Infinite Width Limits of Self Supervised Neural Networks

Maximilian Fleissner
Technical University of Munich

Gautham Govind Anil
IIT Madras

Debarghya Ghoshdastidar
Technical University of Munich

Abstract

The NTK is a widely used tool in the theoretical analysis of deep learning, allowing us to look at supervised deep neural networks through the lenses of kernel regression. Recently, several works have investigated kernel models for self-supervised learning, hypothesizing that these also shed light on the behavior of wide neural networks by virtue of the NTK. However, it remains an open question to what extent this connection is mathematically sound — it is a commonly encountered misbelief that the kernel behavior of wide neural networks emerges irrespective of the loss function it is trained on. In this paper, we bridge the gap between the NTK and self-supervised learning, focusing on two-layer neural networks trained under the Barlow Twins loss. We prove that the NTK of Barlow Twins indeed becomes constant as the width of the network approaches infinity. Our analysis technique is a bit different from previous works on the NTK and may be of independent interest. Overall, our work provides a first justification for the use of classic kernel theory to understand self-supervised learning of wide neural networks. Building on this result, we derive generalization error bounds for kernelized Barlow Twins and connect them to neural networks of finite width.

1 INTRODUCTION

In recent years, self-supervised learning (SSL) has emerged as a powerful paradigm, building the foundation of several modern machine learning models. At its core, SSL relies on the idea of using augmentations to encode a notion of similarity in otherwise unlabeled data. As a typical example, consider an image dataset. Even though labels may not be known, it is reasonable to believe that randomly cropping, slightly rotating, or blurring the images will not change the true underlying class information. Therefore, two augmented versions (x, x^+) of the same image should receive similar representations $f(x), f(x^+)$ in a lower-dimensional ambient space. This constitutes the basic intuition behind non-contrastive SSL, and several loss functions that capture this idea have emerged.[1] Among the most popular losses is Barlow Twins (Zbontar et al., 2021), a loss function that pushes the cross-correlation of the embeddings $f(x), f(x^+)$ towards the identity matrix. This aims to prevent a phenomenon known as dimension collapse, where the learned representations collapse to a single point in the embedding space.

Despite the empirical success of SSL on a range of tasks (Radford et al., 2021; Bachman et al., 2019), it has taken quite some time for deep learning theory to catch up with this innovation. Arguably one of the most promising avenues towards understanding the fundamental principles of SSL is by connecting it to kernel methods (Smola and Schölkopf, 1998). This of course is reminiscent of the supervised setting, where the NTK (Jacot et al., 2018; Lee et al., 2019) provides a powerful framework to understand several phenomena of deep learning with wide neural networks, including generalization (Simon et al., 2021), benign overfitting (Mallinar et al., 2022), and robustness (Bombari et al., 2023). While a significant number of researchers are currently looking at SSL from a kernel perspective (for an overview, see related works), and while almost all of these works are motivated with the NTK, the connection is left implicit. However, it is not a priori clear that neural networks trained under SSL actually behave like kernel machines in the infinite width limit. In fact, Anil et al. (2024) have recently demonstrated that for certain contrastive loss functions, the NTK is in fact **not** constant at infinite width. This casts a shadow of doubt on the validity of kernel approxima-

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

[1] In contrastive SSL, one additionally incorporates a notion of dissimilarity into the model by including negative examples x^- into the training procedure.

tions to SSL, urging us to take a closer look at the matter.

In this paper, we bridge the gap between SSL and the NTK for the Barlow Twins loss. We prove that for neural networks with one hidden layer, the neural tangent kernel indeed becomes constant as the width of the network approaches infinity. The proof technique is different from previous works on the NTK, and leverages Grönwall's inequality (see Appendix F.1). This is necessitated by the training dynamics of the Barlow Twins loss, which make an extension of existing methods difficult. Our work confirms the hypothesized connection between kernel methods and neural network based SSL, and proves that Barlow Twins is akin to one of the most prominent representation learning methods, Kernel PCA (Schölkopf et al., 1997). Building on these insights, we use classic tools from learning theory to derive generalization error bounds for kernel versions of Barlow Twins, and then connect them to neural networks of finite width.

This paper is structured as follows. We discuss related works in Section 2 and present our formal setup in Section 3. Section 4 contains our main result, stated in the general setting of multi-dimensional embeddings $f : \mathbb{R}^d \to \mathbb{R}^K$ for the Barlow twins loss. In Section 5 we give a detailed proof sketch of the one-dimensional case, which is computationally less involved but contains the main technical ideas. Section 6 derives generalization error bounds for the Barlow twins loss in an abstract Hilbert space setting. Using the kernel trick and our newly established validity of the NTK approximation, we then relate these bounds to finite neural networks. Finally, we empirically verify our findings through experiments in Section 7.

2 RELATED WORK

In comparison to supervised learning, the theoretical understanding of self-supervised learning is still at an early stage. Nonetheless, several works have investigated SSL using classic tools from statistical learning theory (Arora et al., 2019; Wei et al., 2020). Furthermore, there have been several successful attempts to look at SSL through the lenses of classic spectral and kernel methods (Kiani et al., 2022; Johnson et al., 2022; HaoChen et al., 2021; Cabannes et al., 2023; Esser et al., 2024). These works provide a number of useful insights into both theoretical as well as practical aspects of training SSL models, but leave the formal connection between the kernel regime and deep learning based SSL implicit.

The idea that neural networks behave like (neural tangent) kernel models in the infinite width limit was first investigated by Jacot et al. (2018). Several later works further explored this connection in various contexts (Arora et al., 2019; Chizat et al., 2019; Lee et al., 2019; Liu et al., 2020a), mostly for the squared (or hinge) loss. In particular, Liu et al. (2020b) develop a general framework for looking at convergence to the NTK using the Hessian. The Barlow Twins loss does not fall under the umbrella of previous analysis, since it is a fourth order loss with very different training dynamics.

Thus, even though Simon et al. (2023) motivate their investigation of kernel versions of Barlow Twins precisely with the NTK, we are not aware of any derivation that proves this analogy is valid. Ziyin et al. (2022) investigate the landscape of several SSL losses, but their theory is stated in the linear setting. Closest to our work is Anil et al. (2024), who manage to bound the evolution of the NTK for certain contrastive loss functions, but do not quite prove constancy until convergence of the loss: Their bounds hold until a time that grows with network width, leaving the possibility that as wider networks are trained, the time till convergence also grows.

3 FORMAL SETUP

In this work, we analyze an idealized version of the popular Barlow Twins loss (Zbontar et al., 2021), as considered in previous theoretical works on SSL (Simon et al., 2023). Training data consists of positive pairs, denoted $\mathcal{X} = \{x_1, \dots, x_N\} \subset \mathbb{R}^d$ and $\mathcal{X}_+ = \{x_1^+, \dots, x_N^+\} \subset \mathbb{R}^d$. We assume data lies on the unit sphere, that is $\|x_n\| = \|x_n^+\| = 1$ for all $n \in [N]$. For example, each pair (x_n, x_n^+) could consist of two (normalized) augmentations of the same underlying image, randomly cropped or blurred. The goal of Barlow Twins is now to learn the parameters of a neural network $f : \mathbb{R}^d \to \mathbb{R}^K$ that embeds each point into a lower-dimensional space. Typically, $K \ll d$. To encourage useful representations, the Barlow Twins loss function pushes the cross-moment matrix of $f(\mathcal{X})$ and $f(\mathcal{X}_+)$ to the identity matrix in $\mathbb{R}^K$. Defining

$$C = \frac{1}{2N} \sum_{n=1}^{N} f(x_n) f(x_n^+)^\top + f(x_n^+) f(x_n)^\top \tag{1}$$

we minimize the loss function $\mathcal{L}(f) = \|C - I\|_F^2$ over the parameters of a neural network f. In this work, we restrict our analysis to two-layer neural networks, that is

$$f(x) = \frac{1}{\sqrt{M}} \sum_{m=1}^{M} w_m \phi(v_m^\top x) \tag{2}$$

where $w_m \in \mathbb{R}^K$ and $v_m \in \mathbb{R}^d$ for all $m \in [M]$, and ϕ is a bounded, smooth activation function, and has

bounded first derivative. For example, we could have $\phi(t) = \tanh(t)$. We denote c_ϕ and $c_{\phi'}$ for the supremum norms of ϕ and its derivative. We will later also discuss the ReLU activation. The weights are initialized as random independent Gaussians with constant variance, and collected in a vector $\theta \in \mathbb{R}^{M(d+K)}$ that is trained under gradient flow

$$\frac{\partial \theta}{\partial t} = \dot{\theta}(t) = -\frac{\partial \mathcal{L}}{\partial \theta} \tag{3}$$

We write θ_0 for the weights at initialization, and sometimes denote $f(x;\theta)$ to emphasize that f is a function both of the inputs as well as of its parameters. The neural tangent kernel is defined as a time-varying, matrix-valued map $K_t : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{K\times K}$, where

$$K_t(x,x') = \left(\left(\frac{\partial f_k(x)}{\partial \theta(t)} \right)^\top \left(\frac{\partial f_l(x')}{\partial \theta(t)} \right) \right)_{k,l=1}^{K} \tag{4}$$

for all $x, x' \in \mathbb{R}^d$, and f_k is the k-th output dimension of f. To underline the dependence of the NTK on the parameters $\theta(t)$ that evolve during training, we sometimes also denote it as K_θ. The key insight of the NTK literature is that the NTK does not change during training if the width of the neural network approaches infinity. Consequently, the training dynamics of f approach those of kernel regression with respect to the (vector-valued) kernel at initialization K_0.

The constancy of the NTK in the infinite width limit essentially relies on three facts: Firstly, the spectral norm of the Hessian of the neural network is $\mathcal{O}(\frac{R}{\sqrt{M}})$ for all weights θ with $\|\theta - \theta_0\| \leq R$. Secondly, the change in the NTK from θ_0 to any θ can be bounded in terms of the Hessian and $\|\theta - \theta_0\|$. And thirdly, R is independent of M because convergence happens in a ball of width-independent radius around θ_0. The first fact is true regardless of the loss function, see Lemma 7 in (Anil et al., 2024). The same is true for the second fact: We recall the following Lemma 8 by Anil et al. (2024) which extends the result by Liu et al. (2020a) to functions f with multivariate outputs.

Theorem 3.1. *Consider an initial parameter $\theta_0 \in \mathbb{R}^p$ and a ball $B(\theta_0, R)$ around θ_0 of radius $R > 0$. Suppose that $\forall k \in [K]$, all inputs a and all $\theta \in B(\theta_0, R)$, the Hessian of $f_k(a;\theta)$ in parameter space satisfies*

$$\left\| \nabla^2_\theta f_k(a;\theta) \right\|_2 \leq \epsilon \tag{5}$$

and that $\|\nabla_\theta f_k(a;\theta)\|_2 \leq c_0$. Then, the change in the neural tangent kernel is bounded by

$$\left| (K_{\theta_0}(a,b))_{k,l} - (K_\theta(a,b))_{k,l} \right| \leq 2\epsilon c_0 R \tag{6}$$

for all $\theta \in B(\theta_0, R)$, all $k, l \in [K]$ and all inputs a, b. Moreover, for the neural networks considered in this paper, $c_0 = \mathcal{O}(1)$ whenever $R < \sqrt{M}$, see Appendix F.2.

However, the third piece in the puzzle is missing: Unless R stays independent of M, we do not obtain constancy of the NTK at large width $M \to \infty$. For the squared error, Liu et al. (2020b) use the Polyak-Lojasiewicz condition to ensure that the weights remain in a bounded ball of width-independent radius R. It is not obvious how to extend this idea to Barlow Twins loss. In this paper, we therefore take a slightly different approach, and look at the evolution of the loss and the parameters *in time*.

4 MAIN RESULT

In light of the aforementioned discussion, we focus solely on proving that the weights of the network stay in a ball of fixed radius R, with high probability over random initialization. To this end, we fix a global constant $\delta > 0$, below which the loss is considered to be zero, and training is stopped. Our strategy is to verify the following two statements, both of which hold at large width M.

1. For any finite time $T > 0$, there exists a width-independent $\kappa > 0$ such that $\sup_{t\leq T} \|\dot{\theta}(t)\| \leq \kappa$ with high probability for any network of sufficiently large width.
2. There exists a finite time T such that $\mathcal{L}(T) \leq \delta$ for any network of sufficiently large width.

Together, both imply that for large enough M, the weights θ remain within a ball of width-independent radius R around the initial θ_0 until T, with high probability. The reason is the following: Writing $u(t) = \|\theta_t - \theta_0\|^2$, Cauchy-Schwartz implies

$$\begin{aligned} \frac{\partial}{\partial t} u(t) &= 2 \sum_{m=1}^{M} (\theta_m(t) - \theta_m(0))\, \dot{\theta_m}(t) \\ &\leq 2\|\theta_t - \theta_0\| \cdot \|\dot{\theta}(t)\| \\ &\leq 2\kappa \sqrt{u(t)} \end{aligned} \tag{7}$$

Since $u(0) = 0$, the comparison principle for ordinary differential equations shows $u(t) \leq \kappa^2 t^2$ for all $t \leq T$. Thus, the weights θ remain in a ball of radius $R = \kappa T$ around θ_0 until convergence. We emphasize that κ can (and will) depend on T as well.

Theorem 4.1. ***(Gradient of weights is bounded)*** *Fix any $T \geq 0$ and any $\epsilon > 0$. Then, there exists $M_0 \in \mathbb{N}$ and some $\kappa > 0$ such that, for all networks of width $M > M_0$, with probability at least $1 - \epsilon$, the weights θ satisfy $\|\dot{\theta}(t)\| \leq \kappa$ for all $t \leq T$ when trained under gradient flow.*

The proof is included in Appendix A.1. The statement is probabilistic because it only holds if the weights are

not too large at initialization (which is certainly true with high probability). It remains to show that for some sufficiently large width M, the convergence of the loss indeed happens within some finite time T, up to our fixed tolerance $\delta > 0$. We therefore turn to the evolution of $\mathcal{L}(t)$ over time. Defining

$$u(t) = \begin{bmatrix} \left(\frac{\partial \mathcal{L}}{\partial f_k(x_n)}\right)_{n,k} \\ \left(\frac{\partial \mathcal{L}}{\partial f_k(x_n^+)}\right)_{n,k} \end{bmatrix} \in \mathbb{R}^{2NK}$$

$$\boldsymbol{K}(t) = \begin{bmatrix} K_t(x_1, x_1) & \dots & K_t(x_1, x_N^+) \\ \dots & \dots & \dots \\ K_t(x_N^+, x_1) & \dots & K_t(x_N^+, x_N^+) \end{bmatrix} \in \mathbb{R}^{2NK \times 2NK} \tag{8}$$

we express the time evolution of the Barlow Twins loss in a more concise manner.

Lemma 4.2. ***(Time evolution of the loss)*** *Under gradient flow, the evolution of $\mathcal{L}(t)$ can be expressed as*

$$\frac{\partial}{\partial t}\mathcal{L}(t) = -u(t)^\top \boldsymbol{K}(t) u(t) \tag{9}$$

The proof is included in Appendix A.2. Our main theorem requires an assumption on the loss at initialization.

Definition 4.3. ***(Definitions and Assumptions at initialization)*** *We assume that, under the initialization that ensures Theorem 4.1 to hold, the matrix $\boldsymbol{K}(0)$ is positive definite, with smallest eigenvalue $\lambda_{min}(\boldsymbol{K}(0)) \geq \lambda > 0$, and that $\mathcal{L}(0) \leq 1 - \rho < 1$ for some small $\rho > 0$. Define*

$$\eta = \frac{4\lambda(1 - \sqrt{1-\rho})}{N} \tag{10}$$

and choose $T = T(\eta)$ such that the solution to the autonomous ODE $\frac{\partial}{\partial t}L = -\eta L$ satisfies $L(T) \leq \delta$, under the initial condition $L(0) = 1 - \rho$.

In particular, η and $T = T(\eta)$ are independent of the network width. We state the main result.

Theorem 4.4. ***(Convergence of the loss in finite time)*** *Under the setting in Definition 4.3, there exists $M_1 \in \mathbb{N}$ such that, for all networks of width $M \geq M_1$, the loss under gradient flow satisfies $\frac{\partial}{\partial t}\mathcal{L}(t) < -\eta\mathcal{L}(t)$ for all time $t \leq T$. This implies that $\mathcal{L}(T) < \delta$.*

The proof of Theorem 4.4 is included in Appendix A.3. As a direct consequence of this result and of Theorem 3.1, we obtain the following corollary.

Corollary 4.5. ***(Convergence of the NTK)*** *Under the conditions of Theorem 4.4, there exists a radius $R > 0$ such that, with probability at least $1 - \epsilon$, the change of the NTK until convergence is $\mathcal{O}(R^2/\sqrt{M})$, with R depending only on δ, η, but independent of the network width.*

For the ReLU activation $\phi(x) = \max(0, x)$, gradient flow itself is ill-defined, due to the non-differentiability of the ReLU at zero. However, if we define the weak derivative $\frac{\partial \phi(0)}{\partial x} = 0$, and equate $\dot{\theta}(t) = -\frac{\partial \mathcal{L}}{\partial \theta}$ as per usual, then it is possible to prove Theorem 4.1 nonetheless, and all other results remain true as well. Of course, this is not entirely rigorous, because $\dot{\theta}(t) = -\frac{\partial \mathcal{L}}{\partial \theta}$ is not actually gradient flow. See Appendix C for details.

Remark 4.6. ***(Dropping the assumption on the loss)*** *The condition $\mathcal{L}(0) < 1$ is most likely not necessary. In Appendix E, we prove that in the linearized regime, the loss converges exponentially to zero if all eigenvalues of the embedding cross-moment matrix $C(0)$ are contained in $(0, 1)$. This suggests that "small" positive definite initialization is sufficient to enter the kernel regime at large width. Our experiments also do not impose it, and yet show convergence to the NTK (see Section 7). For wide ReLU networks, $\mathcal{L}(0) < 1$ can however be guaranteed by suitably scaling the Gaussian weights $v_m \in \mathbb{R}^d$ of the first layer by a data-dependent constant (see Appendix A.4).*

5 PROOF SKETCH FOR ONE DIMENSIONAL EMBEDDINGS

To give a better intuition on our proof strategy, we present a more detailed proof sketch for the simplest possible case of one-dimensional embeddings $f : \mathbb{R}^d \to \mathbb{R}$. We consider a neural network with one hidden layer and d-dimensional inputs x. For one-dimensional embeddings, the Barlow Twins loss function equates to

$$\mathcal{L}(f) = \left(\underbrace{\left(\frac{1}{N} \sum_{n=1}^{N} f(x_n) f(x_n^+) \right)}_{=:C} - 1 \right)^2 \tag{11}$$

For ease of notation we define the cross-moment matrix of the embeddings

$$C = \frac{1}{N} \sum_{n=1}^{N} f(x_n) f(x_n^+) \tag{12}$$

As discussed in Section 4, we aim to verify that at sufficiently large width (i) there exists a width-independent $\kappa > 0$ such that $\sup_{t \leq T} \|\dot{\theta}(t)\| \leq \kappa$ with high probability over random initialization of weights and (ii) $\mathcal{L}(T) \leq \delta$ at a finite time T independent of M. We begin by checking the first condition.

Theorem 5.1. ***(Gradient of weights is bounded)*** *Fix any $T \geq 0$ and any $\epsilon > 0$. Then, there exists $M_0 \in \mathbb{N}$ and some $\kappa > 0$ such that, for all networks of width $M > M_0$, with probability at least $1 - \epsilon$, the weights θ satisfy $\|\dot{\theta}(t)\| \leq \kappa$ for all $t \leq T$ when trained under gradient flow.*

The proof is included in Appendix B.1. For simplicity, the proof assumes $\mathcal{L}(0) < 1$, although this is not necessary here yet. Essentially, the proof proceeds in three steps.

1. Firstly, we show that $\frac{\partial}{\partial t}\left(\sum_{m=1}^{M} w_m^2(t)\right) \leq 8$ for all $t \leq T$. This implies that there exists some $\kappa_1 > 0$ such that for all $t \leq T$, we have $\frac{1}{M}\left(\sum_{m=1}^{M} w_m^2(t)\right) \leq \kappa_1$ with probability $\geq 1-\epsilon$.

2. From there, we bound the maximum squared value that any representation takes until time T, that is

$$|f|^2 = \max_{n\in[N]} \sup_{t\leq T} \max\left(|f(x_n)|^2, |f(x_n^+)|^2\right) \quad (13)$$

by some $\kappa_2 > 0$ (again independent of M). This requires the first part of the proof, as well as Grönwall's inequality (see Appendix F.1). Grönwall's inequality introduces potentially exponential dependencies of $|f|^2$ on T. Note that it is not a priori clear that $|f|^2$ remains bounded: While we know that the products $f(x_n)f(x_n^+)$ cannot explode (otherwise, the loss would also explode, which is impossible under gradient flow), the individual terms could be large.

3. Finally, we show that

$$\|\dot{\theta}(t)\|^2 \leq \frac{16|f|^2}{M}\left(c_\phi^2 + dc_{\phi'}^2\left(\sum_{m=1}^{M} w_m^2\right)\right) \quad (14)$$

for all $t \leq T$. Combining this with the first and second part we conclude that there indeed exists $\kappa > 0$ such that

$$\sup_{t\leq T} \|\dot{\theta}(t)\| \leq \kappa \quad (15)$$

with probability $\geq 1-\epsilon$.

With Theorem 5.1 established, it remains to prove convergence of the loss in finite time, for wide neural networks. To this end, we derive the evolution of $\mathcal{L}(t)$.

Lemma 5.2. ***(Time evolution of the loss)*** *The loss $\mathcal{L}(t)$ evolves over time as*

$$\frac{\partial}{\partial t}\mathcal{L}(t) = -\frac{4\mathcal{L}(t)}{N^2} \cdot \begin{pmatrix} f(\mathcal{X}_+) \\ f(\mathcal{X}) \end{pmatrix}^\top \boldsymbol{K}(t) \begin{pmatrix} f(\mathcal{X}_+) \\ f(\mathcal{X}) \end{pmatrix} \quad (16)$$

where we define

$$\boldsymbol{K}(t) = \begin{pmatrix} K_\theta(\mathcal{X},\mathcal{X}) & K_\theta(\mathcal{X},\mathcal{X}_+) \\ K_\theta(\mathcal{X}_+,\mathcal{X}) & K_\theta(\mathcal{X}_+,\mathcal{X}_+) \end{pmatrix} \quad (17)$$

which is just the kernel matrix of $(\mathcal{X}_+,\mathcal{X})$ at $\theta(t)$.

The proof is included in Appendix B.2. We recognize this as a non-autonomous ODE, linear in $\mathcal{L}$ but multiplied with a time-dependent scalar function

$$g(t) = \frac{4}{N^2}\begin{pmatrix} f(\mathcal{X}_+) \\ f(\mathcal{X}) \end{pmatrix}^\top \boldsymbol{K}(t) \begin{pmatrix} f(\mathcal{X}_+) \\ f(\mathcal{X}) \end{pmatrix} \quad (18)$$

on the right side. Suppose we could find a constant $\eta > 0$ such that $\forall t \leq T : g(t) > \eta$. Then, the solution to the non-autonomous ODE for $\mathcal{L}(t)$ can be upper bounded by the solution to the autonomous ODE

$$\begin{aligned} \dot{L} &= -\eta L \\ L(0) &= 1-\rho \end{aligned} \quad (19)$$

because $\mathcal{L}(0) < L(0)$ and $\dot{\mathcal{L}} < \dot{L} < 0$. Since $L(t)$ certainly converges to zero up to our previously fixed $\delta > 0$ up to T as defined in Definition 4.3, so does $\mathcal{L}(t)$.

Theorem 5.3. ***(Convergence of the loss in finite time)*** *Under the setting in Definition 4.3, there exists $M_1 \in \mathbb{N}$ such that, for all networks of width $M \geq M_1$, the loss under gradient flow satisfies $\frac{\partial}{\partial t}\mathcal{L}(t) < -\eta\mathcal{L}(t)$ for all $t \leq T$, by virtue of $g(t) > \eta$ for all $t \leq T$. This implies that $\mathcal{L}(T) < \delta$.*

The proof is included in Appendix B.3. Essentially, it proceeds as follows: Firstly, the kernel matrix $\boldsymbol{K}(t)$ does not change much up to time T, provided M is large. This is due to Theorem 5.1, which guarantees that the weights remain in a bounded ball of radius R around their initialization (until time T), and Theorem 3.1, which then bounds the change in the NTK matrix as $\mathcal{O}(\frac{R^2}{\sqrt{M}})$. Therefore, the smallest eigenvalue of $\boldsymbol{K}(t)$ stays larger than $\lambda/2$ when M is sufficiently large. Thus, for all $t \leq T$, it holds that

$$\begin{aligned} g(t) &> \frac{2\lambda}{N^2}\left\|\begin{pmatrix} f(\mathcal{X}) \\ f(\mathcal{X}_+) \end{pmatrix}\right\|^2 \\ &\geq \frac{2\lambda}{N^2}\left(\begin{pmatrix} f(\mathcal{X}) \\ f(\mathcal{X}_+) \end{pmatrix}\right)^\top \left(\begin{pmatrix} f(\mathcal{X}_+) \\ f(\mathcal{X}) \end{pmatrix}\right) \\ &= \frac{4C(t)\lambda}{N} \end{aligned} \quad (20)$$

where we used Cauchy-Schwartz inequality in the second line, and plugged in

$$C(t) = \frac{1}{2N}\begin{pmatrix} f(\mathcal{X}) \\ f(\mathcal{X}_+) \end{pmatrix}^\top \begin{pmatrix} f(\mathcal{X}_+) \\ f(\mathcal{X}) \end{pmatrix} \quad (21)$$

in the third line. Finally, gradient flow ensures that $C(t) \geq 1 - \sqrt{\mathcal{L}(0)} \geq 1 - \sqrt{1-\rho}$ for all t, giving the desired lower bound on $g(t)$.

6 IMPLICATIONS FOR THE GENERALIZATION ERROR OF BARLOW TWINS

Having established the validity of the NTK approximation for neural networks trained under the Barlow Twins loss, we leverage our newly found connection into generalization error bounds for (finite) neural networks. We wish to bound $\bar{\mathcal{L}}(f) = \|\bar{C}(f) - I\|_F^2$ where the population cross-moment matrix is

$$\bar{C}(f) = \mathbb{E}_{x,x^+}\left[\frac{f(x)f(x^+)^\top + f(x^+)f(x)^\top}{2}\right] \tag{22}$$

Here, f denotes a neural network with M neurons, and p weights collected in a vector $\theta \in \mathbb{R}^p$. Whenever M is large enough to ensure Corollary 4.5 kicks in, we know that with probability at least $1-\epsilon$

$$\begin{aligned} f_k(x;\theta_T) =& f_k(x;\theta_0) + \langle \theta_T - \theta_0, \nabla_\theta f_k(x;\theta_0)\rangle \\ &+ \zeta \end{aligned} \tag{23}$$

which is simply the first-order Taylor approximation of f_k in parameter space. Note that $\zeta = \mathcal{O}(\frac{R^3}{\sqrt{M}})$ because $\|\theta_T - \theta_0\| \leq R$ and the Hessian scales as $\mathcal{O}(\frac{R}{\sqrt{M}})$. Therefore, the trained neural network $f_k(x;\theta_T)$ can be approximated by a kernel model

$$\begin{aligned} g_k(x) &= \begin{pmatrix} 1 \\ \theta_T - \theta_0 \end{pmatrix}^\top \begin{pmatrix} f_k(x;\theta_0) \\ \nabla_\theta f_k(x;\theta_0)\rangle \end{pmatrix} \\ &=: (\theta'_T - \theta_0)^\top \psi_k(x) \end{aligned} \tag{24}$$

up to some small positive ζ. In this section, we therefore derive generalization error bounds for kernel versions of Barlow Twins.

Generalization error for Barlow Twins in Hilbert spaces. We first place ourselves in an abstract Hilbert space framework. Denoting z, z^+ for positive pairs residing in some Hilbert space $\mathcal{H}$, we wish to learn a bounded linear operator $W : \mathcal{H} \to \mathbb{R}^K$ such as to minimize

$$\begin{aligned} &\bar{\mathcal{L}}(W) = \left\|W\bar{\Gamma}W^* - I_K\right\|_F^2 \text{ , where} \\ &\bar{\Gamma} = \frac{1}{2}\mathbb{E}_{z,z^+}\left[z(z^+)^* + z^+z^*\right] \end{aligned} \tag{25}$$

Here $*$ denotes the adjoint, $\|\cdot\|_{\text{HS}}$ is the Hilbert-Schmidt norm of linear operators on $\mathcal{H}$ (in the finite-dimensional setting, the Frobenius norm $\|\cdot\|_F$), and $\|\cdot\|$ is the operator norm.

Theorem 6.1. ***(Generalization error for Barlow Twins in Hilbert spaces)*** *Let $N \in \mathbb{N}$ and $N' \leq N$. Assume that pairs (z, z^+) are almost surely contained in a ball of radius $S > 0$ in a Hilbert space $\mathcal{H}$. Moreover, assume that for any set of training data, $\mathcal{L}(W) \leq \delta$ for some $\|W\| \leq B$. Then, with probability $\geq 1 - 2\epsilon$, it holds that*

$$\begin{aligned} \bar{\mathcal{L}}(W) \leq 3\delta + \frac{3B^4}{N}\left(\hat{\boldsymbol{V}} + \exp\left(-\frac{(N'-1)^2\epsilon^2}{8S^8N'}\right)\right) \\ + 3\exp\left(-\frac{N\epsilon^2}{B^4S^4}\right) \end{aligned} \tag{26}$$

where we define

$$\begin{aligned} \hat{\boldsymbol{V}} &= \frac{1}{N'(N'-1)} \sum_{\substack{i<j \\ i,j\in[N']}} \|\Gamma_i - \Gamma_j\|_{HS}^2 \\ \Gamma_i &= \frac{1}{2}\left(z_i(z_i^+)^* + z_i^+ z_i^*\right) \end{aligned} \tag{27}$$

and the randomness is over independently sampled positive pairs $(z_1, z_1^+), \ldots, (z_N, z_N^+)$.

The proof relies on McDiarmid's inequality (McDiarmid et al., 1989) and is included in Appendix D.1. The quantity $\hat{\boldsymbol{V}}$ is purely empirical, and may be estimated from fewer samples by choosing $N' < N$. Of course, estimation is a strong word here — in an abstract Hilbert space setting, we may not be able to compute $\|\Gamma_i - \Gamma_j\|_{\text{HS}}^2$. However, when the inner product in $\mathcal{H}$ takes the form of a kernel, $\hat{\boldsymbol{V}}$ can be expressed more explicitly.

Connecting to neural networks. Such is the case for the NTK approximation of the neural network (24). Defining $\psi(x) = (\psi_1(x), \ldots \psi_K(x))$ as the concatenation of all feature maps, we obtain the kernel inner product

$$\hat{K}_{\theta_0}(x, x') = \langle \psi(x), \psi(x')\rangle = \sum_{k=1}^K \psi_k(x)\psi_k(x') \tag{28}$$

Then, defining

$$W = \begin{pmatrix} (\theta'_T - \theta'_0)^\top & 0 & \ldots \\ 0 & (\theta'_T - \theta'_0)^\top & \ldots \\ \ldots & \ldots & \ldots \end{pmatrix} \in \mathbb{R}^{K \times K(p+1)} \tag{29}$$

we obtain $g(x) = W\psi(x)$. Actually, multivariate functions such as g reside in a vector-valued RKHS. This is also clear from the fact that the NTK in the multivariate setting is a matrix-valued kernel K_{θ_0}. In a vector-valued RKHS, we do not have feature maps, but rather feature operators ψ. For the neural network, the feature operator is the Jacobian. Under the trace inner product $\langle \psi(x), \psi(x')\rangle := \text{Trace}(\psi(x)^*\psi(x')) = \text{Trace}(K_{\theta_0}(x, x'))$ we see that this construction is identical to the one given above, in that it gives the same

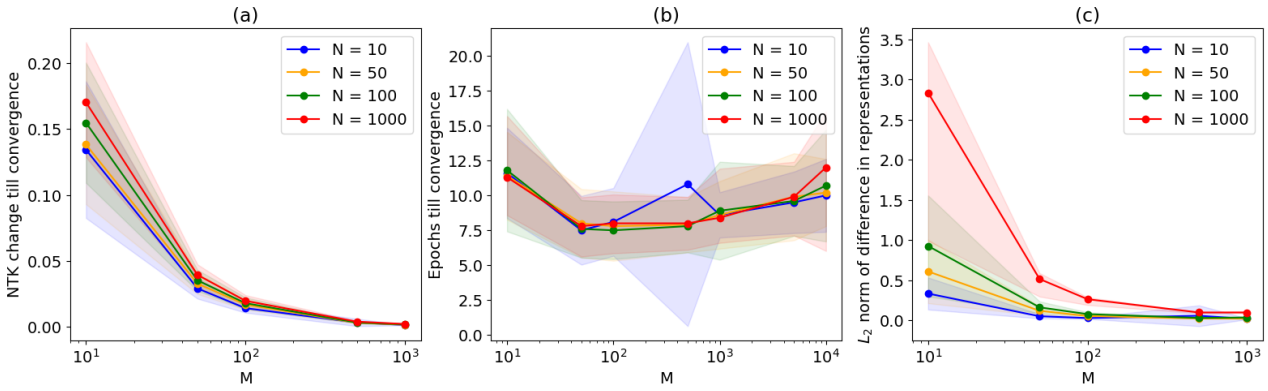

Figure 1: For a fixed sample size N, we plot different quantities for varying network width M. We then vary N and plot: (a) NTK change till convergence (b) Training Epochs till convergence (c) Squared norm of difference between representations of neural network and corresponding kernel model

inner product between $\psi(x), \psi(x')$. By virtue of the kernel trick, we can now estimate $\hat{\boldsymbol{V}}$ without moving into the Hilbert space.

Lemma 6.2. ***(Estimating $\hat{\boldsymbol{V}}$ with kernels)*** *Consider the setting of Theorem 6.1. Suppose that there exists a feature map $\psi : \mathbb{R}^d \to \mathcal{H}$ such that $z_n = \psi(x_n)$ and $z_n^+ = \psi(x_n^+)$. Moreover, assume there exists a kernel $K_{\theta_0} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ with $\langle \psi(x), \psi(x') \rangle = K_{\theta_0}(x, x')$ for all $x, x' \in \mathbb{R}^d$. Then, for all $i \neq j$, it holds that*

$$\begin{aligned} &\|\Gamma_i - \Gamma_j\|_{HS}^2 = \\ &0.5\left(K_{\theta_0}(x_i, x_i)K_{\theta_0}(x_i^+, x_i^+) + K_{\theta_0}(x_i, x_i^+)^2\right) + \\ &0.5\left(K_{\theta_0}(x_j, x_j)K_{\theta_0}(x_j^+, x_j^+) + K_{\theta_0}(x_j, x_j^+)^2\right) - \\ &K_{\theta_0}(x_i, x_j)K_{\theta_0}(x_i^+, x_j^+) - K_{\theta_0}(x_i, x_j^+)K_{\theta_0}(x_i^+, x_j) \end{aligned} \tag{30}$$

This result is proven in Appendix D.2. Furthermore, combining Theorem 6.1 with the fact that the neural network f can be approximated by a kernel model (24), we immediately obtain the following corollary.

Corollary 6.3. *Fix $\epsilon > 0$. Assume f is a neural with M neurons in the hidden layer, where M is large enough for Corollary 4.5 to hold. Suppose f is trained until the Barlow Twins loss is smaller than δ. Then, with probability at least $1 - 3\epsilon$, it holds that*

$$\bar{\mathcal{L}}(f) \leq 2\nu(N, \epsilon, \delta) + 8K^2\zeta^2(2BS + \zeta)^2 \tag{31}$$

where $\nu(N, \epsilon, \delta)$ is the slack term from Theorem 6.1. B can be taken as the norm of $\|\theta'_T - \theta'_0\| \leq R + 1$, and $S = \sup_{x \in \mathbb{R}^d} \|\psi(x)\|$.

The proof is included in Appendix D.3. It uses the fact that f is approximated up to ζ by a kernel model with probability $\geq 1 - \epsilon$. This allows bounding the difference between the loss of f and the loss of g. Then, we use Theorem 6.1, which holds with probability at least $1 - 2\epsilon$.

Remark 6.4. *Our analysis only bounds the pretraining loss. However, this can be passed on to guarantees on the classification error on downstream tasks, provided the augmentations elucidate enough of the underlying class structures. For example, see Section 5.2 of Cabannes et al. (2023).*

7 EXPERIMENTS

In this Section, we verify our theoretical claims on the MNIST dataset (Deng, 2012). Optimization is done using gradient descent with a learning rate of 0.5. The threshold for loss convergence δ is set at 10^{-5}. We use a single-hidden layer neural network with tanh activation unless stated otherwise. All experiments are run 10 times with different random seeds; their means along with standard deviations are plotted in the figures.

We first verify near-constancy of the NTK for Barlow Twins at large width, with one-dimensional embeddings. We do not restrict ourselves to the setting of $\mathcal{L}(0) < 1$. Varying the sample size from 10 to 1000, we plot the norm of the NTK deviation by varying the hidden layer width M from 10 to 10000. Recall that in our theoretical results, η depends on the sample size N, hence does T, hence does R and hence does M. This suggests that as N grows, a larger width is necessary for the kernel regime to kick in. However, as Figure 1 (a) clearly displays, the neural network enters the kernel regime irrespective of the sample size N as M grows, with the change in the NTK near zero at 10^3 neurons in the hidden layer.

Next, we look at the number of training epochs required for the loss to converge below δ. As per Theorem 5.3, we expect this to be independent of the width of the neural network, provided the network is suffi-

ciently large. This is empirically verified in Figure 1 (b).

With the NTK nearly constant at large width, we expect the representations learned by the finite neural network to be close to those learned by a corresponding kernel model, via gradient descent under the NTK at initialization $\boldsymbol{K}(0)$. We verify this in Figure 1 (c), where we see that irrespective of the number of samples, the representations at convergence are closer for wider neural networks Additional experiments are included in Appendix G. There, we also check the variation of these quantities for networks with ReLU activation, and include results for higher-dimensional embeddings.

8 DISCUSSION

In this paper, we connect self-supervised learning with neural networks to the neural tangent kernel regime. We prove that at infinite width, the NTK of a neural network trained under the Barlow Twins loss becomes constant. This is the first result that rigorously justifies the use of traditional kernel methods to understand SSL through the lenses of the NTK. Furthermore, the kernel connection enables us to bound the population Barlow Twins loss in terms of the empirical loss and a slack term that decays with the number of samples. This opens a number of interesting avenues and challenges for future work.

Extension to deep networks, and other losses. It is desirable to extend the convergence results to other, in particular deep, architectures. This requires an extension of Theorem 4.1 to multiple layers. In addition, verifying the validity of the NTK approximation for other commonly used non-contrastive loss functions such as VIC-Reg (Bardes et al., 2021) is a natural next step. Moreover, our proof for the ReLU activations is not entirely rigorous, since $\frac{\partial \mathcal{L}}{\partial \theta}$ can only be understood in the weak sense. Finally, as discussed earlier, we believe the condition $\mathcal{L}(0) < 1$ can be dropped, but a proof remains to be established.

Improving generalization error bounds. Theorem 6.1 still relies on uniform bounds, despite closed-form expressions for kernel versions of Barlow Twins being available. To be precise, Simon et al. (2023) show that the minimum norm W^* that achieves zero loss lies in the top eigenspace of the cross-moment operator Γ. This W^* is referred to as the *spectral solution*. We believe that spectral perturbation bounds provide tighter guarantees on the generalization error.

Does gradient flow approach the spectral solution? Even with improved generalization bounds for the spectral solution W^* (as noted above), there is still a missing link: It is not yet known whether gradient flow actually approaches the spectral solution for linearized networks. Simon et al. (2023) prove it for "aligned" initialization (starting in the top eigenspace), and give heuristics for why it would also hold under "small" initialization.

Conclusion. Our work establishes the significance of the NTK to self-supervised learning, and confirms numerous recent works that have headed into this direction. As Belkin et al. (2018) put it, *to understand deep learning, we need to understand kernel learning.* In light of our results, we are tempted to add: To understand self-supervised learning, we need to understand representation learning with kernels.

Acknowledgements

This paper is supported by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research. We also thank the anonymous reviewers.

References

Anil, G. G., Esser, P., and Ghoshdastidar, D. (2024). When can we approximate wide contrastive models with neural tangent kernels and principal component analysis? *arXiv preprint arXiv:2403.08673*.

Arora, S., Khandeparkar, H., Khodak, M., Plevrakis, O., and Saunshi, N. (2019). A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*.

Bachman, P., Hjelm, R. D., and Buchwalter, W. (2019). Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32.

Bardes, A., Ponce, J., and LeCun, Y. (2021). Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*.

Belkin, M., Ma, S., and Mandal, S. (2018). To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, pages 541–549. PMLR.

Blanchard, G., Bousquet, O., and Zwald, L. (2007). Statistical properties of kernel principal component analysis. *Machine Learning*, 66:259–294.

Bombari, S., Kiyani, S., and Mondelli, M. (2023). Beyond the universal law of robustness: Sharper laws for random features and neural tangent kernels. In *International Conference on Machine Learning*, pages 2738–2776. PMLR.

Cabannes, V., Kiani, B., Balestriero, R., LeCun, Y., and Bietti, A. (2023). The ssl interplay: Augmentations, inductive bias, and generalization. In *International Conference on Machine Learning*, pages 3252–3298. PMLR.

Chizat, L., Oyallon, E., and Bach, F. (2019). On lazy training in differentiable programming. *Advances in neural information processing systems*, 32.

Deng, L. (2012). The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142.

Esser, P., Fleissner, M., and Ghoshdastidar, D. (2024). Non-parametric representation learning with kernels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11910–11918.

HaoChen, J. Z., Wei, C., Gaidon, A., and Ma, T. (2021). Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011.

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31.

Johnson, D. D., El Hanchi, A., and Maddison, C. J. (2022). Contrastive learning can find an optimal basis for approximately view-invariant functions. In *The Eleventh International Conference on Learning Representations*.

Kiani, B. T., Balestriero, R., Chen, Y., Lloyd, S., and LeCun, Y. (2022). Joint embedding self-supervised learning in the kernel regime. *arXiv preprint arXiv:2209.14884*.

Lee, J., Xiao, L., Schoenholz, S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. (2019). Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32.

Liu, C., Zhu, L., and Belkin, M. (2020a). On the linearity of large non-linear models: when and why the tangent kernel is constant. *Advances in Neural Information Processing Systems*, 33:15954–15964.

Liu, C., Zhu, L., and Belkin, M. (2020b). Toward a theory of optimization for over-parameterized systems of non-linear equations: the lessons of deep learning. *arXiv preprint arXiv:2003.00307*, 7.

Mallinar, N., Simon, J., Abedsoltan, A., Pandit, P., Belkin, M., and Nakkiran, P. (2022). Benign, tempered, or catastrophic: Toward a refined taxonomy of overfitting. *Advances in Neural Information Processing Systems*, 35:1182–1195.

McDiarmid, C. et al. (1989). On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Schölkopf, B., Smola, A., and Müller, K.-R. (1997). Kernel principal component analysis. In *International conference on artificial neural networks*, pages 583–588. Springer.

Simon, J. B., Dickens, M., Karkada, D., and DeWeese, M. R. (2021). The eigenlearning framework: A conservation law perspective on kernel regression and wide neural networks. *arXiv preprint arXiv:2110.03922*.

Simon, J. B., Knutins, M., Ziyin, L., Geisz, D., Fetterman, A. J., and Albrecht, J. (2023). On the stepwise nature of self-supervised learning. In *International Conference on Machine Learning*, pages 31852–31876. PMLR.

Smola, A. J. and Schölkopf, B. (1998). *Learning with kernels*, volume 4. Citeseer.

Wei, C., Shen, K., Chen, Y., and Ma, T. (2020). Theoretical analysis of self-training with deep networks on unlabeled data. *arXiv preprint arXiv:2010.03622*.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. (2021). Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pages 12310–12320. PMLR.

Ziyin, L., Lubana, E. S., Ueda, M., and Tanaka, H. (2022). What shapes the loss landscape of self-supervised learning? *arXiv preprint arXiv:2210.00638*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**/No/Not Applicable]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [**Yes**/No/Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [**Yes, we add a link to the code** /No/Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [**Yes**/No/Not Applicable]
   (b) Complete proofs of all theoretical results. [**Yes**/No/Not Applicable]
   (c) Clear explanations of any assumptions. [**Yes**/No/Not Applicable]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [**Yes, we add a link to the code**/No/Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [**Yes, see appendix** /No/Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [**Yes**/No/Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [**Yes**/No/Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [**Yes**/No/Not Applicable]
   (b) The license information of the assets, if applicable. [Yes/No/**Not Applicable**]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/**Not Applicable**]
   (d) Information about consent from data providers/curators. [Yes/No/**Not Applicable**]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/**Not Applicable**]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Yes/No/**Not Applicable**]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/**Not Applicable**]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/**Not Applicable**]

Supplementary Materials

A PROOFS FROM SECTION 4

Appendix B contains a more detailed and computationally less involved derivation of all results for the one-dimensional case. It may be convenient to refer to it first. The proof strategy is identical.

A.1 Proof of Theorem 4.1

Proof. Recall that $T > 0$ is a fixed, width-independent point in time. We show that there exists $\kappa > 0$ such that $\sup_{t \leq T} \|\dot{\theta}\| \leq \kappa$ for all networks of width $M > 8T$, with high probability over randomly initialized weights. Specifically, for any $\epsilon > 0$, we know that there exists $\kappa_1 > 0$ such that

$$\frac{\|\theta(0)\|^2}{M} \leq \kappa_1 - 1 \tag{32}$$

with probability at least $1 - \epsilon$, because the weights are independent Gaussians at initialization. We condition everything on this event. In this proof, we also assume $\mathcal{L}(0) < 1$ under said event. We need this condition anyway for Theorem 4.4 and therefore include it into our bounds here as well to simplify matters. However, any width-independent bound on $\mathcal{L}(0)$ is sufficient. Denote $c_\phi = \max_{t \in \mathbb{R}} |\phi(t)|$ and $c_{\phi'} = \max_{t \in \mathbb{R}} \left|\frac{\partial}{\partial t}\phi(t)\right|$. We will show the following three statements.

1. Firstly, we show that

$$\frac{\partial}{\partial t}\left(\sum_{m=1}^{M} \|w_m(t)\|^2\right) \leq 8 \tag{33}$$

for all $t \leq T$. This immediately implies that there exists some $\kappa_1 > 0$ such that for all $t \leq T$

$$\frac{1}{M}\left(\sum_{m=1}^{M} \|w_m(t)\|^2\right) \leq \kappa_1 \tag{34}$$

holds with probability $\geq 1 - \epsilon$ over random initialization, because

$$\frac{1}{M}\left(\sum_{m=1}^{M} \|w_m(t)\|^2\right) \leq \frac{1}{M}\left(\sum_{m=1}^{M} \|w_m(0)\|^2\right) + \frac{8T}{M} \leq \frac{1}{M}\left(\sum_{m=1}^{M} \|w_m(0)\|^2\right) + 1 \tag{35}$$

where we used $M > 8T$ and the fact that

$$\frac{1}{M}\left(\sum_{m=1}^{M} \|w_m(0)\|^2\right) \leq \frac{\|\theta(0)\|^2}{M} \leq \kappa_1 - 1 \tag{36}$$

with probability at least $1 - \epsilon$.

2. Using part 1, we bound the maximum squared value that any representation takes until time T, that is

$$|f|^2 = \max_{n \in [N]} \sup_{t \leq T} \max\left(\|f(x_n)\|^2, \|f(x_n^+)\|^2\right) \tag{37}$$

by some $\kappa_2 > 0$ (again independent of M).

3. We combine both statements to show that there exists $\kappa > 0$ such that $\sup_{t \leq T} \|\dot{\theta}\| \leq \kappa$ with probability at least $1 - \epsilon$.

Part 1. First of all, we derive the evolution of all weights over time. Using the matrix chain rule,

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial w_{mk}} &= \operatorname{Trace}\left(\left(\frac{\partial \mathcal{L}}{\partial C}\right)^\top \frac{\partial C}{\partial w_{mk}}\right) \\ &= 2 \operatorname{Trace}\left((C - I)\left(\frac{1}{2N}\frac{\partial}{\partial w_{mk}}\left(\sum_{n=1}^{N} f(x_n) f(x_n^+)^\top + f(x_n^+) f(x_n)^\top\right)\right)\right) \end{aligned} \tag{38}$$

and similarly,

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial v_{mr}} &= \operatorname{Trace}\left(\left(\frac{\partial \mathcal{L}}{\partial C}\right)^\top \frac{\partial C}{\partial v_{mr}}\right) \\ &= 2\operatorname{Trace}\left((C-I)\left(\frac{1}{2N}\frac{\partial}{\partial v_{mr}}\left(\sum_{n=1}^N f(x_n)f(x_n^+)^\top + f(x_n^+)f(x_n)^\top\right)\right)\right)\end{aligned} \tag{39}$$

For any $n \in [N]$ and any $i, j, k \in [K]$

$$\begin{aligned}&\frac{\partial}{\partial w_{mk}}\left(f_i(x_n)f_j(x_n^+) + f_j(x_n)f_i(x_n^+)\right) \\ &= \begin{cases} 2f_k(x_n^+)\frac{1}{\sqrt{M}}\phi(v_m^\top x_n) + 2f_k(x_n)\frac{1}{\sqrt{M}}\phi(v_m^\top x_n^+), \text{ if } k=i=j \\ f_j(x_n^+)\frac{1}{\sqrt{M}}\phi(v_m^\top x_n) + f_j(x_n)\frac{1}{\sqrt{M}}\phi(v_m^\top x_n^+), \text{ if } k=i\neq j \\ f_i(x_n^+)\frac{1}{\sqrt{M}}\phi(v_m^\top x_n) + f_i(x_n)\frac{1}{\sqrt{M}}\phi(v_m^\top x_n^+), \text{ if } k=j\neq i \\ 0, \text{ if } k\notin\{i,j\}\end{cases}\end{aligned} \tag{40}$$

Equating $\frac{\partial}{\partial t}w_{mk} = -\frac{\partial \mathcal{L}}{\partial w_{mk}}$ gives

$$\begin{aligned}\frac{\partial}{\partial t}\left(\sum_{m=1}^M \|w_m\|^2\right) &= 2\sum_{m=1}^M\sum_{k=1}^K w_{mk}\left(\frac{\partial}{\partial t}w_{mk}\right) \\ &= 4\operatorname{Trace}\left((I-C)\left(\frac{1}{2N}\sum_{n=1}^N\sum_{k=1}^K A_{k,i,j}(n)\right)_{i,j=1}^K\right)\end{aligned} \tag{41}$$

where we define

$$\begin{aligned}A_{k,i,j}(n) &= \sum_{m=1}^M w_{mk}\cdot\frac{\partial}{\partial w_{mk}}\left(f_i(x_n)f_j(x_n^+) + f_i(x_n^+)f_j(x_n)\right) \\ &= \begin{cases} 2f_k(x_n^+)f_k(x_n) + 2f_k(x_n)f_k(x_n^+), \text{ if } k=i=j \\ f_j(x_n^+)f_k(x_n) + f_j(x_n)f_k(x_n^+), \text{ if } k=i\neq j \\ f_i(x_n^+)f_k(x_n) + f_i(x_n)f_k(x_n^+), \text{ if } k=j\neq i \\ 0, \text{ if } k\notin\{i,j\}\end{cases}\end{aligned} \tag{42}$$

Noticing that

$$\sum_{k=1}^K A_{k,i,j}(n) = \begin{cases} 2f_i(x_n^+)f_i(x_n) + 2f_i(x_n)f_i(x_n^+), \text{ if } i=j \\ 2f_i(x_n^+)f_j(x_n) + 2f_j(x_n)f_i(x_n^+), \text{ if } i\neq j\end{cases} \tag{43}$$

we obtain

$$\frac{\partial}{\partial t}\left(\sum_{m=1}^M \|w_m\|^2\right) = 8\operatorname{Trace}\left((I-C)C\right) \tag{44}$$

This is certainly bounded by some $\kappa_0 > 0$, because the loss is non-increasing under gradient flow, which forces all entries of C to stay bounded. In our case, with $\mathcal{L}(0) = \|I-C\|_F^2 < 1$, we may choose $\kappa_0 = 8$.

Regarding the evolution of the first layer weights v_{mr}, we note that for any $n \in [N]$, $i, j \in [K]$ and $r \in [d]$

$$\begin{aligned}
&\frac{\partial}{\partial v_{mr}} \left(f_i(x_n) f_j(x_n^+) + f_j(x_n) f_i(x_n^+) \right) = \\
&\frac{1}{\sqrt{M}} f_j(x_n^+) w_{mj} \phi'(v_m^\top x_n) x_n^{(r)} + \\
&\frac{1}{\sqrt{M}} f_i(x_n) w_{mi} \phi'(v_m^\top x_n^+) \left(x_n^+\right)^{(r)} + \\
&\frac{1}{\sqrt{M}} f_j(x_n) w_{mi} \phi'(v_m^\top x_n^+) \left(x_n^+\right)^{(r)} + \\
&\frac{1}{\sqrt{M}} f_i(x_n^+) w_{mi} \phi'(v_m^\top x_n) x_n^{(r)}
\end{aligned} \tag{45}$$

Thus, we obtain

$$\dot{v_{mr}}(t) = \frac{1}{N} \text{Trace} \left((C - I) \left(\frac{\partial}{\partial v_{mr}} \left(f_i(x_n) f_j(x_n^+) + f_j(x_n) f_i(x_n^+) \right) \right)_{i,j} \right) \tag{46}$$

Part 2. In this part we bound $|f|^2$. For any x and any $k \in [K]$, the k-th output f_k evolves over time as follows.

$$\begin{aligned}
\frac{\partial}{\partial t} f_k(x) &= \langle \frac{\partial f_k(x)}{\partial \theta}, \frac{\partial}{\partial t} \theta \rangle \\
&= - \sum_{m=1}^{M} \frac{\partial f_k(x)}{\partial \theta_m} \frac{\partial \mathcal{L}}{\partial \theta_m} \\
&= - \sum_{m=1}^{M} \frac{\partial f_k(x)}{\partial \theta_m} \sum_{n=1}^{N} \sum_{l=1}^{K} \frac{\partial \mathcal{L}}{\partial f_l(x_n)} \frac{\partial f_l(x_n)}{\partial \theta_m} + \frac{\partial \mathcal{L}}{\partial f_l(x_n^+)} \frac{\partial f_l(x_n^+)}{\partial \theta_m} \\
&= - \sum_{n=1}^{N} \sum_{l=1}^{K} \frac{\partial \mathcal{L}}{\partial f_l(x_n)} \left(K_\theta(x, x_n) \right)_{k,l} + \frac{\partial \mathcal{L}}{\partial f_l(x_n^+)} \left(K_\theta(x, x_n^+ \right)_{k,l}
\end{aligned} \tag{47}$$

So we need to derive all $\frac{\partial \mathcal{L}}{\partial f_k(x_i)}$. For any $k \in [K]$, and any $i \in [N]$, we see that $\frac{\partial C}{\partial f_k(x_i)} = \frac{1}{2N} A(i,k)$, where we define the matrix

$$A(i,k) = \begin{pmatrix}
0 & \dots & 0 & f_1(x_i^+) & 0 & \dots & 0 \\
0 & \dots & 0 & f_2(x_i^+) & 0 & \dots & 0 \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots \\
f_1(x_i^+) & f_2(x_i^+) & \dots & 2f_k(x_i^+) & \dots & \dots & f_K(x_i^+) \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots \\
0 & \dots & 0 & f_{K-1}(x_i^+) & 0 & \dots & 0 \\
0 & \dots & 0 & f_K(x_i^+) & 0 & \dots & 0
\end{pmatrix} \tag{48}$$

and also let

$$A(i) = \begin{pmatrix}
f_1(x_i^+) & \dots & f_1(x_i^+) \\
\dots & \dots & \dots \\
f_K(x_i^+) & \dots & f_K(x_i^+)
\end{pmatrix} \in \mathbb{R}^{K \times K} \tag{49}$$

Notice that $A(i,k) = e_k e_k^\top A(i)^\top + A(i) e_k e_k^\top$. Using the matrix chain rule, and the fact that $\frac{\partial \mathcal{L}}{\partial C} = 2(C - I)$, we

therefore obtain

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial f_k(x_i)} &= 2\,\mathrm{Trace}\left((C-I)^\top \frac{\partial C}{\partial f_k(x_i)}\right) \\
&= \frac{1}{N}\,\mathrm{Trace}\left((C-I)A(i,k)\right) \\
&= \frac{1}{N}\left(\sum_{l=1}^{K}(C-I)_{l\bullet}A(i,k)_{\bullet l}\right) \\
&= \frac{1}{N}\left(\sum_{l\neq k}(C-I)_{lk}f_l(x_i^+) + (C-I)_{\bullet k}^\top f(x_i^+)\right) \\
&= \frac{1}{N}\left(\sum_{l\neq k}C_{lk}f_l(x_i^+) + (C-I)_{\bullet k}^\top f(x_i^+)\right) \\
&= \frac{2}{N}\left(C_{1k}f_1(x_i^+) + C_{2k}f_2(x_i^+) + \cdots + (C_{kk}-1)f_k(x_i^+) + \cdots + C_{Kk}f_K(x_i^+)\right) \\
&= \frac{2}{N}(C-I)_{k\bullet}f(x_i^+) \\
&= \frac{2}{N}e_k^\top(C-I)f(x_i^+)
\end{aligned} \tag{50}$$

where we exploited the symmetry of $A(i,k)$ in the sixth step. Overall, we have

$$\begin{aligned}
\frac{\partial}{\partial t}f_k(x) = -\frac{2}{N}\sum_{n=1}^{N}\sum_{l=1}^{K} & K_\theta(x,x_n)_{k,l}\cdot e_l^\top(C-I)f(x_n^+) + \\
& K_\theta(x,x_n^+)_{k,l}\cdot e_l^\top(C-I)f(x_n) \\
= -\frac{2}{N}\sum_{l=1}^{K} & K_\theta(x,[\mathcal{X},\mathcal{X}_+])_{k,l}\cdot \boldsymbol{F}([\mathcal{X}_+,\mathcal{X}])^\top(C-I)e_l
\end{aligned} \tag{51}$$

where we write $\boldsymbol{F}([\mathcal{X}_+,\mathcal{X}]) \in \mathbb{R}^{K\times 2N}$ for the matrix that contains the representations at time t as columns. Moreover, with $\boldsymbol{K}_{k,l}$ the kernel matrix of $\mathcal{X},\mathcal{X}_+$ at output k,l, we have

$$\frac{\partial}{\partial t}f_k([\mathcal{X},\mathcal{X}_+]) = \frac{2}{N}\sum_{l=1}^{K}\boldsymbol{K}_{k,l}\cdot\boldsymbol{F}([\mathcal{X}_+,\mathcal{X}])^\top(I-C)e_l \tag{52}$$

Therefore, the squared norm of the representations evolve as

$$\begin{aligned}
\frac{\partial}{\partial t}\left(\sum_{k=1}^{K}\|f_k([\mathcal{X},\mathcal{X}_+])\|^2\right) &= 2\sum_{k=1}^{K}\left(f_k([\mathcal{X},\mathcal{X}_+])\right)^\top\left(\frac{\partial}{\partial t}f_k([\mathcal{X},\mathcal{X}_+])\right) \\
&= \frac{4}{N}\sum_{k=1}^{K}\left(f_k([\mathcal{X},\mathcal{X}_+])\right)^\top\sum_{l=1}^{K}\boldsymbol{K}_{k,l}\cdot\boldsymbol{F}([\mathcal{X}_+,\mathcal{X}])^\top(C-I)e_l \\
&\leq \frac{4}{N}\sum_{k,l=1}^{K}\|f_k([\mathcal{X},\mathcal{X}_+])\|\left\|\boldsymbol{K}_{k,l}\boldsymbol{F}([\mathcal{X}_+,\mathcal{X}])^\top(C-I)e_l\right\| \\
&\leq \frac{4}{N}\sum_{k,l=1}^{K}\|f_k([\mathcal{X},\mathcal{X}_+])\|\cdot\|\boldsymbol{K}_{k,l}\|_2\cdot\left\|\boldsymbol{F}([\mathcal{X}_+,\mathcal{X}])^\top(C-I)e_l\right\| \\
&\leq \frac{4}{N}\sum_{k,l=1}^{K}\|f_k([\mathcal{X},\mathcal{X}_+])\|\cdot\|\boldsymbol{K}_{k,l}\|_2\cdot\|\boldsymbol{F}([\mathcal{X}_+,\mathcal{X}])\|_2
\end{aligned} \tag{53}$$

where we have used Cauchy-Schwartz inequality, and the fact that $\|I - C\|_2 \leq 1$ due to $\mathcal{L}(0) < 1$ in the last line. Continuing,

$$\begin{aligned}
&\frac{4}{N}\sum_{k,l=1}^{K} \|f_k([\mathcal{X}, \mathcal{X}_+])\| \cdot \|\boldsymbol{K}_{k,l}\|_2 \cdot \|\boldsymbol{F}([\mathcal{X}_+, \mathcal{X}])\|_2 \leq \\
&\frac{4}{N}\max_{k,l}\|\boldsymbol{K}_{k,l}\|_2 \cdot \left(\sum_{k=1}^{K}\|f_k([\mathcal{X}, \mathcal{X}_+])\| \left(\sum_{l=1}^{K}\|f_l([\mathcal{X}, \mathcal{X}_+])\|^2\right)^{1/2}\right) \leq \\
&\frac{4}{N}\max_{k,l}\|\boldsymbol{K}_{k,l}\|_2 \cdot \left(\sum_{k,l=1}^{K}\|f_k([\mathcal{X}, \mathcal{X}_+])\| \cdot \|f_l([\mathcal{X}, \mathcal{X}_+])\|\right) \leq \\
&\frac{4K}{N}\max_{k,l}\|\boldsymbol{K}_{k,l}\|_2 \cdot \left(\sum_{k=1}^{K}\|f_k([\mathcal{X}, \mathcal{X}_+])\|^2\right)
\end{aligned} \tag{54}$$

where in the second line we bound the spectral norm of $\boldsymbol{F}([\mathcal{X}_+, \mathcal{X}])$ by its Frobenius norm, in the third line we bound the 2-norm by the 1-norm, and in the fourth line we use Cauchy-Schwartz again, which introduces a factor of K. We proceed to bound the spectral norm of the kernel matrices $\boldsymbol{K}_{k,l}$.

Note that for any k, l, and any $(a, b) \in [\mathcal{X}, \mathcal{X}_+] \times [\mathcal{X}, \mathcal{X}_+]$, we have

$$\begin{aligned}
K_{k,l}(a,b) &= \left(\frac{\partial f_k(a)}{\partial \theta}\right)^\top \left(\frac{\partial f_l(a)}{\partial \theta}\right) \\
&= \frac{1}{M}\left(\sum_{m=1}^{M} \mathbf{1}_{k=l} \cdot \phi(v_m^\top a)\phi(v_m^\top b)\right) + \frac{1}{M}\left(\sum_{m=1}^{M} w_{mk}w_{ml}\phi'(v_m^\top a)\phi'(v_m^\top b)a^\top b\right) \\
&\leq c_\phi^2 + \frac{1}{M}\left(\sum_{m=1}^{M}|w_{mk}w_{ml}| \cdot c_{\phi'}^2\right) \\
&\leq c_\phi^2 + \frac{c_{\phi'}^2}{M} \cdot \left(\sum_{m=1}^{M} w_{mk}^2 + w_{ml}^2\right) \\
&\leq c_\phi^2 + c_{\phi'}^2\kappa_1
\end{aligned} \tag{55}$$

where the final line uses Part 1 of this proof, which bounds $\frac{1}{M}\sum_{m=1}^{M}\|w_m(t)\|^2$ uniformly in time by κ_1. Also, recall that data lies on the unit sphere, so $a^\top b \leq 1$. Overall, we see that the spectral norm of each kernel matrix $\boldsymbol{K}_{k,l}$ is bounded by

$$\|\boldsymbol{K}_{k,l}\|_2 \leq 2N\left(c_\phi^2 + c_{\phi'}^2\kappa_1\right) \tag{56}$$

Plugging this in back into our evolution of the representations,

$$\frac{\partial}{\partial t}\sum_{k=1}^{K}\|f_k([\mathcal{X}, \mathcal{X}_+])\|^2 \leq \left(8Kc_\phi^2 + 8Kc_{\phi'}^2\kappa_1\right)\left(\sum_{k=1}^{K}\|f_k([\mathcal{X}, \mathcal{X}_+])\|^2\right) \tag{57}$$

We are thus in the setting of Grönwall's inequality, and obtain

$$\sum_{k=1}^{K}\|f_k([\mathcal{X}, \mathcal{X}_+])\|^2 \leq \left(\sum_{k=1}^{K}\|f_k([\mathcal{X}, \mathcal{X}_+])(0)\|^2\right) \cdot \exp\left(T\left(8Kc_\phi^2 + 8Kc_{\phi'}^2\kappa_1\right)\right) =: \kappa_2 \tag{58}$$

for all time $t \leq T$. When $\frac{\|\theta(0)\|^2}{M} \leq \kappa_1 - 1$ this also gives a bound on $\|f_k([\mathcal{X}, \mathcal{X}_+])(0)\|^2$.

Part 3. We now control $\|\dot{\theta}(t)\|^2$. Using the time derivatives for w_{mk} and v_{mr} derived in Part 1,

$$\begin{aligned}
|\dot{w_{mk}}(t)|^2 &= \frac{1}{N^2}\,\mathrm{Trace}\left((C-I)\left(\sum_{n=1}^{N}\frac{\partial}{\partial w_{mk}}\left(f_i(x_n)f_j(x_n^+)+f_j(x_n)f_i(x_n^+)\right)\right)_{i,j}\right)^2 \\
&\leq \frac{1}{N^2}\left\|\sum_{n=1}^{N}\left(\frac{\partial}{\partial w_{mk}}\left(f_i(x_n)f_j(x_n^+)+f_j(x_n)f_i(x_n^+)\right)\right)_{i,j}\right\|_F^2 \\
&= \mathcal{O}\left(\frac{\kappa_2 c_\phi^2}{M}\right)
\end{aligned} \tag{59}$$

where we used Cauchy-Schwartz for the trace inner product and $\|C-I\|_F<1$ in the second line, and $|f|^2\leq\kappa_2$ from Part 2 in the third line. Summing up,

$$\sum_{m=1}^{M}\|\dot{w_m}(t)\|^2 = \mathcal{O}\left(\kappa_2 c_\phi^2\right) \tag{60}$$

Similarly, we find that

$$\begin{aligned}
|\dot{v_{mr}}(t)|^2 &= \frac{1}{N^2}\,\mathrm{Trace}\left((C-I)\left(\sum_{n=1}^{N}\frac{\partial}{\partial v_{mr}}\left(f_i(x_n)f_j(x_n^+)+f_j(x_n)f_i(x_n^+)\right)\right)_{i,j}\right)^2 \\
&\leq \frac{1}{N^2}\left\|\sum_{n=1}^{N}\left(\frac{\partial}{\partial v_{mr}}\left(f_i(x_n)f_j(x_n^+)+f_j(x_n)f_i(x_n^+)\right)\right)_{i,j}\right\|_F^2 \\
&= \mathcal{O}\left(\frac{\kappa_2\|w_m\|^2 c_{\phi'}^2}{M}\right)
\end{aligned} \tag{61}$$

Summing up and exploiting part 1 to bound $\frac{1}{M}\sum_{m=1}^{M}\|w_m(t)\|^2$ uniformly in time by κ_1, we find that

$$\sum_{m=1}^{M}\|\dot{v_m}\|^2 = \mathcal{O}(\kappa_1\kappa_2 c_{\phi'}^2) \tag{62}$$

Combining both results, we see that there exists $\kappa>0$ such that $\sup_{t\leq T}\|\dot{\theta}(t)\|\leq\kappa$ with probability $\geq 1-\epsilon$ over random initialization. □

A.2 Proof of Lemma 4.2

Proof. Recall that for any x and any $k\in[K]$, the k-th output f_k evolves over time as follows.

$$\begin{aligned}
\frac{\partial}{\partial t}f_k(x) &= \langle\frac{\partial f_k(x)}{\partial\theta},\frac{\partial}{\partial t}\theta\rangle \\
&= -\sum_{m=1}^{M}\frac{\partial f_k(x)}{\partial\theta_m}\frac{\partial\mathcal{L}}{\partial\theta_m} \\
&= -\sum_{m=1}^{M}\frac{\partial f_k(x)}{\partial\theta_m}\sum_{n=1}^{N}\sum_{l=1}^{K}\frac{\partial\mathcal{L}}{\partial f_l(x_n)}\frac{\partial f_l(x_n)}{\partial\theta_m}+\frac{\partial\mathcal{L}}{\partial f_l(x_n^+)}\frac{\partial f_l(x_n^+)}{\partial\theta_m} \\
&= -\sum_{n=1}^{N}\sum_{l=1}^{K}\frac{\partial\mathcal{L}}{\partial f_l(x_n)}\left(K_\theta(x,x_n)\right)_{k,l}+\frac{\partial\mathcal{L}}{\partial f_l(x_n^+)}\left(K_\theta(x,x_n^+\right)_{k,l}
\end{aligned} \tag{63}$$

From here, we write down the time evolution of the loss. We denote $*, +$ for the symbols that refer to an anchor sample $x = x^*$, or an augmentation x^+.

$$\begin{aligned}
\frac{\partial}{\partial t}\mathcal{L}(t) &= \sum_{i=1}^{N} \frac{\partial \mathcal{L}}{\partial f(x_i)} \frac{\partial}{\partial t} f(x_i) + \frac{\partial \mathcal{L}}{\partial f(x_i^+)} \frac{\partial}{\partial t} f(x_i^+) \\
&= -\sum_{i,n=1}^{N} \sum_{k,l=1}^{K} \sum_{\alpha,\beta=*,+} \frac{\partial \mathcal{L}}{\partial f_k(x_i^\alpha)} K_{k,l}(x_i^\alpha, x_n^\beta) \frac{\partial \mathcal{L}}{\partial f_k(x_i^\beta)} \\
&= -\sum_{i,n=1}^{N} \left(\frac{\partial \mathcal{L}}{\partial f(x_i)}\right)^\top K(x_i, x_n) \left(\frac{\partial \mathcal{L}}{\partial f(x_n)}\right) + \\
&\qquad \left(\frac{\partial \mathcal{L}}{\partial f(x_i^+)}\right)^\top K(x_i^+, x_n) \left(\frac{\partial \mathcal{L}}{\partial f(x_n)}\right) + \\
&\qquad \left(\frac{\partial \mathcal{L}}{\partial f(x_i)}\right)^\top K(x_i, x_n^+) \left(\frac{\partial \mathcal{L}}{\partial f(x_n^+)}\right) + \\
&\qquad \left(\frac{\partial \mathcal{L}}{\partial f(x_i^+)}\right)^\top K(x_i^+, x_n^+) \left(\frac{\partial \mathcal{L}}{\partial f(x_n^+)}\right)
\end{aligned} \tag{64}$$

this can be expressed as

$$\frac{\partial}{\partial t}\mathcal{L}(t) = -u(t)^\top \boldsymbol{K}(t) u(t) \tag{65}$$

where we have defined

$$u(t) = \begin{pmatrix} \frac{\partial \mathcal{L}}{\partial f_1(x_1)} & \cdots & \frac{\partial \mathcal{L}}{\partial f_K(x_1)} & \cdots & \frac{\partial \mathcal{L}}{\partial f_K(x_N)} & \frac{\partial \mathcal{L}}{\partial f_1(x_1^+)} & \cdots & \frac{\partial \mathcal{L}}{\partial f_K(x_N^+)} \end{pmatrix}^\top \tag{66}$$

$$\boldsymbol{K}(t) = \begin{pmatrix} K(x_1,x_1) & \dots & K(x_1,x_N) & \dots & K(x_1,x_N^+) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ K(x_N,x_1) & \dots & K(x_N,x_N) & \dots & K(x_N,x_N^+) \\ K(x_1^+,x_1) & \dots & K(x_1^+,x_N) & \dots & K(x_1^+,x_N^+) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ K(x_N^+,x_1) & \dots & K(x_N^+,x_N) & \dots & K(x_N^+,x_N^+) \end{pmatrix} \tag{67}$$

□

A.3 Proof of Theorem 4.4

Proof. Observe that

$$\frac{\partial}{\partial t}\mathcal{L}(t) = -u(t)^\top \boldsymbol{K}(t) u(t) \leq -\|u(t)\|^2 \lambda_{min}(\boldsymbol{K}(t)) \tag{68}$$

with $\boldsymbol{K}$ and $u(t)$ as defined earlier. Recall from Appendix A.1 (part 3 of the proof) that

$$\frac{\partial \mathcal{L}}{\partial f_k(x_i)} = \frac{2}{N} e_k^\top (C - I) f(x_i^+) \tag{69}$$

Thus, we can lower bound the squared Euclidean norm of $u(t)$ as

$$\begin{aligned}
\|u(t)\|^2 &= \sum_{k=1}^{K}\sum_{i=1}^{N}\left(\frac{\partial \mathcal{L}}{\partial f_k(x_i)}\right)^2 + \left(\frac{\partial \mathcal{L}}{\partial f_k(x_i^+)}\right)^2 \\
&= \frac{4}{N^2}\sum_{k=1}^{K}\sum_{i=1}^{N} f(x_i^+)^\top (C-I)e_k e_k^\top (C-I) f(x_i^+) + f(x_i)^\top (C-I)e_k e_k^\top (C-I) f(x_i) \\
&= \frac{4}{N^2}\sum_{i=1}^{N} f(x_i^+)^\top (C-I)^2 f(x_i^+) + f(x_i)^\top (C-I)^2 f(x_i) \\
&\geq \frac{4}{N^2}\sum_{i=1}^{N} f(x_i^+)(C-I)^2 f(x_i) + f(x_i)(C-I)^2 f(x_i^+) \\
&= \frac{4}{N^2}\,\mathrm{Trace}\left((C-I)^2\left(\sum_{i=1}^{N} f(x_i)f(x_i^+)^\top + f(x_i^+)f(x_i)^\top\right)\right) \\
&= \frac{8}{N}\,\mathrm{Trace}((C-I)^2 C)
\end{aligned} \tag{70}$$

The inequality used above relies on the following fact: For any positive semi-definite matrix $A = Q^\top \Lambda Q \in \mathbb{R}^K$ (where Q is orthonormal and Λ is diagonal with non-negative entries), and for any $v, w \in \mathbb{R}^K$, Cauchy-Schwartz yields

$$\begin{aligned}
v^\top A w + w^\top A v &= (Qv)^\top \Lambda (Qw) + (Qw)^\top \Lambda (Qv) \\
&= 2\sum_{k=1}^{K} \lambda_k (Qv)_k (Qw)_k \\
&\leq \sum_{k=1}^{K} \lambda_k \left((Qv)_k^2 + (Qw)_k^2\right) \\
&= v^\top Q^\top \Lambda Q v + w^\top Q^\top \Lambda Q w \\
&= v^\top A v + w^\top A w
\end{aligned} \tag{71}$$

All that remains to be done is to show that

$$\|u(t)\|^2 \geq \kappa \|C - I\|_F^2 \tag{72}$$

for some κ that is time-independent. By Von Neumann's trace inequality we can certainly choose $\kappa = \lambda_{min}(C(t))$ at any given t. Thus, as long as the smallest eigenvalue of $C(t)$ is lower bounded until convergence, we are fine. Recall that we assume $\mathcal{L}(0) \leq 1 - \rho < 1$. Now assume that at some time t' we have $\lambda_{min}(C(t')) < 1 - \sqrt{\mathcal{L}(0)}$. Then, denoting $\lambda_1, \ldots, \lambda_K$ for the eigenvalues of $C(t')$, it follows that

$$\mathcal{L}(t') = \mathrm{Trace}\left((C-I)^2\right) = \sum_{k=1}^{K}(\lambda_k - 1)^2 \geq (\lambda_{min}(C(t')) - 1)^2 > \mathcal{L}(0) \tag{73}$$

This is a contradiction to the non-increasing nature of the loss under gradient flow. Hence, as long as $\mathcal{L}(0) < 1$, we can lower bound $\lambda_{min}(C)$ uniformly in time by $1 - \sqrt{\mathcal{L}(0)}$ from below. Thus, it holds that

$$\|u(t)\|^2 > \frac{8}{N}(1 - \sqrt{\mathcal{L}(0)}) \cdot \mathcal{L}(t) \tag{74}$$

and hence

$$\frac{\partial}{\partial t}\mathcal{L}(t) < -\frac{8(1-\sqrt{\mathcal{L}(0)}) \cdot \lambda_{min}(\boldsymbol{K}(t))}{N} \cdot \mathcal{L}(t) \leq -\frac{8(1-\sqrt{1-\rho}) \cdot \lambda_{min}(\boldsymbol{K}(t))}{N} \cdot \mathcal{L}(t) \tag{75}$$

holds for all t. We now take M to be so large that the entries of the kernel matrix change by less than some $\gamma > 0$ for all $t \leq T$, where $\gamma > 0$ is small enough to ensure that $\lambda_{min}(\boldsymbol{K}(t)) \geq \lambda/2$ for all $t \leq T$. Such M certainly exists, because Theorem 3.1 guarantees that the change in the entries of the kernel matrix $\boldsymbol{K}(t)$ are no more than $\mathcal{O}(\frac{R^2}{\sqrt{M}})$ until time T, where $R = \kappa T$ following the discussion preceding Theorem 4.1. Overall, this implies that whenever $M \geq M_1$ for some $M_1 \in \mathbb{N}$, we have $\frac{\partial}{\partial t}\mathcal{L}(t) < -\eta \mathcal{L}(t)$ for all $t \leq T$. □

A.4 Ensuring $\mathcal{L}(0) < 1$ for ReLU

Lemma A.1. ***(Scaling the first layer ensures small loss)*** *Consider the ReLU activation function. For any dataset $\mathcal{X}, \mathcal{X}_+$, there exist $s > 0$ and $M_0 \in \mathbb{N}$ such that for all $M \geq M_0$ and under $w_{mk} \sim \mathcal{N}(0,1), v_{mr} \sim \mathcal{N}(0, s^2)$, the loss at initialization satisfies $\mathcal{L}(0) < 1$ with high probability.*

Proof. In expectation, we have the following expression for the cross-moments at initialization.

$$\begin{aligned}\mathbb{E}[C_{kl}] &= \frac{1}{2N}\sum_{n=1}^{N}\mathbb{E}[f_k(x_n)f_l(x_n^+)] + \mathbb{E}[f_k(x_n^+)f_l(x_n)] \\ &= \frac{1}{2N}\sum_{n=1}^{N}\frac{1}{M}\sum_{m,m'=1}^{M}\mathbb{E}\left[w_{mk}\left(\phi(v_m^\top x_n)\phi(v_{m'}^\top x_n^+) + \phi(v_m^\top x_n^+)\phi(v_{m'}^\top x_n)\right)w_{m'l}\right]\end{aligned} \tag{76}$$

For $k \neq l$, the fact that all w_{mk} are independent and zero mean shows that $\mathbb{E}[C_{kl}] = 0$. When $k = l$, we have $\mathbb{E}[w_{mk}w_{m'l}] = \mathbf{1}(m = m')$, since the weights w_{mk} are standard Gaussians. Therefore,

$$\mathbb{E}[C_{kk}] = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{M}\sum_{m}^{M}\mathbb{E}\left[\phi(v_m^\top x_n)\phi(v_m^\top x_n^+)\right] \tag{77}$$

For ReLU, the expectation is strictly positive and scales as s^2. Thus, there certainly exists $s > 0$ such that

$$\mathbb{E}[C_{kk}] = \frac{2K-1}{2K} \tag{78}$$

For large M, the variance of each C_{kl} is $\mathcal{O}(1/M)$ because the weights are independent. Thus, the matrix $I - C$ concentrates around $(2K)^{-1}I_K$, so $\mathcal{L}(0) = \|C(0) - I\|_F^2 < 1$ with high probability. □

B PROOFS FROM SECTION 5

B.1 Proof of Theorem 5.1

Proof. Recall that $T > 0$ is a fixed, width-independent point in time. We show that there exists $\kappa > 0$ such that $\sup_{t \leq T} \|\dot{\theta}\| \leq \kappa$ for all networks of width $M > 8T$, with high probability over randomly initialized weights. Specifically, for any $\epsilon > 0$, we know that there exists $\kappa_1 > 0$ such that

$$\frac{\|\theta(0)\|^2}{M} \leq \kappa_1 - 1 \tag{79}$$

with probability at least $1 - \epsilon$, because the weights are independent Gaussians at initialization. We condition everything on this event. In this proof, we also assume $\mathcal{L}(0) < 1$ under said event. We need this condition anyway for Theorem 4.4 and therefore include it into our bounds here as well to simplify matters. However, any width-independent bound on $\mathcal{L}(0)$ is sufficient. Denote $c_\phi = \max_{t \in \mathbb{R}} |\phi(t)|$ and $c_{\phi'} = \max_{t \in \mathbb{R}} \left|\frac{\partial}{\partial t}\phi(t)\right|$. We will show the following three statements.

1. Firstly, we show that

$$\frac{\partial}{\partial t}\left(\sum_{m=1}^{M} w_m^2(t)\right) \leq 8 \tag{80}$$

   for all $t \leq T$. This immediately implies that, with probability $\geq 1 - \epsilon$ over random initialization,

$$\sup_{t \leq T} \frac{1}{M}\left(\sum_{m=1}^{M} w_m^2(t)\right) \leq \kappa_1 \tag{81}$$

   because for any t, we can write

$$\frac{1}{M}\left(\sum_{m=1}^{M} w_m^2(t)\right) \leq \frac{1}{M}\left(\sum_{m=1}^{M} w_m^2(0)\right) + \frac{8T}{M} \leq \frac{1}{M}\left(\sum_{m=1}^{M} w_m^2(0)\right) + 1 \tag{82}$$

   where we used $M > 8T$ and the fact that $\frac{\|\theta(0)\|^2}{M} \leq \kappa_1 - 1$ with high probability.

2. From there, we bound the maximum squared value that any representation takes until time T, that is

$$|f|^2 = \max_{n \in [N]} \sup_{t \leq T} \max\left(|f(x_n)|^2, |f(x_n^+)|^2\right) \tag{83}$$

   by some $\kappa_2 > 0$ (again independent of M).

3. We combine both statements to show that there exists $\kappa > 0$ such that $\sup_{t \leq T} \|\dot{\theta}\| \leq \kappa$ with probability at least $1 - \epsilon$.

Part 1. We begin by computing

$$\frac{\partial \mathcal{L}}{\partial w_m} = 2(C-1)\left(\frac{1}{N}\sum_{n=1}^{N}\left[f(x_n)\frac{1}{\sqrt{M}}\phi(v_m^\top x_n^+) + f(x_n^+)\frac{1}{\sqrt{M}}\phi(v_m^\top x_n)\right]\right) \tag{84}$$

which equates to $-\frac{\partial}{\partial t} w_m$ under gradient flow, and

$$\frac{\partial \mathcal{L}}{\partial v_{mj}} = 2(C-1)\left(\frac{1}{N}\sum_{n=1}^{N}\left[f(x_n)\frac{1}{\sqrt{M}} w_m \phi'(v_m^\top x_n^+)(x_n^+)^{(j)} + f(x_n^+)\frac{1}{\sqrt{M}} w_m \phi'(v_m^\top x_n)(x_n)^{(j)}\right]\right) \tag{85}$$

which equates to $-\frac{\partial}{\partial t} v_{mj}$ under gradient flow. Clearly,

$$\frac{\partial}{\partial t}\left(w_m^2\right) = 4(1-C)\left(\frac{1}{N}\sum_{n=1}^{N}\left[f(x_n) f_m(x_n^+) + f(x_n^+) f_m(x_n)\right]\right) \tag{86}$$

where we write $f_m(x) = \frac{1}{\sqrt{M}} w_m \phi(v_m^\top x)$. Noticing that $f(x) = \sum_{m=1}^M f_m(x)$, we arrive at

$$\begin{aligned}\frac{\partial}{\partial t}\left(\sum_{m=1}^M w_m^2\right) &= 4(1-C)\left(\frac{1}{N}\sum_{n=1}^N\left[f(x_n)\left(\sum_{m=1}^M f_m(x_n^+)\right) + f(x_n^+)\left(\sum_{m=1}^M f_m(x_n)\right)\right]\right) \\ &= 8(1-C)C\end{aligned} \tag{87}$$

The loss $\mathcal{L}(t) = (C-1)^2$ is non-increasing under gradient flow. Thus, since $\mathcal{L}(0) < 1$, we have $C \in (0,1)$. We obtain $\frac{\partial}{\partial t}\left(\sum_{m=1}^M w_m^2(t)\right) \le 8$ for all $t \le T$ as desired.

Part 2. We consider the evolution of the representations $f(x)$ over time, where $x \in \mathcal{X}, \mathcal{X}_+$ Note that for any x,

$$\begin{aligned}\frac{\partial}{\partial t} f(x) &= \langle \frac{\partial}{\partial \theta} f(x), \frac{\partial}{\partial t}\theta\rangle \\ &= -\sum_{m=1}^M \frac{\partial f(x)}{\partial \theta_m}\frac{\partial \mathcal{L}}{\partial \theta_m} \\ &= -\sum_{m=1}^M \frac{\partial}{\partial \theta_m} f(x)\left[\sum_{n=1}^N\left(\frac{\partial \mathcal{L}}{\partial f(x_n)}\frac{\partial f(x_n)}{\partial \theta_m} + \frac{\partial \mathcal{L}}{\partial f(x_n^+)}\frac{\partial f(x_n^+)}{\partial \theta_m}\right)\right] \\ &= -\sum_{n=1}^N\left[K_\theta(x,x_n)\frac{\partial \mathcal{L}}{\partial f(x_n)} + K_\theta(x,x_n^+)\frac{\partial \mathcal{L}}{\partial f(x_n^+)}\right]\end{aligned} \tag{88}$$

In our setting

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial f(x_n)} &= 2(C-1)\left(\frac{1}{N} f(x_n^+)\right) \\ \frac{\partial \mathcal{L}}{\partial f(x_n^+)} &= 2(C-1)\left(\frac{1}{N} f(x_n)\right)\end{aligned} \tag{89}$$

Hence, for any $i \in [N]$, we have

$$\begin{aligned}\frac{\partial}{\partial t} f(x_i) &= \frac{2(1-C)}{N}\cdot\sum_{n=1}^N\left[K_\theta(x_i,x_n) f(x_n^+) + K_\theta(x_i,x_n^+) f(x_n)\right] \\ &= \frac{2(1-C)}{N}\cdot K_\theta(x_i,[\mathcal{X},\mathcal{X}_+])\begin{pmatrix} f(\mathcal{X}_+) \\ f(\mathcal{X})\end{pmatrix}\end{aligned} \tag{90}$$

and therefore we can write

$$\frac{\partial}{\partial t}\begin{pmatrix} f(\mathcal{X}) \\ f(\mathcal{X}_+)\end{pmatrix} = \frac{2(1-C)}{N}\cdot K_\theta([\mathcal{X},\mathcal{X}_+],[\mathcal{X},\mathcal{X}_+])\begin{pmatrix} f(\mathcal{X}_+) \\ f(\mathcal{X})\end{pmatrix} \tag{91}$$

Thus,

$$\frac{\partial}{\partial t}\left\|\begin{pmatrix} f(\mathcal{X}) \\ f(\mathcal{X}_+)\end{pmatrix}\right\|^2 = 2\begin{pmatrix} f(\mathcal{X}) \\ f(\mathcal{X}_+)\end{pmatrix}^\top\left(\frac{\partial}{\partial t}\begin{pmatrix} f(\mathcal{X}) \\ f(\mathcal{X}_+)\end{pmatrix}\right) \tag{92}$$

$$= \frac{4(1-C)}{N}\begin{pmatrix} f(\mathcal{X}) \\ f(\mathcal{X}_+)\end{pmatrix}^\top K_\theta([\mathcal{X},\mathcal{X}_+],[\mathcal{X},\mathcal{X}_+])\begin{pmatrix} f(\mathcal{X}_+) \\ f(\mathcal{X})\end{pmatrix} \tag{93}$$

By Cauchy-Schwartz inequality, for any positive semi-definite matrix A, it holds that

$$|y^\top A z|^2 \le (y^\top A y)\cdot(z^\top A z) \le \|y\|^2\|z\|^2\|A\|_2^2 \tag{94}$$

In our case, this implies that

$$\begin{pmatrix} f(\mathcal{X}) \\ f(\mathcal{X}_+)\end{pmatrix}^\top K_\theta([\mathcal{X},\mathcal{X}_+],[\mathcal{X},\mathcal{X}_+])\begin{pmatrix} f(\mathcal{X}_+) \\ f(\mathcal{X})\end{pmatrix} \le \left\|\begin{pmatrix} f(\mathcal{X}) \\ f(\mathcal{X}_+)\end{pmatrix}\right\|^2\cdot\|K_\theta([\mathcal{X},\mathcal{X}_+],[\mathcal{X},\mathcal{X}_+])\|_2 \tag{95}$$

where we used the fact that

$$\left\| \begin{pmatrix} f(\mathcal{X}) \\ f(\mathcal{X}_+) \end{pmatrix} \right\| = \left\| \begin{pmatrix} f(\mathcal{X}_+) \\ f(\mathcal{X}) \end{pmatrix} \right\| \tag{96}$$

since both vectors are just permutations of one another. It remains to show that the spectral norm of the time-dependent kernel matrix remains bounded until time T. Note that for any pair $(a, b) \in [\mathcal{X}, \mathcal{X}_+] \times [\mathcal{X}, \mathcal{X}_+]$, and for any time $t \leq T$, we have $|a^\top b| \leq 1$. Therefore, every entry of the kernel matrix is bounded via

$$\begin{aligned} K_\theta(a,b) &= \left(\frac{\partial f(a)}{\partial \theta}\right)^\top \left(\frac{\partial f(b)}{\partial \theta}\right) \\ &= \sum_{m=1}^{M} \left(\frac{\partial f(a)}{\partial w_m} \frac{\partial f(b)}{\partial w_m} + \sum_{j=1}^{d} \frac{\partial f(a)}{\partial v_{mj}} \frac{\partial f(b)}{\partial v_{mj}} \right) \\ &= \frac{1}{M} \sum_{m=1}^{M} \phi(v_m^\top a)\phi(v_m^\top b) + w_m^2 \phi'(v_m^\top a)\phi'(v_m^\top b) a^\top b \\ &\leq \frac{1}{M} \sum_{m=1}^{M} c_\phi^2 + w_m^2 c_{\phi'}^2 \\ &\leq c_\phi^2 + c_{\phi'}^2 \left(\frac{1}{M} \sum_{m=1}^{M} w_m^2(t) \right) \\ &\leq c_\phi^2 + c_{\phi'}^2 \kappa_1 \end{aligned} \tag{97}$$

where we used Part 1 in the final inequality. Since the spectral norm is upper bounded by the trace, we obtain

$$\|K_\theta([\mathcal{X}, \mathcal{X}_+], [\mathcal{X}, \mathcal{X}_+])\|_2 \leq 2N \left(c_\phi^2 + c_{\phi'}^2 \kappa_1\right) \tag{98}$$

Consequently, going back to the evolution of the representations, we see that

$$\frac{\partial}{\partial t} \left\| \begin{pmatrix} f(\mathcal{X}) \\ f(\mathcal{X}_+) \end{pmatrix} \right\|^2 \leq 8(1-C) \left(c_\phi^2 + c_{\phi'}^2 \kappa_1\right) \left\| \begin{pmatrix} f(\mathcal{X}) \\ f(\mathcal{X}_+) \end{pmatrix} \right\|^2 \leq 8 \left(c_\phi^2 + c_{\phi'}^2 \kappa_1\right) \left\| \begin{pmatrix} f(\mathcal{X}) \\ f(\mathcal{X}_+) \end{pmatrix} \right\|^2 \tag{99}$$

Grönwall's inequality now ensures that there exists $\kappa_2 > 0$ such that

$$\left\| \begin{pmatrix} f(\mathcal{X}) \\ f(\mathcal{X}_+) \end{pmatrix}(t) \right\|^2 \leq \left\| \begin{pmatrix} f(\mathcal{X}) \\ f(\mathcal{X}_+) \end{pmatrix}(0) \right\|^2 \exp\left(8T \left(c_\phi^2 + c_{\phi'}^2 \kappa_1\right)\right) =: \kappa_2 \tag{100}$$

for all time $t \leq T$, where we used the fact that $\left\| \begin{pmatrix} f(\mathcal{X}) \\ f(\mathcal{X}_+) \end{pmatrix}(0) \right\|^2$ can be bounded in terms of $\frac{\|\theta(0)\|^2}{M}$. Since

$$|f|^2 = \max_{n \in [N]} \sup_{t \leq T} \max\left(|f(x_n)|^2, |f(x_n^+)|^2\right) \leq \sup_{t \leq T} \left\| \begin{pmatrix} f(\mathcal{X}) \\ f(\mathcal{X}_+) \end{pmatrix}(t) \right\|^2 \tag{101}$$

this concludes part 2, giving us a bound κ_2 on $|f|^2$.

Part 3. We finally control $\|\dot{\theta}(t)\|^2$. Using the time derivatives for w_m, v_{mj} we derived in Part 1,

$$\begin{aligned} |\dot{w_m}(t)|^2 &= 4(1-C)^2 \left(\frac{1}{N} \sum_{n=1}^{N} \left[f(x_n) \frac{1}{\sqrt{M}} \phi(v_m^\top x_n^+) + f(x_n^+) \frac{1}{\sqrt{M}} \phi(v_m^\top x_n) \right] \right)^2 \\ &\leq 4(1-C)^2 \left(2|f| c_\phi \frac{1}{\sqrt{M}} \right)^2 \\ &= 16(1-C)^2 \frac{|f|^2 c_\phi^2}{M} \\ &\leq \frac{16 \kappa_2 c_\phi^2}{M} \end{aligned} \tag{102}$$

for all $m \in [M]$. Similarly, it holds that

$$\begin{aligned} |\dot{v_{mj}}(t)|^2 &= 4(1-C)^2 \left(\frac{1}{N} \sum_{n=1}^{N} \left[f(x_n) \frac{1}{\sqrt{M}} w_m \phi'(v_m^\top x_n^+)(x_n^+)^{(j)} + f(x_n^+) \frac{1}{\sqrt{M}} w_m \phi'(v_m^\top x_n)(x_n)^{(j)} \right] \right)^2 \\ &\leq 4(1-C)^2 \left(\frac{2|f||w_m| c_{\phi'}}{\sqrt{M}} \right)^2 \\ &\leq \frac{16 \kappa_2 w_m^2 c_{\phi'}^2}{M} \end{aligned} \tag{103}$$

for all $m \in [M], j \in [d]$. Thus, we obtain

$$\begin{aligned} \|\dot{\theta}(t)\|^2 &= \sum_{m=1}^{M} \sum_{j=1}^{d} (\dot{w_m}(t))^2 + (\dot{v_{mj}}(t))^2 \\ &\leq \frac{1}{M} \left(\sum_{m=1}^{M} 16 \kappa_2 c_\phi^2 + 16 d \kappa_2 w_m^2 c_{\phi'}^2 \right) \end{aligned} \tag{104}$$

From part 1, $\frac{1}{M} \sum_{m=1}^{M} w_m^2(t) \leq \kappa_1$ holds for all $t \leq T$. Thus, defining

$$\kappa = 4 \sqrt{\kappa_2 c_\phi^2 + d \kappa_1 \kappa_2 c_{\phi'}^2} \tag{105}$$

we obtain the desired high-probability bound on $\sup_{t \leq T} \|\dot{\theta}(t)\|$ and this conclude the proof. □

B.2 Proof of Lemma 5.2

Proof. First recall what we showed in Appendix B.1. The function representations $f(x)$ evolve as

$$\frac{\partial}{\partial t} f(x) = - \sum_{n=1}^{N} K_\theta(x, x_n) \frac{\partial \mathcal{L}}{\partial f(x_n)} + K_\theta(x, x_n^+) \frac{\partial \mathcal{L}}{\partial f(x_n^+)} \tag{106}$$

for any $i \in [N]$. From there, it is apparent that

$$\begin{aligned} \frac{\partial}{\partial t} \mathcal{L}(t) &= \sum_{i=1}^{N} \frac{\partial \mathcal{L}}{\partial f(x_i)} \frac{\partial}{\partial t} f(x_i) + \frac{\partial \mathcal{L}}{\partial f(x_i^+)} \frac{\partial}{\partial t} f(x_i^+) \\ &= - \sum_{i,n=1}^{N} \frac{\partial \mathcal{L}}{\partial f(x_i)} \frac{\partial \mathcal{L}}{\partial f(x_n)} K(x_n, x_i) + \frac{\partial \mathcal{L}}{\partial f(x_i^+)} \frac{\partial \mathcal{L}}{\partial f(x_n)} K(x_n, x_i^+) + \\ &\qquad \frac{\partial \mathcal{L}}{\partial f(x_i)} \frac{\partial \mathcal{L}}{\partial f(x_n^+)} K(x_n^+, x_i) + \frac{\partial \mathcal{L}}{\partial f(x_i^+)} \frac{\partial \mathcal{L}}{\partial f(x_n^+)} K(x_n^+, x_i^+) \\ &= - \frac{4(C-1)^2}{N^2} \sum_{i,n=1}^{N} f(x_i^+) f(x_n^+) K(x_n, x_i) + f(x_i) f(x_n^+) K(x_n, x_i^+) + \\ &\qquad f(x_i^+) f(x_n) K(x_n^+, x_i) + f(x_i) f(x_n) K(x_n^+, x_i^+) \\ &= - \frac{4}{N^2} \cdot \mathcal{L}(t) \cdot \begin{pmatrix} f(\mathcal{X}) \\ f(\mathcal{X}_+) \end{pmatrix}^\top \boldsymbol{K}(t) \begin{pmatrix} f(\mathcal{X}) \\ f(\mathcal{X}_+) \end{pmatrix} \end{aligned} \tag{107}$$

where the kernel matrix $\boldsymbol{K}(t)$ is defined as

$$\boldsymbol{K}(t) = \begin{pmatrix} K_\theta(\mathcal{X}_+, \mathcal{X}_+) & K_\theta(\mathcal{X}_+, \mathcal{X}) \\ K_\theta(\mathcal{X}, \mathcal{X}_+) & K_\theta(\mathcal{X}, \mathcal{X}) \end{pmatrix} \tag{108}$$

□

B.3 Proof of Theorem 5.3

Proof. Note that the entries of the kernel matrix $\boldsymbol{K}$ change by no more than $\mathcal{O}(\frac{\kappa^2 T^2}{\sqrt{M}})$ up to time T. This is a consequence of Theorem 3.1 which states that the change is $\mathcal{O}(\frac{R^2}{\sqrt{M}})$, and Theorem 5.1 which asserts that $R = \mathcal{O}(\kappa T)$. Since the spectral norm is upper bounded by the trace of a matrix, for any $\gamma > 0$ there exists some large M such that

$$\|\boldsymbol{K}(t) - \boldsymbol{K}(0)\|_2 \leq \text{Trace}\left(\boldsymbol{K}(t) - \boldsymbol{K}(0)\right) \leq \gamma \tag{109}$$

for all $t \leq T$. We pick $\gamma = \frac{\lambda}{2}$, which ensures that

$$\lambda_{min}(\boldsymbol{K}(t)) \geq \frac{\lambda}{2} \tag{110}$$

Define

$$z(t) = \begin{pmatrix} f(\mathcal{X}) \\ f(\mathcal{X}_+) \end{pmatrix}(t) \tag{111}$$

We claim that $\frac{\|z(t)\|^2}{2N} \geq C(0)$ for all time $t \geq 0$. Indeed, assume this was not the case at a certain time $t' > 0$, where instead

$$\frac{\|z(t')\|^2}{2N} < C(0) \tag{112}$$

Then,

$$\begin{aligned} C(t') &= \frac{1}{2N}\sum_{n=1}^{N} f(x_n)f(x_n^+) + f(x_n^+)f(x_n) \\ &= \frac{1}{2N}\left(\begin{pmatrix} f(\mathcal{X}) \\ f(\mathcal{X}_+) \end{pmatrix}(t')\right)^\top \left(\begin{pmatrix} f(\mathcal{X}_+) \\ f(\mathcal{X}) \end{pmatrix}(t')\right) \\ &\leq \frac{1}{2N}\|z(t')\|^2 \\ &< C(0) \end{aligned} \tag{113}$$

The first inequality is Cauchy-Schwartz, using the fact that

$$\left\|\begin{pmatrix} f(\mathcal{X}) \\ f(\mathcal{X}_+) \end{pmatrix}(t')\right\| = \left\|\begin{pmatrix} f(\mathcal{X}_+) \\ f(\mathcal{X}) \end{pmatrix}(t')\right\| = \|z(t')\| \tag{114}$$

The second inequality plugs in our assumption (112). But if $C(t') < C(0)$, then this leads to a contradiction. Since $C(0) \in (0, 1)$, we would obtain

$$\mathcal{L}(t') = (1 - C(t'))^2 > (1 - C(0))^2 = \mathcal{L}(0) \tag{115}$$

This is impossible, because gradient flow monotonically decreases $\mathcal{L}$. Thus, for all $t \leq T$, we know that

$$\frac{\|z(t)\|^2}{2N} \geq C(0) \tag{116}$$

From there, we complete the proof.

$$\begin{aligned} g(t) &= \frac{4}{N^2} \cdot z(t)^\top \boldsymbol{K}(t) z(t) \\ &\geq \frac{4}{N^2} \cdot \|z(t)\|^2 \cdot \lambda_{min}(\boldsymbol{K}(t)) \\ &> \frac{2\lambda}{N^2} \cdot \|z(t)\|^2 \\ &\geq \frac{2\lambda}{N^2} \cdot (2NC(0)) \\ &= \frac{4\lambda C(0)}{N} \\ &= \frac{4\lambda(1 - \sqrt{\mathcal{L}(0)})}{N} \\ &\geq \frac{4\lambda(1 - \sqrt{1-\rho})}{N} \\ &= \eta \end{aligned} \tag{117}$$

as desired. □

C EXTENDING THEOREM 5.1 TO RELU ACTIVATIONS

In this section, we prove that the weights remain in a ball of bounded radius even for ReLU activations, until any fixed time $T > 0$. The proof is not entirely rigorous, because gradient flow w.r.t. the weights is ill-defined due to the non-differentiability of the ReLU function. We restrict ourselves to an analysis of the one-dimensional setting.

Proof. Similar to the case for smooth bounded activations with bounded first derivatives (Appendix B.1), our proof consists of three parts. Remember that $T > 0$ is a fixed, width-independent point in time. We will show the following three statements, and assume $\mathcal{L}(0) < 1$ throughout. However, as before, all that we really need is boundedness of the loss at initialization. Note that our arguments are slightly reshuffled compared to the proof of Theorem 5.1. This is due to unboundedness of the ReLU function.

1. Firstly, we show that $\frac{\partial}{\partial t}\|\theta(t)\|^2 \leq 16$ for all $t \leq T$.
2. Secondly, we verify that for any $t \leq T$, it holds that

$$\|\dot{\theta}(t)\|^2 \leq \frac{16|f|^2 d\|\theta\|^2}{M} \tag{118}$$

   where we denote

$$|f|^2 = \max_{n\in[N]} \sup_{t\leq T} \max\left(|f(x_n)|^2, |f(x_n^+)|^2\right) \tag{119}$$

   for the maximum squared value that any representation takes until time T.
3. Thirdly, we show that there exists a constant κ_1 (again depending only on T) such that $|f|^2 \leq \kappa_1$.

Together, this will be enough. Part 1 implies that for all $t \leq T$, we have

$$\frac{\|\theta(t)\|^2}{M} \leq \frac{\|\theta(0)\|^2 + 16T}{M} \leq \frac{\|\theta(0)\|^2}{M} + 2 \tag{120}$$

since $M > 8T$. Then, combining part 2 and part 3, we arrive at

$$\|\dot{\theta}(t)\|^2 \leq \frac{16d\kappa_1(\|\theta(0)\|^2 + 16T)}{M} \leq 16d\kappa_1\left(\frac{\|\theta(0)\|^2}{M} + 2\right) \tag{121}$$

for all $t \leq T$. Note that for any $\epsilon > 0$ we can choose κ such that this expression is smaller than κ with probability at least $1 - \epsilon$. This is possible because the weights at initialization are independent Gaussians.

Part 1. We begin by computing

$$\frac{\partial \mathcal{L}}{\partial w_m} = 2(C-1)\left(\frac{1}{N}\sum_{n=1}^{N}\left[f(x_n)\frac{1}{\sqrt{M}}\phi(v_m^\top x_n^+) + f(x_n^+)\frac{1}{\sqrt{M}}\phi(v_m^\top x_n)\right]\right) \tag{122}$$

which equates to $-\frac{\partial}{\partial t}w_m$ under gradient flow, and

$$\frac{\partial \mathcal{L}}{\partial v_{mj}} = 2(C-1)\left(\frac{1}{N}\sum_{n=1}^{N}\left[f(x_n)\frac{1}{\sqrt{M}}w_m\phi'(v_m^\top x_n^+)(x_n^+)^{(j)} + f(x_n^+)\frac{1}{\sqrt{M}}w_m\phi'(v_m^\top x_n)(x_n)^{(j)}\right]\right) \tag{123}$$

which equates to $-\frac{\partial}{\partial t}v_{mj}$ under gradient flow. Thus, writing $f_m(x) = \frac{1}{\sqrt{M}}w_m\phi(v_m^\top x)$, it holds that

$$\frac{\partial}{\partial t}\left(w_m^2\right) = \frac{\partial}{\partial t}\left(\|v_m\|^2\right) = 4(1-C)\left(\frac{1}{N}\sum_{n=1}^{N}\left[f(x_n)f_m(x_n^+) + f(x_n^+)f_m(x_n)\right]\right) \tag{124}$$

where we used properties of the ReLU function ϕ, namely that

$$\phi(v_m^\top x) = \phi'(v_m^\top x)\cdot v_m^\top x \tag{125}$$

Hence, noticing that $f(x) = \sum_{m=1}^{M} f_m(x)$, we arrive at

$$\begin{aligned}\frac{\partial}{\partial t}\left(\|\theta\|^2\right) &= \frac{\partial}{\partial t}\left(\sum_{m=1}^{M}\left(w_m^2 + \|v_m\|^2\right)\right) \\ &= 8(1-C)\left(\frac{1}{N}\sum_{n=1}^{N}\left[f(x_n)\left(\sum_{m=1}^{M} f_m(x_n^+)\right) + f(x_n^+)\left(\sum_{m=1}^{M} f_m(x_n)\right)\right]\right) \\ &= 16(1-C)C\end{aligned} \tag{126}$$

The loss is decreasing under gradient flow. Due to $\mathcal{L}(0) < 1$, we have $C \in (0,1)$. We obtain $\frac{\partial}{\partial t}\left(\|\theta\|^2\right) \leq 16$ as claimed.

Part 2. We now control $\|\dot{\theta}(t)\|$. First note that $|\phi(v_m^\top x)| \leq \|v_m\|$ due to $\|x\| = 1$ for all $x \in (\mathcal{X}, \mathcal{X}_+)$. Moreover,

$$\begin{aligned}|\dot{w_m}(t)|^2 &= 4(1-C)^2\left(\frac{1}{N}\sum_{n=1}^{N}\left[f(x_n)\frac{1}{\sqrt{M}}\phi(v_m^\top x_n^+) + f(x_n^+)\frac{1}{\sqrt{M}}\phi(v_m^\top x_n)\right]\right)^2 \\ &\leq 4(1-C)^2\left(\frac{\|v_m\|}{N\sqrt{M}}\sum_{n=1}^{N}\left[|f(x_n)| + |f(x_n^+)|\right]\right)^2 \\ &\leq 4(1-C)^2\left(\frac{2\|v_m\||f|}{\sqrt{M}}\right)^2 \\ &= \frac{16(1-C)^2\|v_m\|^2|f|^2}{M} \\ &\leq \frac{16\|v_m\|^2|f|^2}{M}\end{aligned} \tag{127}$$

holds for all $m \in [M]$. Similarly, it holds that

$$|\dot{v_{mj}}(t)|^2 \leq \frac{16 w_m^2 |f|^2}{M} \tag{128}$$

for all $m \in [M], j \in [d]$. Thus, we obtain

$$\begin{aligned}\|\dot{\theta}(t)\|^2 &= \sum_{m=1}^{M}\left[|\dot{w_m}(t)|^2 + \sum_{j=1}^{d}|\dot{v_{mj}}(t)|^2\right] \\ &\leq \frac{16|f|^2}{M}\left[\sum_{m=1}^{M}\left(\|v_m\|^2 + d w_m^2\right)\right] \\ &\leq \frac{16|f|^2 d\|\theta\|^2}{M}\end{aligned} \tag{129}$$

as desired.

Part 3. Just as in Appendix B.1, we look at the evolution of the representations over time. We obtain

$$\frac{\partial}{\partial t}\begin{pmatrix} f(\mathcal{X}) \\ f(\mathcal{X}_+)\end{pmatrix} = \frac{2(1-C)}{N}\cdot K_\theta([\mathcal{X}, \mathcal{X}_+], [\mathcal{X}, \mathcal{X}_+])\begin{pmatrix} f(\mathcal{X}_+) \\ f(\mathcal{X})\end{pmatrix} \tag{130}$$

and thus

$$\begin{aligned}\frac{\partial}{\partial t}\left\|\begin{pmatrix} f(\mathcal{X}) \\ f(\mathcal{X}_+)\end{pmatrix}\right\|^2 &= 2\begin{pmatrix} f(\mathcal{X}) \\ f(\mathcal{X}_+)\end{pmatrix}^\top\left(\frac{\partial}{\partial t}\begin{pmatrix} f(\mathcal{X}) \\ f(\mathcal{X}_+)\end{pmatrix}\right) \\ &= \frac{4(1-C)}{N}\begin{pmatrix} f(\mathcal{X}) \\ f(\mathcal{X}_+)\end{pmatrix}^\top K_\theta([\mathcal{X}, \mathcal{X}_+], [\mathcal{X}, \mathcal{X}_+])\begin{pmatrix} f(\mathcal{X}_+) \\ f(\mathcal{X})\end{pmatrix}\end{aligned} \tag{131}$$

By Cauchy-Schwartz inequality,

$$\begin{pmatrix} f(\mathcal{X}) \\ f(\mathcal{X}_+) \end{pmatrix}^\top K_\theta([\mathcal{X}, \mathcal{X}_+], [\mathcal{X}, \mathcal{X}_+]) \begin{pmatrix} f(\mathcal{X}_+) \\ f(\mathcal{X}) \end{pmatrix} \leq \left\| \begin{pmatrix} f(\mathcal{X}) \\ f(\mathcal{X}_+) \end{pmatrix} \right\|^2 \cdot \|K_\theta([\mathcal{X}, \mathcal{X}_+], [\mathcal{X}, \mathcal{X}_+])\|_2 \tag{132}$$

Again, we must bound the spectral norm of the time-dependent kernel matrix until time T. This is now slightly different for ReLU. Note that for any pair $(a, b) \in [\mathcal{X}, \mathcal{X}_+] \times [\mathcal{X}, \mathcal{X}_+]$, we have

$$\begin{aligned} K_\theta(a, b) &= \left(\frac{\partial f(a)}{\partial \theta} \right)^\top \left(\frac{\partial f(b)}{\partial \theta} \right) \\ &= \sum_{m=1}^{M} \frac{\partial f(a)}{\partial w_m} \frac{\partial f(b)}{\partial w_m} + \sum_{j=1}^{d} \frac{\partial f(a)}{\partial v_{mj}} \frac{\partial f(b)}{\partial v_{mj}} \\ &= \frac{1}{M} \sum_{m=1}^{M} \phi(v_m^\top a)\phi(v_m^\top b) + w_m^2 \phi'(v_m^\top a)\phi'(v_m^\top b) a^\top b \\ &\leq \frac{1}{M} \sum_{m=1}^{M} \|v_m\|^2 + w_m^2 \\ &= \frac{\|\theta(t)\|^2}{M} \\ &\leq \frac{\|\theta(0)\|^2}{M} + 1 \end{aligned} \tag{133}$$

using part 1. Since the spectral norm is upper bounded by the trace, we obtain

$$\|K_\theta([\mathcal{X}, \mathcal{X}_+], [\mathcal{X}, \mathcal{X}_+])\|_2 \leq N \left(\frac{\|\theta(0)\|^2}{M} + 1 \right) \tag{134}$$

which is $\mathcal{O}(1)$. Consequently, going back to the evolution of the representations, we see that

$$\frac{\partial}{\partial t} \left\| \begin{pmatrix} f(\mathcal{X}) \\ f(\mathcal{X}_+) \end{pmatrix} \right\|^2 \leq 4 \left(\frac{\|\theta(0)\|^2}{M} + 1 \right) \left\| \begin{pmatrix} f(\mathcal{X}) \\ f(\mathcal{X}_+) \end{pmatrix} \right\|^2 \tag{135}$$

Grönwall's inequality now ensures that

$$\left\| \begin{pmatrix} f(\mathcal{X}) \\ f(\mathcal{X}_+) \end{pmatrix}(t) \right\|^2 \leq \left\| \begin{pmatrix} f(\mathcal{X}) \\ f(\mathcal{X}_+) \end{pmatrix}(0) \right\|^2 \exp\left(4T \left(\frac{\|\theta(0)\|^2}{M} + 1 \right) \right) =: \kappa_1 \tag{136}$$

for all time $t \leq T$. Since

$$|f|^2 = \max_{n \in [N]} \sup_{t \leq T} \max\left(|f(x_n)|^2, |f(x_n^+)|^2 \right) \leq \sup_{t \leq T} \left\| \begin{pmatrix} f(\mathcal{X}) \\ f(\mathcal{X}_+) \end{pmatrix}(t) \right\|^2 \tag{137}$$

this concludes part 3 and finishes the proof. □

D PROOFS FROM SECTION 6

D.1 Proof of Theorem 6.1

Proof. Our proof is based on previous works on the theoretical analysis of Kernel PCA (Blanchard et al., 2007). Denote by $\bar{L}(W) = \|W\bar{\Gamma}W^* - I\|_F$ the square root of the population Barlow Twins loss. Denote by $L(W) = \|W\Gamma W^* - I\|_F$ the empirical version, computed with respect to the empirical cross-moment matrix

$$\Gamma = \frac{1}{2N}\sum_{n=1}^{N} z_n(z_n^+)^* + z_n^+ z_n^* \tag{138}$$

Given independently drawn positive pairs $(z_1, z_1^+), \ldots, (z_N, z_N^+)$, define the map

$$\psi(z_1, \ldots, z_N) = \sup_{\|W\|\leq B} \left|\bar{L}(W) - L(W)\right| \tag{139}$$

where $W : \mathcal{H} \to \mathbb{R}^K$ is a bounded linear operator. The function ψ satisfies a bounded differences inequality, because for any (y_n, y_n^+) (and with Γ' denoting the cross-moment matrix w.r.t. the new sample) it holds that

$$\begin{aligned}
\left|\psi(z_1, \ldots, z_n, \ldots, z_N) - \psi(z_1, \ldots, y_n, \ldots, z_N)\right| &\leq \sup_{\|W\|\leq B} \left||\bar{L}(W) - \|W\Gamma W^* - I\|_F| - |\bar{L}(W) - \|W\Gamma' W^* - I\|_F|\right| \\
&\leq \sup_{\|W\|\leq B} \left|\|W\Gamma W^* - I\|_F - \|W\Gamma' W^* - I\|_F\right| \\
&\leq \sup_{\|W\|\leq B} \|W(\Gamma - \Gamma')W^*\|_F \\
&\leq B^2 \left\| \frac{1}{2N}\left(z_n(z_n^+)^* + (z_n^+)z_n^* - y_n(y_n^+)^* - (y_n^+)y_n^*\right)\right\|_{\text{HS}} \\
&\leq \frac{B^2}{N}\left(\|z_n(z_n^+)^*\|_{\text{HS}} + \|y_n(y_n^+)^*\|_{\text{HS}}\right) \\
&\leq \frac{2B^2S^2}{N}
\end{aligned} \tag{140}$$

We bounded the difference of suprema by the supremum of the difference, used the reverse triangle inequality twice, and exploited $\|AB\|_{\text{HS}} \leq \|A\| \cdot \|B\|_{\text{HS}}$ for operators A, B. In the final step, we use

$$\|z(z^+)^*\|_{\text{HS}} = \sqrt{\text{Trace}\left(z(z^+)^*(z^+)z^*\right)} = \sqrt{\|z\|^2\|z^+\|^2} \leq S^2 \tag{141}$$

for any z, z^* (recall that they lie in a ball of radius no more than S in $\mathcal{H}$). Overall, McDiarmid's inequality yields that for any $\epsilon > 0$

$$\mathbb{P}\left(\psi(z_1, \ldots, z_n) - \mathbb{E}_{z_1,\ldots,z_N}\left[\psi(z_1, \ldots, z_n)\right] > \epsilon\right) \leq \exp\left(-\frac{2\epsilon^2}{N \cdot (\frac{2B^2S^2}{N})^2}\right) = \exp\left(-\frac{N\epsilon^2}{2B^4S^4}\right) \tag{142}$$

The expectation of ψ (w.r.t. samples $z_1, \ldots, z_N$) can be bounded as follows.

$$\begin{aligned}
\mathbb{E}_{z_1,\ldots,z_N}\left[\psi(z_1, \ldots, z_n)\right] &= \mathbb{E}_{z_1,\ldots,z_N}\left[\sup_{\|W\|\leq B}\left|\bar{L}(W) - L(W)\right|\right] \\
&\leq \mathbb{E}_{z_1,\ldots,z_N}\left[\sup_{\|W\|\leq B}\left|\|W\bar{\Gamma}W^* - I\|_F - \|W\Gamma W^* - I\|_F\right|\right] \\
&\leq B^2\mathbb{E}_{z_1,\ldots,z_N}\left[\|\Gamma - \bar{\Gamma}\|_{\text{HS}}\right]
\end{aligned} \tag{143}$$

where we used the reverse triangle inequality again. Let us write

$$\Gamma_i = \frac{1}{2}\left(z_i(z_i^+)^* + z_i^+ z_i^*\right) \tag{144}$$

so that $\Gamma = \frac{1}{N}\sum_{i=1}^{N}\Gamma_i$. Continuing with Jensen's inequality,

$$\begin{aligned}
B^2\mathbb{E}_{z_1,\ldots,z_N}\left[\|\Gamma-\bar{\Gamma}\|_{\mathrm{HS}}\right] &\leq B^2\left(\mathbb{E}_{z_1,\ldots,z_N}\left[\left\|\Gamma-\bar{\Gamma}\right\|_{\mathrm{HS}}^2\right]\right)^{1/2} \\
&= B^2\left(\mathbb{E}_{z_1,\ldots,z_N}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\left(\Gamma_i-\bar{\Gamma}\right)\right\|_{\mathrm{HS}}^2\right]\right)^{1/2} \\
&= B^2\left(\mathbb{E}_{z_1,\ldots,z_N}\left[\frac{1}{N^2}\sum_{i,j=1}^{N}\mathrm{Trace}\left((\Gamma_i-\bar{\Gamma})^*(\Gamma_j-\bar{\Gamma})\right)\right]\right)^{1/2} \\
&= B^2\left(\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}_{\Gamma_i}\left[\left\|\Gamma_i-\bar{\Gamma}\right\|_{\mathrm{HS}}^2\right]\right)^{1/2} \\
&= B^2\left(\frac{1}{N}\cdot\mathbb{E}_{\Gamma_i}\left[\left\|\Gamma_i-\bar{\Gamma}\right\|_{\mathrm{HS}}^2\right]\right)^{1/2} \\
&= \frac{B^2}{\sqrt{N}}\left(\mathbb{E}_{\Gamma_i}\left[\left\|\Gamma_i-\bar{\Gamma}\right\|_{\mathrm{HS}}^2\right]\right)^{1/2}
\end{aligned} \tag{145}$$

In the third line, we rewrote the Hilbert-Schmidt norm in terms of the trace inner product, then used independence and zero mean of all $\Gamma_i-\bar{\Gamma}, \Gamma_j-\bar{\Gamma}$ to drop the traces of $i \neq j$, and finally wrote everything in terms of a single expectation.

$$\boldsymbol{V} = \mathbb{E}\left[\left\|\Gamma_i-\bar{\Gamma}\right\|_{\mathrm{HS}}^2\right] \tag{146}$$

is the variance of the random operator Γ_i w.r.t. the Hilbert-Schmidt norm. An unbiased estimator of $\boldsymbol{V}$ is given by the empirical variance

$$\hat{\boldsymbol{V}} = \frac{1}{N'(N'-1)}\sum_{\substack{i<j \\ i,j\in[N']}}\|\Gamma_i-\Gamma_j\|_{\mathrm{HS}}^2 = \frac{1}{2N'(N'-1)}\sum_{i\neq j}\|\Gamma_i-\Gamma_j\|_{\mathrm{HS}}^2 \tag{147}$$

This is a function of independent random operators $\Gamma_1,\ldots\Gamma_{N'}$ that again satisfies a bounded differences inequality. For any collection of $\Gamma_1,\ldots,\Gamma_n$ and any "new" element Γ_n', it holds that

$$\begin{aligned}
&\left|\hat{\boldsymbol{V}}(\Gamma_1,\ldots,\Gamma_n,\ldots,\Gamma_{N'}) - \hat{\boldsymbol{V}}(\Gamma_1,\ldots,\Gamma_n',\ldots,\Gamma_{N'})\right| \\
&\leq \frac{1}{2N'(N'-1)}\sum_{i=1}^{N'}\left|\|\Gamma_i-\Gamma_n\|_{\mathrm{HS}}^2 - \|\Gamma_i-\Gamma_n'\|_{\mathrm{HS}}^2\right| \\
&\leq \frac{1}{2N'(N'-1)}\sum_{i=1}^{N'}4\|\Gamma_i\|_{\mathrm{HS}}^2 + 2\|\Gamma_n\|_{\mathrm{HS}}^2 + 2\|\Gamma_n'\|_{\mathrm{HS}}^2 \\
&= \frac{4S^4}{N'-1}
\end{aligned} \tag{148}$$

where we again used the fact that data is contained in a ball of radius S, and hence $\|\Gamma_i\|_{\mathrm{HS}}^2 \leq S^4$ for all i. Using McDiarmid's inequality once again, we conclude that with probability $\geq 1-\epsilon$ over random training data,

$$\boldsymbol{V} \leq \hat{\boldsymbol{V}} + \exp\left(-\frac{(N'-1)^2\epsilon^2}{8S^8N'}\right) \tag{149}$$

Overall, this shows that with probability $\geq 1-\epsilon$, the expected value of $\psi(z_1,\ldots,z_N)$ is bounded as

$$\mathbb{E}\left[\psi(z_1,\ldots,z_n)\right] \leq \frac{B^2}{\sqrt{N}}\left(\hat{\boldsymbol{V}} + \exp\left(-\frac{(N'-1)^2\epsilon^2}{8S^8N'}\right)\right)^{1/2} \tag{150}$$

By a union bound, we conclude that with probability $\geq 1 - 2\epsilon$ over training samples

$$\left|\bar{L}(W) - L(W)\right| \leq \frac{B^2}{\sqrt{N}} \left(\hat{\boldsymbol{V}} + \exp\left(-\frac{(N'-1)^2\epsilon^2}{8S^8N'} \right) \right)^{1/2} + \exp\left(-\frac{N\epsilon^2}{2B^4S^4} \right) =: \nu(N, \epsilon) \tag{151}$$

holds uniformly over all W with $\|W\| \leq B$. Using the fact that $\mathcal{L}(W) \leq \delta$ almost surely over randomly drawn samples, we push $L(W) \leq \sqrt{\delta}$ to the right. Then, we square both sides and use $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$. This gives

$$\bar{\mathcal{L}}(W) \leq 3\delta + \frac{3B^4}{N} \left(\hat{\boldsymbol{V}} + \exp\left(-\frac{(N'-1)^2\epsilon^2}{8S^8N'} \right) \right) + 3\exp\left(-\frac{N\epsilon^2}{B^4S^4} \right) =: \nu(N, \epsilon, \delta) \tag{152}$$

with probability at least $1 - 2\epsilon$. This concludes the proof. □

D.2 Proof of Lemma 6.2

Proof. For all $i \neq j$, we can expand the Hilbert-Schmidt norm and use the cyclic property of the trace to obtain

$$\begin{aligned} \|\Gamma_i - \Gamma_j\|_{\text{HS}}^2 &= \|\Gamma_i\|_{\text{HS}}^2 + \|\Gamma_j\|_{\text{HS}}^2 - 2\,\text{Trace}\left(\Gamma_i^*\Gamma_j\right) \\ &= \frac{1}{2}\left(\|\psi(x_i)\|^2\|\psi(x_i^+)\|^2 + \langle\psi(x_i), \psi(x_i^+)\rangle^2 + \|\psi(x_j)\|^2\|\psi(x_j^+)\|^2 + \langle\psi(x_j), \psi(x_j^+)\rangle^2 \right) \\ &\quad - \left(\langle\psi(x_i), \psi(x_j)\rangle \cdot \langle\psi(x_i^+), \psi(x_j^+)\rangle + \langle\psi(x_i), \psi(x_j^+)\rangle \cdot \langle\psi(x_i^+), \psi(x_j)\rangle \right) \end{aligned} \tag{153}$$

which can be written purely in terms of the kernel, giving

$$\begin{aligned} \|\Gamma_i - \Gamma_j\|_{\text{HS}}^2 =& 0.5\left(K(x_i, x_i)K(x_i^+, x_i^+) + K(x_i, x_i^+)^2 \right) + \\ & 0.5\left(K(x_j, x_j)K(x_j^+, x_j^+) + K(x_j, x_j^+)^2 \right) - \\ & K(x_i, x_j)K(x_i^+, x_j^+) - K(x_i, x_j^+)K(x_i^+, x_j) \end{aligned} \tag{154}$$

which concludes the proof. □

D.3 Proof of Corollary 6.3

Proof. Denote f for the neural network and $g = W\psi(x)$ for the corresponding NTK model from Equation (24). We begin by bounding the difference in the loss in terms of the difference between f and g. The reverse triangle inequality gives

$$\begin{aligned} \left| \bar{\mathcal{L}}(f)^{1/2} - \bar{\mathcal{L}}(g)^{1/2} \right| &= \left| \|C(f) - I\|_F - \|C(g) - I\|_F \right| \\ &\leq \|C(f) - C(g)\|_F \\ &= \left\| \mathbb{E}\left[\frac{1}{2}\left(f(x)f(x^+)^\top + f(x^+)f(x)^\top \right) - \frac{1}{2}\left(g(x)g(x^+)^\top - g(x^+)g(x)^\top \right) \right] \right\|_F \\ &= 0.5\left(\sum_{i,j=1}^{K} \mathbb{E}\left[f_i(x)f_j(x^+) + f_i(x^+)f_j(x) - g_i(x)g_j(x^+) - g_i(x^+)g_j(x) \right]^2 \right)^{1/2} \end{aligned} \tag{155}$$

Adding and subtracting terms of the form $f_i(x) - g_i(x)$, we see that this expression is upper bounded by

$$\begin{aligned}
&0.5 \left(\sum_{i,j=1}^{K} \mathbb{E}\left[\left| f_i(x) f_j(x^+) + f_i(x^+) f_j(x) - g_i(x) g_j(x^+) - g_i(x^+) g_j(x) \right| \right]^2 \right)^{1/2} \leq \\
&0.5 \left(\sum_{i,j=1}^{K} \zeta^2 \mathbb{E}\left[|f_i(x)| + |g_j(x^+)| + |f_j(x^+)| + |g_j(x)| \right]^2 \right)^{1/2} \leq \\
&0.5 \left(\sum_{i,j=1}^{K} \zeta^2 \mathbb{E}\left[|g_i(x)| + |g_j(x^+)| + |g_j(x^+)| + |g_j(x)| + 2\zeta \right]^2 \right)^{1/2} \leq \\
&0.5 \sqrt{\sum_{i,j=1}^{K} \zeta^2 \left(4BS + 2\zeta\right)^2} = \\
&K\zeta(2BS + \zeta)
\end{aligned} \tag{156}$$

where we used the fact that $|f_i(x) - g_i(x)| \leq \zeta$ for all $i \in [K]$ and that $g_i(x) = \langle w_i, \phi(x) \rangle \leq \|w_i\| \|\psi(x)\| \leq BS$ since the data is contained in a ball of radius S in the Hilbert space, and $\|W\| \leq B$. The same derivation works for the (square root of the) empirical loss, so

$$\left| \mathcal{L}(f)^{1/2} - \mathcal{L}(g)^{1/2} \right| \leq K\zeta(2BS + \zeta) \tag{157}$$

From there, we bound $\bar{\mathcal{L}}(f)^{1/2}$ as follows.

$$\begin{aligned}
\bar{\mathcal{L}}(f)^{1/2} &\leq \bar{\mathcal{L}}(g)^{1/2} + K\zeta(2BS + \zeta) \\
&\leq \mathcal{L}(g)^{1/2} + \nu(N, \epsilon) + K\zeta(2BS + \zeta) \qquad \text{with probability} \geq 1 - 2\epsilon. \\
&\leq \mathcal{L}(f)^{1/2} + \nu(N, \epsilon) + 2K\zeta(2BS + \zeta) \\
&\leq \sqrt{\delta} + \nu(N, \epsilon) + 2K\zeta(2BS + \zeta)
\end{aligned} \tag{158}$$

Here, we plugged in the slack term $\nu(N, \epsilon)$ from Theorem 6.1, see Equation (151). This leads to high-probability bounds in the second step. Moreover, we exploited $\mathcal{L}(f) \leq \delta$ in the final step. Squaring both sides, plugging in

$$\left(\sqrt{\delta} + \nu(N, \epsilon) \right)^2 \leq \nu(N, \epsilon, \delta) \tag{159}$$

from Equation (152), and using $(a + b)^2 \leq 2(a^2 + b^2)$ we obtain

$$\bar{\mathcal{L}}(f) \leq 2\nu(N, \epsilon, \delta) + 8K^2\zeta^2(2BS + \zeta)^2 \tag{160}$$

with probability at least $1 - 2\epsilon$. Since the approximation of f through g up to ζ holds with probability at least $1 - \epsilon$ we get the desired statement. □

E DYNAMICS OF LINEAR BARLOW TWINS — CONVERGENCE FROM SMALL INITIALIZATION

We write the empirical linearized Barlow Twins loss as

$$\mathcal{L}(W) = \left\| W\Gamma W^\top - I_K \right\|_F^2 \tag{161}$$

where Γ is the empirical cross-moment matrix between the anchors and the augmentations, either in $\mathbb{R}^d$ or mapped into some RKHS.

It is known from recent works (Simon et al., 2023) that $\frac{\partial \mathcal{L}}{\partial W} = 4(W\Gamma W^\top - I)W\Gamma$ and therefore, under gradient flow, it holds that $\frac{\partial}{\partial t}W = 4(I - W\Gamma W^\top)W\Gamma$. We decompose the matrix W into $W = Q + W_0$, where the rows of W_0 are in the nullspace of Γ, and the rows of Q are in its orthogonal complement. Since $W\Gamma = Q\Gamma$, we have $\mathcal{L}(W) = \mathcal{L}(Q)$, and $\frac{\partial}{\partial t}W_0 = 0$. Therefore, the part of W that starts in the nullspace Γ stays completely unchanged. We may therefore w.l.o.g. restrict ourselves to the setting where Γ has only nonzero eigenvalues. We obtain

$$\frac{\partial}{\partial t}C = 4(I - C)W\Gamma^2 W^\top + 4W\Gamma^2 W^\top (I - C) \tag{162}$$

For an eigendecomposition $\Gamma = UDU^\top$ with diagonal D and orthogonal U, we write $|\Gamma| = U|D|U^\top$ where $|D|_{ii} = |D_{ii}|$. Moreover, denote by $\mu_\Gamma > 0$ the smallest nonzero eigenvalue of Γ in absolute value. Then,

$$\begin{aligned}
\lambda_{min}(W\Gamma^2 W^\top) &= \min_{\|u\|=1} u^\top W|\Gamma|^{1/2}|\Gamma||\Gamma|^{1/2}W^\top u \\
&\geq \mu_\Gamma \||\Gamma|^{1/2}W^\top u\|^2 \\
&\geq \mu_\Gamma \lambda_{min}(W|\Gamma|W^\top) \\
&\geq \mu_\Gamma \lambda_{min}(C)
\end{aligned} \tag{163}$$

Therefore, for any x having unit norm, the quadratic form $x^\top Cx$ satisfies

$$\begin{aligned}
\frac{\partial}{\partial t}\left(x^\top Cx\right) &\geq 8\mu_\Gamma \lambda_{min}(C)(1 - x^\top Cx), \\
\frac{\partial}{\partial t}\left(x^\top Cx\right) &\leq 8\lambda_{max}(\Gamma)\lambda_{max}(C)(1 - x^\top Cx)
\end{aligned} \tag{164}$$

where we used Von-Neumann's trace inequality on the product of the two matrices $I - C$ and $W\Gamma^2 W^\top$. This shows that as long as all eigenvalues of $C(0)$ are contained in $(0, 1)$, the map $t \mapsto x^\top C(t)x$ is strictly increasing for all x, and has an equilibrium at 1. Consequently, we see that the smallest eigenvalue of C is strictly increasing, and that no eigenvalue of C can ever exceed 1. In addition,

$$\begin{aligned}
\frac{\partial}{\partial t}\mathcal{L}(t) &= 2\,\mathrm{Trace}\left((C - I)\frac{\partial C}{\partial t}\right) \\
&= -16\,\mathrm{Trace}\left((C - I)^2 W\Gamma^2 W^\top\right) \\
&\leq -16\lambda_{min}(C)\mu_\Gamma \cdot \mathcal{L}(t) \\
&\leq -16\lambda_{min}(C(0))\mu_\Gamma \cdot \mathcal{L}(t) \\
&= -\eta\mathcal{L}(t)
\end{aligned} \tag{165}$$

where $\eta = 16\lambda_{min}(C(0))\mu_\Gamma$. Therefore, $\mathcal{L}(t) \leq \mathcal{L}(0)\exp(-\eta t)$ and we see that after $T = \frac{-\log\delta}{\eta}$ time, the loss is smaller than δ.

F TECHNICAL RESULTS

F.1 Grönwall's Inequality

Lemma F.1. ***(Grönwall's inequality)*** *Let $u(t)$ and $\beta(t)$ be two continuous functions on an interval $[a, b]$. Suppose $u(t)$ is differentiable in (a, b) and satisfies*

$$\frac{\partial}{\partial t}u(t) \leq \beta(t)u(t) \tag{166}$$

for all $t \in (a, b)$. Then,

$$u(t) \leq u(0) \exp\left(\int_a^t \beta(s)ds\right) \tag{167}$$

for all $t \in [a, b]$.

F.2 Bounding c_0 from Theorem 3.1

Recall that the activation function ϕ and its derivative ϕ' are bounded by c_ϕ and $c_{\phi'}$ respectively, and that we assume $\theta \in B(\theta_0, R)$, and that inputs have unit norm. Thus, for all $k \in [k]$,

$$\begin{aligned}
\|\nabla_\theta f_k(a;\theta\|^2 &= \sum_m \left(\frac{\partial f_k(a;\theta)}{\partial w_{m,k}}\right)^2 + \sum_{m,j}\left(\frac{\partial f_k(a;\theta)}{\partial v_{m,j}}\right)^2 \\
&= \frac{1}{M}\sum_m \phi(v_m^\top a)^2 + \frac{1}{M}\sum_{m,j} w_{m,k}^2 \left(\phi'(v_m^T a)a^{(j)}\right)^2 \\
&\leq c_\phi^2 + \frac{c_{\phi'}^2}{M}\sum_m w_m^2 \|a\|^2 \\
&\leq c_\phi^2 + \frac{c_{\phi'}^2\|\theta\|^2}{M} \\
&\leq c_\phi^2 + \frac{c_{\phi'}^2\|\theta-\theta_0\|^2}{M} + \frac{c_{\phi'}^2\|\theta_0\|^2}{M} \\
&\leq c_\phi^2 + \frac{c_{\phi'}^2 R^2}{M} + \frac{c_{\phi'}^2\|\theta_0\|^2}{M}
\end{aligned} \tag{168}$$

Whenever $R < \sqrt{M}$, this expression is $\mathcal{O}(1)$ independent of the network width, because $\|\theta_0\|^2 = \mathcal{O}(M)$ with high probability.

F.3 McDiarmid's inequality

Lemma F.2. ***(McDiarmid's inequality)*** *Let $(Z_1, \ldots, Z_N) = Z$ be a finite sequence of independent random variables, each with values in $\mathcal{Z}$ and let $\phi : \mathcal{Z}^N \to \mathbb{R}$ be a measurable function such that $|\phi(Z) - \phi(Z')| \leq \nu_n$ whenever Z, Z' differ only differ n-th coordinate. Then, for every $\epsilon > 0$, it holds that*

$$\mathbb{P}\left(\phi(Z) - \mathbb{E}[\phi(Z)] > \epsilon\right) \leq \exp\left(-\frac{2\epsilon^2}{\|\nu\|_2^2}\right) \tag{169}$$

G EXPERIMENTAL DETAILS & ADDITIONAL EXPERIMENTS

All reported experiments have been run 10 times, across random seeds 0-9. The mean as well as standard deviation have been reported through plots for all experiments. All experiments were run using the publicly available Google Colaboratory (`https://colab.research.google.com/drive/12weqAhMLbv5KJ8MlUi80iaEjtDn8lMcv?usp=sharing`). The experiments were run using CUDA enabled PyTorch on a T4 GPU with 15 GB memory. In this appendix, we include some additional experiments.

Experiments for ReLU. We first recreate the three plots provided in the main paper (NTK change till convergence, epochs till convergence and L_2 norm of representation difference) for the case of a single hidden layer neural network with ReLU activation:

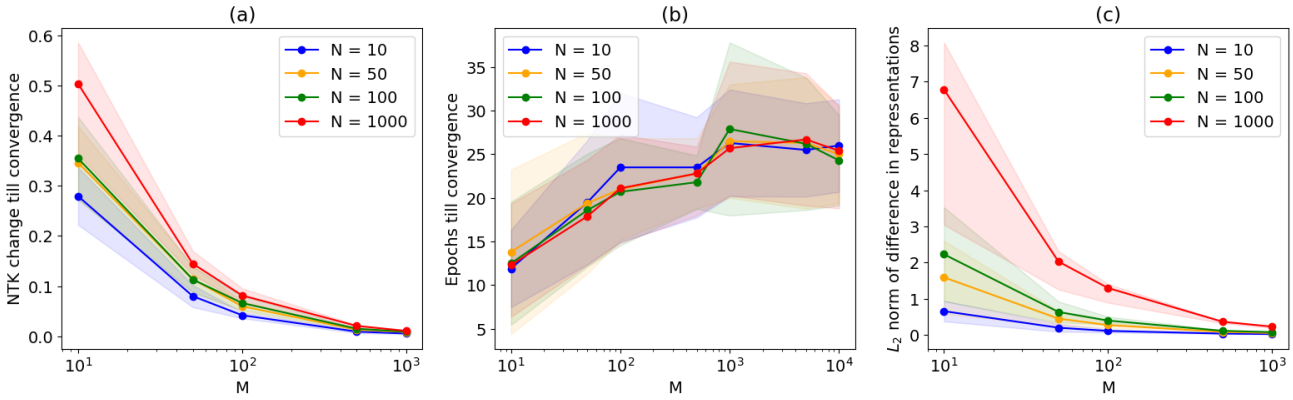

Figure 2: For a fixed sample size N, we plot different quantities for varying network width M. We then vary N and plot: (a) NTK change till convergence (b) Training Epochs till convergence (c) Squared norm of difference between representations of neural network and corresponding kernel model.

As we can see in Figure 2, all three plots closely resemble the corresponding plots with TanH activation. This empirically validates that our claims hold for ReLU activation function as well, for which our derivation was not entirely rigorous (due to the non-differentiability of ReLU at zero).

Deeper ReLU networks. We repeat the same experiment for a 3 hidden layer neural network with ReLU activation:

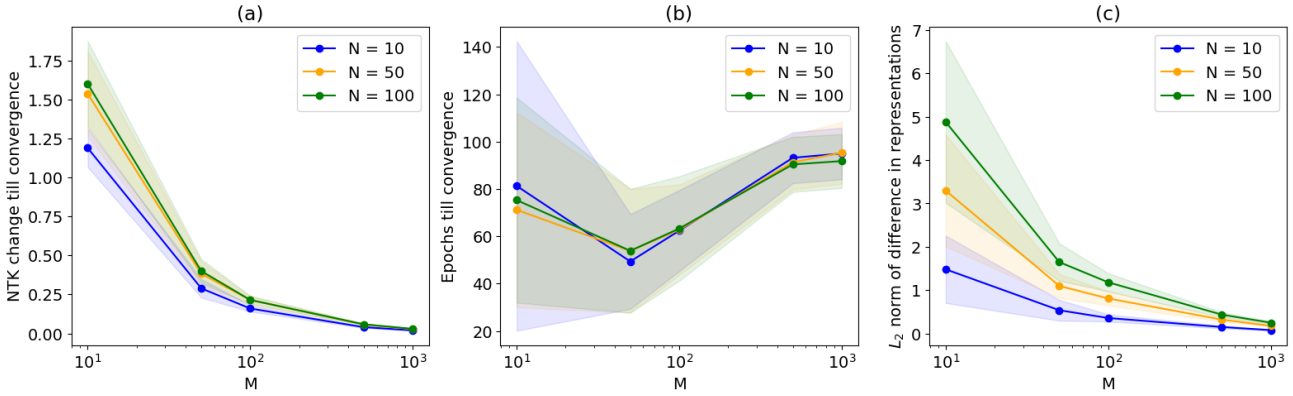

Figure 3: For a fixed sample size N, we plot different quantities for varying network width M. We then vary N and plot: (a) NTK change till convergence (b) Training Epochs till convergence (c) Squared norm of difference between representations of neural network and corresponding kernel model.

While our proof is only for single-hidden layer neural networks, Figure 3 suggests that the analysis holds for deeper networks as well. We leave this for future work to justify theoretically.

Multivariate embeddings. We also verify the constancy of the NTK for multivariate embeddings. Table G shows the change of the NTK for $K = 5$ output dimensions, for varying width M and sample sizes N.

N	M	Change in NTK ($\times 10^2$)
10	50	9.50 ± 2.26
10	100	5.50 ± 1.00
10	500	0.95 ± 0.28
10	1000	0.45 ± 0.12
50	50	17.19 ± 2.35
50	100	10.14 ± 1.17
50	500	2.26 ± 0.42
50	1000	1.06 ± 0.21
100	50	20.47 ± 1.77
100	100	11.83 ± 2.13
100	500	2.73 ± 0.46
100	1000	1.26 ± 0.25

Table 1: Change in NTK values for different N and M

Clearly, as M grows, the change in the NTK until convergence decreases.

Other non-contrastive loss functions. Our code also includes additional experiments for a theory-friendly version of the VIC-Reg loss, as well as a simplified version of BYOL. Just as for Barlow Twins, we observe near-constancy of the NTK at large width.