)^{1/d}}{\alpha}.$$

□

E. Proofs of Lemmas from Appendix C and Proposition 2

Proof of Proposition 2:
The proof of proposition 2, follows the proofs of Theorems 4.1, 4.2, and relies on Theorem C.1. In particular for each $\alpha > 0$ at initialization, take $\omega_\alpha(s) = \lambda_{\min}(\mathbf{\Lambda}(s))$ and define the auxiliary $\omega_{\alpha,0} = \lambda_{\min}(\mathbf{V}^\infty/\alpha^2 + \mathbf{G}^\infty)$. Then we have that

$$\omega_{\alpha,0} \geq \lambda_0/\alpha^2 + \mu_0 > 0.$$

Hence, by the same arguments of Theorem 4.1, 4.2 for $\omega_\alpha(s)$ if $m = \left(n^4\log(n/\delta)/\alpha^4\omega_{\alpha,0}^4\right)$, then we have that the conditions of Theorem C.1 are satisfied, namely, $\lambda(s) \geq \frac{\lambda_0}{2}$ and $\mu(s) \geq \frac{\mu_0}{2}$. Taking $\eta_\alpha = O\left(\frac{1}{\|\mathbf{\Lambda}(s)\|_2}\right)$, then the required

step-size for convergence is satisfied. This follows from the same argument of Theorems 4.1, 4.2 and depends on the fact that $\|\mathbf{\Lambda}(s) - \mathbf{\Lambda}(0)\|_2 \leq \frac{1}{\alpha^2}\|\mathbf{V}(s) - \mathbf{V}^\infty(0)\|_2 + \|\mathbf{G}(s) - \mathbf{G}(0)\|_2$. Now we consider the term, $\alpha\omega_{\alpha,0}$. For $\alpha = 1$,

$$\alpha\omega_{\alpha,0} = \lambda_{\min}(\mathbf{H}^\infty).$$

Which matches the results of un-normalized convergence. In general, we have that

$$\alpha\omega_{\alpha,0} \geq \alpha(\lambda_0/\alpha^2 + \mu_0) \geq \min\{\lambda_0, \mu_0\}.$$

Therefore the bound on m is taken to be independent of α as $m = \Omega\left(\frac{n^4\log(n/\delta)}{\min\{\mu_0^4, \lambda_0^4\}}\right)$ which simplifies the presentation. Now for each α the effective convergence rate is dictated by the least eigenvalue ω_α and the allowed step-size η_α as,

$$\left(1 - \eta_\alpha\omega_\alpha\right).$$

Then taking $\alpha^* = \text{argmin}_{\alpha>0}(1 - \eta_\alpha\omega_\alpha)$ we have that

$$(1 - \eta_{\alpha^*}\omega_{\alpha^*}) \leq (1 - \eta_1\omega_1).$$

which corresponds to the un-normalized converegence rate. Therefore as compared with un-normalized training we have that for α^*, WN enables a faster convergence rate. □

Proof of Lemma C.1:
Fix R, without the loss of generality we write S_i for $S_i(R)$. For each k, $\mathbf{v}_k(0)$ is initialized independently via $\sim N(0, \alpha^2\mathbf{I})$, and for a given k, the event $\mathbb{1}_{ik}(0) \neq \mathbb{1}\{\mathbf{v}^\top\mathbf{x}_i \geq 0\}$ corresponds to $|\mathbf{v}_k(0)^\top\mathbf{x}_i| \leq R$. Since $\|\mathbf{x}_i\|_2 = 1$, $\mathbf{v}_k(0)^\top\mathbf{x}_i \sim N(0, \alpha^2)$. Denoting the event that an index $k \in S_i$ as $A_{i,k}$, we have

$$\mathbb{P}(A_{i,k}) \leq \frac{2R}{\alpha\sqrt{2\pi}}.$$

Next the cardinality of S_i is written as

$$|S_i| = \sum_{k=1}^{m} \mathbb{1}_{A_{i,k}}.$$

Applying Lemma D.1, with probability $1 - \delta/n$,

$$|S_i| \leq \frac{2mR}{\alpha\sqrt{2\pi}} + \frac{16\log(n/\delta)}{3}.$$

Taking a union bound, with probability $1 - \delta$, for all i we have that

$$|S_i| \leq \frac{2mR}{\alpha\sqrt{2\pi}} + \frac{16\log(n/\delta)}{3}.$$

□

Proof of Lemma C.2:
To show this we bound the difference $g_k(s) - g_k(0)$ as the sum of the iteration updates. Each update is written as

$$\left|\frac{\partial L(s)}{\partial g_k}\right| = \left|\frac{1}{\sqrt{m}}\sum_{i=1}^{n}(f_i(s) - y_i)\frac{c_k}{\|\mathbf{v}_k(s)\|_2}\sigma(\mathbf{v}_k(s)^\top\mathbf{x}_i)\right|.$$

As $\left|c_k\sigma\left(\frac{\mathbf{v}_k(s)^\top\mathbf{x}_i}{\|\mathbf{v}_k(s)\|_2}\right)\right| \leq 1$,

$$\left|\frac{\partial L(s)}{\partial g_k}\right| \leq \frac{1}{\sqrt{m}}\sum_{i}^{n}|f_i(s) - y_i| \leq \frac{\sqrt{n}}{\sqrt{m}}\|\mathbf{f}(s) - \mathbf{y}\|_2.$$

By the assumption in the statement of the lemma,

$$\left|\frac{\partial L(s)}{\partial g_k}\right| \leq \frac{\sqrt{n}(1-\frac{\eta\omega}{2})^{s/2}\|\mathbf{f}(0)-\mathbf{y}\|_2}{\sqrt{m}}.$$

Hence bounding the difference by the sum of the gradient updates:

$$|g_k(K+1)-g_k(0)| \leq \eta\sum_{s=0}^{K}\left|\frac{\partial L(s)}{\partial g_k}\right| \leq \frac{4\eta\sqrt{n}\|\mathbf{f}(0)-\mathbf{y}\|_2}{\sqrt{m}}\sum_{s=0}^{K}(1-\frac{\eta\omega}{2})^{s/2}.$$

The last term yields a geometric series that is bounded as

$$\frac{1}{1-\sqrt{1-\frac{\eta\omega}{2}}} \leq \frac{4}{\eta\omega},$$

Hence

$$|g_k(K+1)-g_k(0)| \leq \frac{4\sqrt{n}\|\mathbf{f}(0)-\mathbf{y}\|_2}{\omega\sqrt{m}}.$$

□

Proof of Lemma C.3:

To show this we write $\mathbf{v}_k(s)$ as the sum of gradient updates and the initial weight $\mathbf{v}_k(0)$. Consider the norm of the gradient of the loss with respect to $\mathbf{v}_k$,

$$\|\nabla_{\mathbf{v}_k}L(s)\|_2 = \left\|\frac{1}{\sqrt{m}}\sum_{i=1}^{n}(f_i(s)-y_i)\frac{c_k g_k(s)}{\|\mathbf{v}_k(s)\|_2}\mathbb{1}_{ik}(s)\mathbf{x}_i^{\mathbf{v}_k(s)^\perp}\right\|_2.$$

Since $\|\mathbf{v}_k(s)\|_2 \geq \|\mathbf{v}_k(0)\|_2 \geq \alpha(\delta/m)^{1/d}$ with probability $1-\delta$ over the initialization, applying Cauchy Schwartz's inequality gives

$$\|\nabla_{\mathbf{v}_k}L(s)\|_2 \leq \frac{(1+R_g(m/\delta)^{1/d})\sqrt{n}\|\mathbf{f}(s)-\mathbf{y}\|_2}{\alpha\sqrt{m}}. \tag{E.1}$$

By the assumption on $\|\mathbf{f}(s)-\mathbf{y}\|_2$ this gives

$$\|\nabla_{\mathbf{v}_k}L(s)\|_2 \leq \frac{2\sqrt{n}(1-\frac{\eta\omega}{2})^{s/2}\|\mathbf{f}(0)-\mathbf{y}\|_2}{\alpha\sqrt{m}}.$$

Hence bounding the parameter trajectory by the sum of the gradient updates:

$$\|\mathbf{v}_k(K+1)-\mathbf{v}_k(0)\|_2 \leq \eta\sum_{s=0}^{K}\|\nabla_{\mathbf{v}_k}L(s)\|_2 \leq \frac{2\sqrt{n}\|\mathbf{f}(0)-\mathbf{y}\|_2}{\alpha\sqrt{m}}\sum_{s=1}^{K}\left(1-\frac{\eta\omega}{2}\right)^{s/2}$$

yields a geometric series. Now the series is bounded as

$$\frac{1}{1-\sqrt{1-\frac{\eta\omega}{2}}} \leq \frac{4}{\eta\omega},$$

which gives

$$\|\mathbf{v}_k(K+1)-\mathbf{v}_k(0)\|_2 \leq \frac{8\sqrt{n}\|\mathbf{f}(0)-\mathbf{y}\|_2}{\alpha\sqrt{m}\omega}.$$

□