Quantitative convergence of trained single layer neural networks to Gaussian processes

Eloy Mosig García
Department of Mathematics
University of Pisa
Largo Bruno Pontecorvo, 5, 56127 Pisa PI, Italia
eloy.mosig@phd.unipi.it

Andrea Agazzi
Department of Mathematics and Statistics
University of Bern
Alpeneggstrasse 22, 3012 Bern
andrea.agazzi@unibe.ch

Dario Trevisan
Department of Mathematics
University of Pisa
Largo Bruno Pontecorvo, 5, 56127 Pisa PI, Italia
dario.trevisan@unipi.it

Abstract

In this paper, we study the quantitative convergence of shallow neural networks trained via gradient descent to their associated Gaussian processes in the infinite-width limit. While previous work has established qualitative convergence under broad settings, precise, finite-width estimates remain limited, particularly during training. We provide explicit upper bounds on the quadratic Wasserstein distance between the network output and its Gaussian approximation at any training time $t \geq 0$, demonstrating polynomial decay with network width. Our results quantify how architectural parameters, such as width and input dimension, influence convergence, and how training dynamics affect the approximation error.

1 Introduction

Deep neural networks have achieved remarkable success across a wide range of tasks in computer vision, natural language processing, and scientific computing, often surpassing traditional models by large margins LeCun et al. [2015], Goodfellow et al. [2016]. This empirical progress has sparked substantial interest in understanding the theoretical principles underlying their behavior, particularly in the overparameterized regime, where the number of parameters is larger than the one of training samples.

A major avenue of theoretical investigation in this sense focuses on studying the properties of neural networks in the infinite-width limit. For instance, when the network's parameters at initialization are sampled from a Gaussian distribution, it was shown Neal [1996], de G. Matthews et al. [2018] that the network's output converges to a Gaussian process as width tends to infinity, providing a tractable framework for theoretical analysis.

This perspective was significantly extended by the introduction of the Neural Tangent Kernel (NTK) framework Jacot et al. [2018], allowing to characterize the training dynamics of infinitely wide neural networks under gradient descent in function space. In this limit, the network evolves approximately linearly around its initialization, and training can be understood as kernel regression with a fixed kernel which depends on the the architecture only, and is evaluated on input data. This linearization dramatically simplifies the analysis of generalization and convergence, and has led to a large body of theoretical work on the expressivity and limitations of infinitely wide networks.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

However, the practical relevance of NTK-based analyses hinges on the accuracy of their approximation at finite width. While existing literature has established qualitative convergence of wide neural networks in the NTK regime to Gaussian processes at positive training time Lee et al. [2020], rigorous quantitative results — providing explicit finite-width error bounds — remain scarce. This gap limits the applicability of NTK theory to realistic settings where network width is large but finite.

Indeed, quantitative convergence guarantees are crucial to bridge theory and practice. They allow one to bound the discrepancy between the predictions of a finite-width network and its infinite-width NTK counterpart, thus enabling quantitative uncertainty quantification estimates and the safe deployment of theoretical insights to real-world architectures. Moreover, such estimates reveal how network width, depth, initialization, and training hyperparameters impact the validity of linear approximations during training. These insights are essential for developing principled training strategies and for diagnosing when the NTK regime offers a reliable approximation, or when nonlinear effects beyond NTK must be taken into account.

1.1 Our contributions

This paper provides rigorous quantitative estimates for the convergence of trained shallow neural networks towards their Gaussian process counterparts, measured in terms of quadratic Wasserstein distances. Specifically, we extend previously established convergence bounds at initialization obtained by Basteri and Trevisan [2024], Favaro et al. [2025] and Trevisan [2023] to arbitrary positive training times. Our results deliver explicit convergence rates that decay polynomially with network width, clearly delineating how the approximation error evolves during training.

Concretely, we demonstrate that the distance of the distribution of a shallow neural network's output trained via gradient descent to its Gaussian process approximation at any training time $t > 0$ satisfies explicit quantitative bounds dependent on network width. Our main theorem (3.4) shows that for any test point x , under mild assumptions on the hidden layer width n_1 and on the regularity of the activation function we have:

$$\mathcal{W}_2^2(f_t(x), G_t(x)) = \mathcal{O}\left(\frac{\log n_1}{n_1}\right).$$

We also address long-term training dynamics explicitly, characterizing convergence rates as training time diverges. Indeed, the above result continues to hold on timescales t growing polynomially in the network width n_1 , as discussed in Remark 3.5.

These results significantly refine prior qualitative statements, providing actionable quantitative guidelines on how network parameters and training duration determine the extent to which finite networks emulate their infinite-width limits.

1.2 Related work

The convergence of randomly initialized neural networks to Gaussian processes in the infinite-width limit has been a foundational result in the theory of neural networks. This phenomenon was first suggested by Neal [1996] and later formalized for deep architectures by de G. Matthews et al. [2018]. The key insight that this correspondence extends beyond initialization was introduced by Jacot et al. [2018], who demonstrated that training dynamics in the infinite-width limit are governed by the so-called Neural Tangent Kernel (NTK), a deterministic kernel that linearizes the training trajectory. This sparked significant interest in the use of kernel methods to analyze deep learning models.

Following these developments, several works studied the convergence of finite-width neural networks to their limiting Gaussian processes. In particular, Lee et al. [2020] established that gradient descent dynamics in the NTK regime converge to those of a linearized model. More recently, the works of Basteri and Trevisan [2024] and Trevisan [2023] provided quantitative convergence rates at initialization, measured in Wasserstein distance, which laid the groundwork for a more refined understanding of the finite-width behavior of neural networks. Moreover, in the also recent work by Favaro et al. [2025], additional quantitative results were obtained for total variation and convex distances. Complementary to these works, Bordino et al. [2025] used second-order Poincaré inequalities to derive QCLTs for Gaussian neural networks, obtaining a general but suboptimal convergence rate compared to the optimal n^{-1}.

However, these results were largely confined to the initialization regime. To this day, extensions to the full training trajectory remained limited, with few works addressing how approximation errors evolve over time or depend on architectural features such as width and depth. The present work builds on this gap by extending the quantitative convergence discussed above to trained networks, providing explicit bounds on the Wasserstein distance between the network output and the associated Gaussian process for any positive training time.

From a spectral perspective, the NTK's conditioning plays a central role in understanding convergence rates and generalization. Lower bounds on the smallest eigenvalue of the empirical NTK have been derived under various conditions. For instance, Karhadkar et al. [2024] and Bombari et al. [2022] provide sharp bounds in the context of ReLU and smooth activation functions, respectively.

Additionaly, Carvalho et al. [2025] showed that under very mild assumptions on the non-linearity and non-proportionality of the training data, the analytic NTK is not degenerate. These results are essential for establishing the stability of the gradient flow and, hence, for deriving quantitative convergence guarantees.

Our results are closely related to the work of de G. Matthews et al. [2018], who proved weak convergence of fully-connected BNNs at initialization to a Gaussian process under the metric $\rho(f, f') = \sum_{i\in\mathbb{N}} 2^{-i} \min\{1, |f(x_i) - f'(x_i)|\}$, defined on a countable input set. In our setting, the input set is finite; considering the restriction ρ_F, it follows that convergence in $\mathcal{W}_2$ implies convergence in ρ_F. de G. Matthews et al. [2018] also analyzed convergence under the maximum mean discrepancy (MMD). While MMD is not generally controlled by Wasserstein distances, connections have been established via regularized OT divergences (Feydy et al. [2019], Nietert et al. [2021]). Moreover, Vayer and Gribonval [2023] identified conditions on the RKHS kernel k_H under which $MMD \lesssim \mathcal{W}_2$. Consequently, our bounds also imply MMD convergence under these conditions. The metric ρ_F which offers a notion of pointwise convergence and is oblivious of the tails of the distributions, which helps stablish the results in de G. Matthews et al. [2018]. On the other hand, $\mathcal{W}_2$ captures the geometric structure and scaling of the output space. Finally, while de G. Matthews et al. [2018] address the more general setting deep networks, our analysis focuses on the shallow case, yielding new quantitative rates which improve previous ones in our setting.

A foundational stream of research has shown that, under sufficient overparameterization, gradient-based training of neural networks converges to a global minimum. Seminal results by Du et al. [2019] and Arora et al. [2019] established that for wide two-layer networks with ReLU activation, the empirical NTK remains well-conditioned, enabling convergence via kernel regression. Subsequent advances generalized these results to deep architectures in different directions, such as Allen-Zhu et al. [2019], Zou and Gu [2019], Sankararaman et al. [2020], Wu et al. [2019], Wei et al. [2019], Zou et al. [2020], which provide guarantees that hold with high probability over parameter initalization. These works reinforce that in the NTK regime, the network trajectory stays close to its linearization around initialization. Our contributions align with this body of work and further extend this literature by providing novel finite-sample quantitative bounds on the Wasserstein-2 distance between neural network outputs and their Gaussian process approximations.

1.3 Structure of the paper

Section 2 introduces our notation and mathematical preliminaries. In Section 3, we present our primary theoretical contributions, including our main quantitative convergence theorem. The key technical proofs and intermediate results are outlined succinctly, referring the reader to the relevant lemmas in the Supplementary Material. Numerical experiments validating our theoretical predictions appear in Section 4. Section 5 discusses implications and future research directions.

2 Notation

In the following, given a matrix $A \in \mathbb{R}^{p\times q}$ we will denote by $\|A\|$ its Frobenius norm and by $\|A\|_{op}$ its operator norm. $A_{i\_} \in \mathbb{R}^q$ will denote the i-th row of A and $A_{\_j} \in \mathbb{R}^p$ will denote the j-th column of A, for $1 \le i \le p$ and $1 \le j \le q$. The symbol $\cdot$ denotes the usual matrix product. $\sigma_{\min}(A), \sigma_{\max}(A)$ are the smallest and largest singular value of A; and if $p = q$, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the smallest and largest eigenvalue of A, respectively. For any vector-valued function f, $f(z)_u$ denotes

the u-th coordinate of $f(z)$. For any Polish metric space X, $\mathcal{P}(X)$ will denote the space of Borel probability measures on X.

2.1 Shallow neural networks and associated Gaussian process

We consider a fully connected, shallow (i.e. single hidden layer) neural network of *width* n_1 and input dimension n_0. We assume the output dimension n_2 to be equal to 1 for simplicity. The output of the neural network as a function of its parameters is given by:

$$f(x;\theta) = \frac{1}{\sqrt{n_1}}\Phi\left(\frac{1}{\sqrt{n_0}}x\theta^{(0)}\right)\theta^{(1)} \in \mathbb{R},$$

where $\theta^{(0)} \in \mathbb{R}^{n_0\times n_1}$ and $\theta^{(1)} \in \mathbb{R}^{n_1}$ denote the inner and outer (respectively) *weights* or *parameters*, $\Phi(z)$ is the *activation function*, which acts entrywise on its input, and $x \in \mathbb{R}^{n_0}$ from now on denotes a test input. Note that our model implicitly covers neural networks with biases $b^{(0)} \in \mathbb{R}^{n_1}, b^{(1)} \in \mathbb{R}$, by substituting the input x with $(x,1)$, using the parameters $\tilde{\theta}^{(0)} = (\theta^{(0)}, b^{(0)}) \in \mathbb{R}^{(n_0+1)\times n_1}, \tilde{\theta}^{(1)} = (\theta^{(1)}, b^{(1)}) \in \mathbb{R}^{n_1+1}$ and using the activation function $\tilde{\Phi}(z) = (\Phi(z), 1)$. We will denote by $N = n_0n_1 + n_1$ the total dimension of the parameters. For any ordered set of inputs $X = (x_1, \dots, x_d) \in \mathbb{R}^{n_0\times d}$ we will use the notation $f(X;\theta) = (f(x_1;\theta), \dots, f(x_d;\theta)) \in \mathbb{R}^{n_0\times d}$. In what follows, parameters $\theta^{(0)}_{ij}, \theta^{(1)}_j$, for $1 \le i \le n_0$ and $1 \le j \le n_1$, are drawn independent and identically distributed (i.i.d.) from standard Gaussian random variables at initialization.

We will denote by h_i the *preactivation* of i-th hidden neuron, for $1 \le i \le n_1$:

$$h_i(x;\theta) = \frac{1}{\sqrt{n_0}}x(\theta^{(0)})_{\_i} \in \mathbb{R}.$$

Now we introduce the *Gaussian approximation* G of the neural network f as the centered Gaussian process associated to the covariance operator $\mathcal{K}$ given by:

$$\begin{aligned}\tilde{\mathcal{K}}(x,x') &= \frac{1}{n_0}x^\top x',\\ \mathcal{K}(x,x') &= \mathbb{E}_{(u,v)\sim\mathcal{N}(0,\mathcal{T}(x,x'))}[\Phi(u)\Phi(v)],\end{aligned}$$

with

$$\mathcal{T}(x,x') = \begin{pmatrix}\tilde{\mathcal{K}}(x,x) & \tilde{\mathcal{K}}(x,x')\\ \tilde{\mathcal{K}}(x',x) & \tilde{\mathcal{K}}(x',x')\end{pmatrix}.$$

Explicit convergence rates for the Gaussian approximation at initialization can be found in Basteri and Trevisan [2024], Favaro et al. [2025].

2.2 Training, empirical NTK and limiting kernel

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^{n_0} \times \mathbb{R}$ be a given dataset. Denote by $\mathcal{X} = (x_1, \dots, x_n) \in \mathbb{R}^{n_0\times n}$ the vector of training inputs, and by $y = (y_1, \dots, y_n) \in \mathbb{R}^n$ the vector of outputs. From now on, we will consider the parameters $\theta = \theta_t$ to evolve on training time. Consider the empirical risk for the mean squared error loss:

$$\mathcal{R}_{\mathcal{D}}[f,\theta] = \frac{1}{2}(f(\mathcal{X};\theta) - y)^\top(f(\mathcal{X};\theta) - y).$$

Continuous time gradient descent with respect to this objective function yields the following dynamics for the parameters and the network:

$$\begin{aligned}\frac{\partial}{\partial t}\theta_t &= -\nabla_\theta\mathcal{R}_{\mathcal{D}}[f,\theta_t] = -\nabla_\theta f(\mathcal{X};\theta_t)(f(\mathcal{X};\theta_t) - y),\\ \frac{\partial}{\partial t}f(\mathcal{X};\theta) &= -\nabla_\theta f(\mathcal{X};\theta_t)^\top\nabla_\theta f(\mathcal{X};\theta_t)(f(\mathcal{X};\theta_t) - y),\\ \frac{\partial}{\partial t}f(x;\theta) &= -\nabla_\theta f(x;\theta_t)^\top\nabla_\theta f(\mathcal{X};\theta_t)(f(\mathcal{X};\theta_t) - y).\end{aligned} \tag{2.1}$$

In Lee et al. [2020] the authors showed that the dynamics in (2.1) converge to those of a linearized network, which we now introduce.

Definition 2.1. Given a neural network f we define its associated *linearized network*:

$$f^{\text{lin}}(x;\theta_t) = f(x;\theta_0) + \nabla_\theta f(x;\theta_0)|_{\theta=\theta_0}\omega_t$$

with the change of parameters $\omega_t = \theta_t - \theta_0$. In the following, we will also consider the *linearized gradient flow*, given by

$$\frac{\partial}{\partial t}\overline{\theta}_t = -\nabla_\theta f(\mathcal{X};\theta_0)\cdot(f^{\text{lin}}(\mathcal{X};\overline{\theta}_t) - y).$$

The linearized network is known to approximate arbitrarily well the real training dynamics in the wide limit under some stability conditions and positive-definiteness of the NTK (see Theorem 5.1 in Bartlett et al. [2021] and Theorem 2.2 in Chizat et al. [2019]). In the Supplementary Material we prove an analogue result (Proposition B.9) adapted to our setting, In particular, our result applies to shallow networks with inner and outer weights and for the MSE loss without scaling over the number of training points, as opposed to the cited results.

An alternative formulation of this asymptotic linearzation phenomenon is provided in Jacot et al. [2018], where the neural tangent kernel (NTK) was introduced. The authors showed that under Gaussian initialization, the dynamics (2.1) are governed by this operator which we now define in our setting. Define the *empirical kernel*, or *NTK* $k\colon \mathbb{R}^{n_0}\times\mathbb{R}^{n_0}\times\mathbb{R}^N\to\mathbb{R}$ as:

$$k(x,x';\theta) = \nabla_\theta f(x;\theta)\nabla_\theta f(x';\theta)^\top,$$

where $\theta\in\mathbb{R}^N$ denotes the matrix of *parameters*. The empirical kernel at the hidden layer can also be defined as a function of $\mathbb{R}^{n_0}\times\mathbb{R}^{n_0}\times\mathbb{R}^N\to\mathbb{R}^{n_1\times n_1}$:

$$\tilde{k}_{uv}(x,x';\theta) = \nabla_{\theta^{(0)}}h(x;\theta)_u\nabla_{\theta^{(0)}}h(x';\theta)_v^\top,\quad 1\le u,v\le n_1,$$

During training, the parameter θ_t will be omitted when it is clear from the context and we will simply write $k_t(x,x') = k_t(x,x';\theta_t)$ and $\tilde{k}_t(x,x') = \tilde{k}_t(x,x';\theta_t)$.

The convergence of the NTK to this limiting kernel was first proven by the authors of Jacot et al. [2018]. Define the *analytical*, or *limiting kernel* k_∞ as follows:

$$k_\infty(x,x') = \mathcal{K}(x,x') + \tilde{\mathcal{K}}(x,x')\mathbb{E}_{(u,v)\sim\mathcal{N}(0,\mathcal{T}(x,x'))}[\Phi'(u)\Phi'(v)],$$

with $\mathcal{T}(x,x') = \begin{pmatrix}\tilde{\mathcal{K}}(x,x) & \tilde{\mathcal{K}}(x,x')\\ \tilde{\mathcal{K}}(x',x) & \tilde{\mathcal{K}}(x',x')\end{pmatrix}$. From now on, let $k_\infty = k_\infty(\mathcal{X},\mathcal{X})$ denote the limiting kernel valued on the training set.

Note that, in the linearized regime, Equation (2.1) can be solved analytically. In effect, letting $k_{\mathcal{X}\mathcal{X}} = \nabla_\theta f(\mathcal{X};\theta_0)\nabla_\theta f_0(\mathcal{X};\theta_0)^\top$ and $k_{x\mathcal{X}} = \nabla_\theta f(x;\theta_0)\nabla_\theta f(\mathcal{X};\theta_0)^\top$, the gradient flow equations (2.1) for f^{lin} can be rewritten as:

$$\begin{aligned}\frac{\partial}{\partial t}\overline{\theta}_t &= -\nabla_\theta f(\mathcal{X};\theta_0)^\top(f^{\text{lin}}(\mathcal{X};\overline{\theta}_t) - y),\\ \frac{\partial}{\partial t}f^{\text{lin}}(\mathcal{X};\overline{\theta}_t) &= -k_{\mathcal{X}\mathcal{X}}(f^{\text{lin}}(\mathcal{X};\overline{\theta}_t) - y),\\ \frac{\partial}{\partial t}f^{\text{lin}}(x;\overline{\theta}_t) &= -k_{x\mathcal{X}}(f^{\text{lin}}(\mathcal{X};\overline{\theta}_t) - y).\end{aligned}\tag{2.2}$$

The inverse of matrix $k_{\mathcal{X}\mathcal{X}}$, which is random in the initialization of the network and may not be positive definite for some θ, appears in the solution of Equation (2.2). Thus we introduce the following auxiliary operator:

Definition 2.2. For any symmetric, invertible matrix $B\in\mathbb{R}^{n\times n}$ and for any $t>0$, define the $n\times n$ real matrix

$$I_t(B) = (\mathbb{1}_n - e^{-Bt})B^{-1}.$$

Note that $I_t(B)$ is invertible and symmetric since the matrix exponential of B is positive definite. The operator I_t can be extended to general symmetric matrices in the following way: define, for each $a\in\mathbb{R}$,

$$I_t(a) = \int_0^t e^{-as}ds = \begin{cases}\frac{1-e^{-at}}{a} & \text{if } a\neq 0,\\ t & \text{if } a = 0.\end{cases}$$

Let $B \in \mathbb{R}^{n\times n}$ be symmetric, not necesarily non-degenerate. Consider the eigenvalue decomposition of B, $B = UDU^\top$ with $D = \text{diag}(a_1, \dots a_n)$ and U orthogonal, then put $I_t(B) = U\text{diag}(I_t(a_1), \dots I_t(a_n))U^\top$.

Lemma A.1 shows some properties and well-posedness of the operator I_t defined in Definition 2.2. The matrix $I_t(B)$ can be thought of as the analytic continuation of the matrix function $(\mathbb{1}_n - e^{-Bt})B^{-1}$, for any $B \in \mathbb{R}^{n\times n}$ symmetric. With this definition, the solution to (2.2) reads:

$$f^{\text{lin}}(\mathcal{X}; \overline{\theta}_t) = \exp(-k_{\mathcal{X}\mathcal{X}}t)f(\mathcal{X}; \theta_0) + (\mathbb{1}_n - \exp(-k_{\mathcal{X}\mathcal{X}}t))y, \tag{2.3}$$

$$f^{\text{lin}}(x; \overline{\theta}_t) = f(x; \theta_0) - k_{x\mathcal{X}}I_t(k_{\mathcal{X}\mathcal{X}})(f(x; \theta_0) - y). \tag{2.4}$$

The computation leading to this expression is contained in Supplementary Material A.

In Lee et al. [2020], the authors showed that when f is linear in its parameters, its output distribution at a test point $x \in \mathbb{R}^{n_0}$ converges weakly, as n_1 diverges, to a Gaussian process G_t with mean and covariance given by:

$$\begin{aligned}\mu_t(x) &= k_\infty(x, \mathcal{X})I_t(k_\infty)y,\\ \Sigma_t(x, x') &= \mathcal{K}(x, x') - \mathcal{K}(x, \mathcal{X})I_t(k_\infty)k_\infty(\mathcal{X}, x') - k_\infty(x, \mathcal{X})I_t(k_\infty)\mathcal{K}(\mathcal{X}, x')\\ &\quad + k_\infty(x, \mathcal{X})I_t(k_\infty)\mathcal{K}(\mathcal{X}, \mathcal{X})I_t(k_\infty)k_\infty(\mathcal{X}, x'),\end{aligned} \tag{2.5}$$

for every positive training time t. For the sake of completeness we included a proof for the above formula in Supplementary Material A.1.

The solution of (2.2) completely characterizes the dynamics of the Gaussian process G_t for any time $t \geq 0$, as a consequence of the Central Limit Theorem. In the rest of the paper we will assume that $k_\infty(\mathcal{X}, \mathcal{X})$ is positive definite. This is a mild assumption and holds in a very general setting. Indeed, the authors of Carvalho et al. [2025] showed that when the training data is in general position in $\mathbb{R}^{n_0}$ and Φ is not a polynomial the smallest eigenvalue of $k_\infty(\mathcal{X}, \mathcal{X})$ is strictly greater than zero.

3 Assumptions and main result

In this section we state our assumptions and main result.

Assumption 1. The parameters $\theta_{ij}^{(0)}, \theta_j^{(1)}$, for $1 \leq i \leq n_0$ and $1 \leq j \leq n_1$, are drawn independent and identically distributed (i.i.d.) from standard Gaussian random variables at initialization.

Assumption 2. The limiting kernel $k_\infty(\mathcal{X}, \mathcal{X})$ is positive definite.

Assumption 3. Φ and Φ' are Lipschitz continuous and bounded.

Assumption 4. For some fixed $r \geq 5$ the following inequality holds:

$$\frac{4\|\mathcal{X}\|(\sqrt{5}\|\Phi\|_\infty + \|y\|)}{\sqrt{n_1 n_0}}\left(\|\Phi'\|_\infty + \text{Lip}\Phi + \frac{\|\mathcal{X}\|\text{Lip}\Phi'\sqrt{r\log n_1}}{\sqrt{n_0}}\right) < \lambda_{\text{min}}^\infty,$$

Remark 3.1. Note that these are rather mild assumptions. Assumption 2 is mild and holds in a very general setting. Indeed, the authors of Carvalho et al. [2025] showed that when the training data is in general position in $\mathbb{R}^{n_0}$ and Φ is not a polynomial the smallest eigenvalue of $k_\infty(\mathcal{X}, \mathcal{X})$ is strictly greater than zero. Assumption 3 is standard in literature and is satisfied by most activation functions, such as the sigmoid function and other logistic activations, hyperbolic tangent, Gaussian activation or sinusoid, among others. The ReLu family is a notable exception, although we expect our result to also hold in that case. As for Assumption 4, notice that the left hand side tends to zero as $\min\{n_1, n_0\}$ grows. This implies that our hypothesis is satisfied for sufficiently large n_0 or n_1. In particular, it holds for sufficiently overparametrized networks.

Remark 3.2. Assumption 4 may appear somewhat artificial, so we provide an informal intuition on its use. This assumption is needed to control the fluctuations of $\|k_0 - k_\infty\|$ with the (deterministic) smallest eigenvalue of the limiting kernel. This enables the use of Proposition B.9, which provides a quenched estimation of the L^2 distance between f and f^{lin}, which in turn plays a central role in the proof of Theorem 3.4. In particular, the L^2 norm of this difference is controlled with a function of the Lipschitz constant of the Jacobian $\nabla_\theta f$ at initialization, and this Lipschitz constant is estimated with an expression in terms of $\|k_0 - k_\infty\|$ bounded by the left-hand side in Assumption 4. This is the content of Lemmas B.15 and B.16.

To measure how well the Gaussian process G_t approximates the network f at time t, we will use a well-known family of metrics between probability distributions, the Wasserstein distances:

Definition 3.3. Let $p \in [1, \infty]$ and let μ, ν be two probability measures defined on a Polish space (M, d_M) with finite p-moment. The p-Wasserstein distance between μ and ν is given by

$$\mathcal{W}_p(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu,\nu), X\sim\mu, Y\sim\nu} \left(\mathbb{E}_\gamma \left[d_M(X, Y)^p\right]\right)^{\frac{1}{p}},$$

where $\Gamma(\mu, \nu)$ denotes the set of joint probability measures γ defined on $M \times M$ with marginal laws μ and ν. With a slight abuse of notation, we will often write $\mathcal{W}_p(X, Y) = \mathcal{W}_p(\mu, \nu)$ for any $X \sim \mu$ and $Y \sim \nu$.

Now we can state our main theorem:

Theorem 3.4. *Under Assumptions 1, 2, 3 and 4, for each test point $x \in \mathbb{R}^{n_0}$ there exist positive constants a_1 and a_2 not depending on n_0, n_1 nor t such that:*

$$\mathcal{W}_2^2(f(x; \theta_t), G_t(x)) \leq r \left(\frac{a_1 \log n_1}{(\lambda_{\text{min}}^\infty)^3 n_1 n_0} + \frac{a_2 n_0}{(\lambda_{\text{min}}^\infty)^r n_1^{\frac{r}{4}}} (1 + t^8) \right).$$

Here r is the constant appearing in Assumption 4.

Remark 3.5. Note that our result is not limited to fixed training time t, but holds for values of t growing polynomially on n_1. Indeed, provided that t grows at most polynomially in n_1, the constant r can be chosen arbitrarily big making the term dependent on time negligible. In particular, as long as t grows polynomially on n_1, a sufficiently big r can be chosen so that the right-hand side in Theorem 3.4 tends to zero as n_1 diverges.

The term t^8 in the right hand side of the inequality is due to Lemma B.12 in Supplementary Material B. Lemma B.12 provides upper bounds of the entries $\theta_t^{(0)}$ and $\theta_t^{(1)}$ that account for perturbations that occur on tail events with respect to the initalization distribution (i.e. in the "bad event" S^C). A finer control is possible if one is interested in a result that holds S on only, which has high probability, such as the ones in Bartlett et al. [2021], Chizat et al. [2019]. This finer control corresponds to Theorem B.9.

3.1 Sketch of proof of Theorem 3.4

The proof of our main theorem is as follows: we bound by triangle inequality

$$\mathcal{W}_2(f(x; \theta_t), G_t(x)) \leq \mathcal{W}_2(f(x; \theta_t), f^{\text{lin}}(x; \overline{\theta}_t)) + \mathcal{W}_2(f^{\text{lin}}(x; \overline{\theta}_t), G_t(x)),$$

and proceed to control the two terms appearing in the right-hand side separately.

To bound the first summand, we partition $\mathbb{R}^N$ into a "good" event $S \subset \mathbb{R}^N$, in which the assumptions of Proposition B.9 hold, along some other concentration properties of the parameters, and a "bad" event S^C in which they do not; so the estimation of the first summand reduces to the estimation of integrals over S and S^C, respectively. Moreover, as n_1 diverges, $\mathbb{P}(S)$ converges to 1. Proposition B.9 consists on an upper bound of $\|f(x; \theta_t) - f^{\text{lin}}(x; \overline{\theta}_t)\|^2$ on this "good event"; which is a version of Theorem 5.1 in Bartlett et al. [2021] adapted to our setting. This allows to bound the first integral. In S^C the strategy is to show that the measure of S^C decreases faster than how the upper bounds in Theorem B.10 grow. Theorem B.10 provides estimations for $\|f(x; \theta_t) - f^{\text{lin}}(x; \overline{\theta}_t)\|^2$ that are rougher than the ones in Proposition B.9 in the sense that do not vanish in the wide limit, but hold in S^C. We estimate the second integral by partitioning S^C into countably many discs parametrized by $\gamma \in \mathbb{N}$, and summing over γ while exploiting concentration inequalities that hold on each disc. The result of this estimation is summarized in the following Theorem:

Proposition 3.6. *On the hypothesis of Theorem 3.4, there exist positive constants a_1 and a_2 not depending on n_0, n_1 nor t such that:*

$$\mathcal{W}_2^2(f(x; \theta_t), f^{\text{lin}}(x; \overline{\theta}_t)) \leq r \left(\frac{a_1 \log n_1}{(\lambda_{\text{min}}^\infty)^3 n_1 n_0} + \frac{a_2 n_0}{(\lambda_{\text{min}}^\infty)^r n_1^{\frac{r}{4}}} (1 + t^8) \right).$$

The dependence on time of the right-hand side of statement of the Theorem comes from Lemma B.12. In particular, in the "bad" event S^C a lower bound of the smallest eigenvalue of the random matrix k_0 is not available, and hence by Definiton 2.2 the sharpest upper bound for $\|I_t(k_0)\|$ is t.

The second summand, instead, is estimated with the following result:

Proposition 3.7. *Let f^{lin} be the linearization of f, and let $x \in \mathbb{R}^{n_0}$ be a test point. Then, under assumptions 1 and 3, there exist positive constants $C, \overline{C}, \overline{D}$ not depending on n_1 nor t such that:*

$$\mathcal{W}_2^2(f^{\text{lin}}(\mathcal{X}; \overline{\theta}_t), G_t(\mathcal{X})) \leq \frac{1}{n_1} C(t+1) e^{-\lambda_{\min}^{\infty} t},$$
$$\mathcal{W}_2^2(f^{\text{lin}}(x; \overline{\theta}_t), G_t(x)) \leq \frac{1}{n_1} (\overline{C} + \overline{D}(t+1) e^{-\lambda_{\min}^{\infty} t}).$$

The first statement in Proposition 3.7 is proven by bounding $\frac{\partial}{\partial t}\|f^{\text{lin}}(\mathcal{X}; \overline{\theta}_t) - G_t(\mathcal{X})\|^2$ with Young's inequality and gradient flow equations. Next, we apply Grönwall and Hölder's inequalities to decompose the problem in some expected values of L^2 and L^4 norms of the differences between kernels and between f^{lin} and G, both at initialization. Lastly, the main result in Basteri and Trevisan [2024] and Proposition B.4 providing L^p estimations for the difference between the empirical NTK and the limiting kernel complete the proof. The proof for the second statement in Proposition 3.7 uses the bound for training points in a triangular system of differential inqualities, inherited by the solution of Equation (2.2).

Theorem 3.4 and Propositions 3.6 and 3.7 are proven in full detail in the Supplementary Material C and D.

4 Numerical Experiments

We conduct some numerical experiments in Figure 1 to support our theoretical results. In both experiments, t is taken as the product of the learning rate and the number of iterations or epochs. For simplicity, we considered training and test inputs on the real line. The training set and test set were drawn from a uniform distribution on a fixed interval, and the labels y of the training points correspond to a sine function with additive noise. The code is available at `https://github.com/emosig/quantitative_gaussian_trainedNN`.

4.1 Experiment 1: Gaussian approximation of f

The leftmost and center plots in Figure 1 represent 100 trained shallow neural networks with sigmoid activation of width $n_1 = 700$ on the leftmost plot and $n_1 = 1000$ in the central plot. The networks have been trained for $2 \cdot 10^4$ epochs with learning rates of $\frac{1}{700}$ and $\frac{7}{1000}$ (hence $t \approx 28.571$ on the leftmost plot and $t = 140$ in the central one) to fit two training points, corresponding to different random seeds on each case. Together with the networks, the plot depicts the mean (in black) of G_t and a 95% confidence interval (in grey) over 200 equally spaced test points on the interval $[-10, 10]$. The networks were programmed with PyTorch 2.6.0 Paszke et al. [2019], and the Gaussian process G_t was constructed using the library *neural tangents 0.6.5* Novak et al. [2020] for the kernels $\mathcal{K}$ and k_∞, needed to construct μ_t and Σ_t in (2.5). The operator $I_t(B)$ was programmed by solving the linear system of equations $BX = \mathbb{1}_n - e^{-Bt}$ with the *linalg* package of *NumPy* Harris et al. [2020].

4.2 Experiment 2: $\mathcal{W}_2(f(x; \theta_t), G_t(x))$ decays with n_1 for any x

In our second experiment (Figure 1, right), we compute the quadratic Wasserstein distance between the trained shallow network and G_t for a variety of widths, ranging from 2 to 256. To do this we drew 10^4 samples of G_t, given by the mean μ_t and covariance Σ_t in (2.5), which were computed with the *neural tangents 0.6.5* library Novak et al. [2020]. Then, we trained indepently over a single training point, for 100 epochs and with a learning rate of 0.1 (hence $t = 10$), 10^4 neural networks for each width and then calculated the empirical Wasserstein distance with the *Python Optimal Transport 0.9.5* library Flamary et al. [2021]. As in Experiment 1, *PyTorch* was used to construct the neural networks, and the activation chosen for f and G is once again the sigmoid.

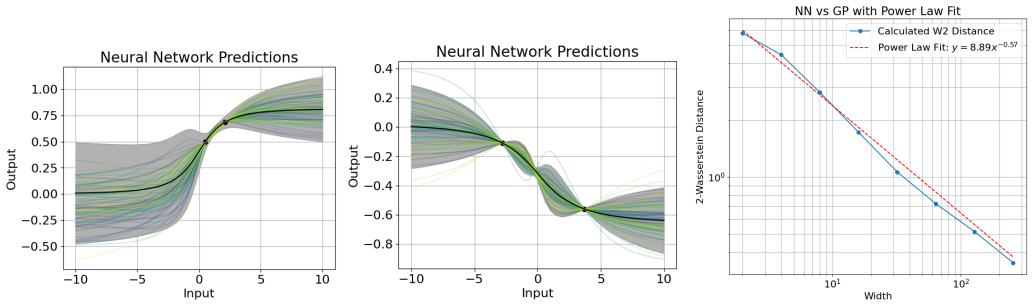

Figure 1: The Gaussian process approximates the neural networks during training (left and center images), and it converges in 2-Wasserstein space to f_t (right image). On the rightmost image, the blue points represent the empirical Wasserstein distance between f and G for increasing widths, and the red plot is the power-law fit between the blue points.

Discussion on the choice of n_1 and the number of samples The authors of Fournier and Guillin [2015] provide an estimation of the error between a probability measure μ and its empirical counterpart $\hat{\mu}_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{X_i}$, given an i.i.d. sequence $(X_i)_{i=1}^N$, $X_i \sim \mu$. More precisely, Theorem 1 in Fournier and Guillin [2015], for $p = 2$ and $n_0 = 1$, provided that μ has finite third moment, reads:

$$\mathbb{E}[\mathcal{W}_2^2(\mu, \hat{\mu}_N)] \leq \frac{c_1}{\sqrt{N}},$$

for some constant c_1. This estimation applied to f and G can be combined with our main result to find a minimal ratio of samples per width needed for our numerical estimations to be noise-free. There exist constants c_2, c_3 not depending on n_1, N nor t such that:

$$\mathcal{W}_2^2(f(x;\theta_t), G_t(x)) \approx \mathcal{W}_2^2(\widehat{f(x;\theta_t)}_N, \widehat{G_t(x)}_N) + \frac{c_2}{\sqrt{N}} \leq \frac{c_3 \log n_1}{n_1}.$$

Hence our computations make sense when $N \gg \left(\frac{n_1}{\log n_1}\right)^2$. For $N = 10^4$, this means an upper bound for the widths for which our experiments make sense is, approximately, 650.

5 Discussion

In this paper we provided quantitative convergence rates in 2-Wasserstein distance between trained shallow neural networks with standard Gaussian initialization and an appropriate Gaussian process, for any positive training time. This was proven for Lipschitz, bounded activations with bounded derivative and for sufficiently large hidden layer width. We now address some limitations of our work and possible future research directions:

1. Our main result is not uniform in time. Although the dependence on time can be minimized at the price of including a sufficiently big multiplicative constant in the right-hand side of our inequality as discussed in Remark 3.5, a general result holding uniformly in $t > 0$, in the limit when t tends to infinity exponentially on n_1 is not available.

   This dependence on time could be related to the transition from the NTK regime to a feature-learning regime, as suggested by the work of Huang and Yau [2020]. Their analysis, however, does not address the tails of the distributions, which in our proof correspond to the set S^C and are responsible for the t^8 scaling. Moreover, Yang and Hu [2021] show that under standard and NTK parameterizations, wide networks cannot perform feature learning in the infinite-width limit.

   This suggests that our observed t^8 scaling might reflect the boundary of the NTK regime: in the "bad event" S^C or for sufficiently large times, the training dynamics may drift into feature-learning, where purely kernel-based control breaks down. Our main result remains consistent with works such as Bartlett et al. [2021], Chizat et al. [2019], which hold with high probability, whereas our analysis explicitly incorporates the contribution of the event S^C. At present, it is unclear whether the t^8 scaling is sharp.

   We would like to address this problem in future work.

2. The bound in our Theorem 3.4 depends on the test point x. This dependence is explicitly stated on the proof of the auxiliary results Proposition 3.7 and Theorem B.10 in the Supplementary Material. Locally uniform bounds on the test point x might follow from functional inequalities such as the ones found by Favaro et al. [2025] if extended to the NTK regime.
3. We conjecture that our main result remains valid even without Assumption 3, as suggested by our numerical experiments with the ReLU activation. In this work, we deliberately focused on a specialized setting with mild hypotheses to obtain a novel and technically precise result while maintaining a clear exposition. Future research will aim to relax the regularity assumptions on the activation and extend our analysis to a more general setting.
4. When Assumption 1 and 2 hold, $k_\infty(\mathcal{X}, \mathcal{X})$ is strictly positive definite and Φ is Lipschitz, the rate of convergence at initialization found in Trevisan [2023] for the squared 2-Wasserstein distance is of n_1^{-2}. This fact suggests that, in the proof of our main Theorem, a better estimation of $\mathcal{W}_2(f(x;\theta_t), f^{\text{lin}}(x;\overline{\theta}_t))$ can be found; either by improving Proposition 3.6 or by improving the estimation of the Lipschitz constant and the norm of the Jacobian $\nabla_\theta f(x;\theta_0)$ in the "good event" of the proof of Theorem 3.4, possibly by choosing a more restrictive "good event" that still makes the infinite sums in the proof of Theorem 3.4 converge. This last possibility calls for finer concentration inequalities for the parameters and the difference between the empirical NTK and its infinite-width limit.
5. Our results could be extended to deep, fully connected neural networks, as done for initialization in Trevisan [2023], Basteri and Trevisan [2024] and Favaro et al. [2025]. We hypothesize that Proposition 3.7 can be easily adapted to the deep setting exploiting recursive characterizations of k_t and k_∞ as the ones available in Jacot et al. [2018], Nguyen et al. [2021] or Lee et al. [2020]. The other half of our proof, though, relies in Proposition 3.6, which would need a new proof for the deep setting.
6. Another desirable step could be to study how well our result extends to other architectures, such as convolutional neural networks or the more modern attention-based architectures. This approach is present in Yang and Littwin [2021] but to the best of our knowledge no quantitative results in this direction are available.

6 Acknowledgements

E.M.G. and D.T. are members of GNAMPA group of the Istituto Nazionale di Alta Matematica (INdAM). All authors acknowledge the support of GNAMPA Project CUP E53C22001930001. E.M.G also gratefully acknowledges the hospitality and support of the Institute of Mathematical Statistics and Actuarial Sciences at University of Bern during part of this work. This research was supported by a Ph.D. fellowship funded by the Italian Ministry of University and Research (MUR) under the PNRR program "Transizioni digitali e ambientali (TDA)" (D.M. n. 118/2023, CUP I51J23000400007), within the project "Limiti di scala di dinamiche stocastiche."

D.T. acknowledges the MUR Excellence Department Project awarded to the Department of Mathematics, University of Pisa, CUP I57G22000700001, the HPC Italian National Centre for HPC, Big Data and Quantum Computing - Proposal code CN1 CN00000013, CUP I53C22000690001, the PRIN 2022 Italian grant 2022WHZ5XH - "understanding the LEarning process of QUantum Neural networks (LeQun)", CUP J53D23003890006, the project G24-202 "Variational methods for geometric and optimal matching problems" funded by Università Italo Francese. Research also partly funded by PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", funded by the European Commission under the NextGeneration EU programme. This research benefitted from the support of the FMJH Program Gaspard Monge for optimization and operations research and their interactions with data science.

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, page 242–252. PMLR, 2019. URL `https://proceedings.mlr.press/v97/allen-zhu19a.html`.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 322–332. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/arora19a.html.

Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta Numerica*, 30:87–201, 2021. doi: 10.1017/S0962492921000027.

Andrea Basteri and Dario Trevisan. Quantitative gaussian approximation of randomly initialized deep neural networks. *Machine Learning*, 113:1–21, 06 2024. doi: 10.1007/s10994-024-06578-z.

Simone Bombari, Mohammad Hossein Amani, and Marco Mondelli. Memorization and optimization in deep neural networks with minimum over-parameterization. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=x8DNliTBSYY.

Alberto Bordino, Stefano Favaro, and Sandra Fortini. Non-asymptotic approximations of gaussian neural networks via second-order poincaré inequalities, 2025. URL https://arxiv.org/abs/2304.04010.

Luís Carvalho, João L. Costa, José Mourão, and Gonçalo Oliveira. The positivity of the neural tangent kernel. *SIAM Journal on Mathematics of Data Science*, 7(2):495–515, 2025. doi: 10.1137/24M1659534. URL https://doi.org/10.1137/24M1659534.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. *On lazy training in differentiable programming*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Alexander G. de G. Matthews, Mark Rowland, Jiri Hron, Richard E. Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. *ArXiv*, abs/1804.11271, 2018. URL https://api.semanticscholar.org/CorpusID:13757156.

Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks, 2019.

S. Favaro, B. Hanin, D. Marinucci, I. Nourdin, and G. Peccati. Quantitative clts in deep neural networks. *Probability Theory and Related Fields*, 191(3):933–977, Apr 2025. ISSN 1432-2064. doi: 10.1007/s00440-025-01360-1. URL https://doi.org/10.1007/s00440-025-01360-1.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89 of *Proceedings of Machine Learning Research*, pages 2681–2690. Proceedings of Machine Learning Research, April 2019. URL https://proceedings.mlr.press/v89/feydy19a.html.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, AurÃ©lie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, LÃ©o Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL http://jmlr.org/papers/v22/20-451.html.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, Aug 2015. ISSN 1432-2064. doi: 10.1007/s00440-014-0583-7. URL https://doi.org/10.1007/s00440-014-0583-7.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández

del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

Jiaoyang Huang and Horng-Tzer Yau. Dynamics of deep neural networks and neural tangent hierarchy. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4542–4551. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/huang20l.html.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf.

Kedar Karhadkar, Michael Murray, and Guido Montufar. Bounds for the smallest eigenvalue of the NTK for arbitrary spherical data of arbitrary dimension. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=mHVmsy9len.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302 – 1338, 2000. doi: 10.1214/aos/1015957395. URL https://doi.org/10.1214/aos/1015957395.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015. ISSN 1476-4687. doi: 10.1038/nature14539. URL https://doi.org/10.1038/nature14539.

Jaehoon Lee, Lechao Xiao, Samuel S Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent *. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12): 124002, December 2020. ISSN 1742-5468. doi: 10.1088/1742-5468/abc62b. URL http://dx.doi.org/10.1088/1742-5468/abc62b.

R. M. Neal. *Bayesian Learning for Neural Networks, Vol. 118 of Lecture Notes in Statistics*. Springer-Verlag, 1996.

Quynh Nguyen, Marco Mondelli, and Guido F. Montúfar. Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep relu networks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8119–8129. PMLR, 2021. URL http://proceedings.mlr.press/v139/nguyen21g.html.

Sloan Nietert, Ziv Goldfeld, and Kengo Kato. Smooth p-wasserstein distance: Structure, empirical approximation, and statistical applications. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8172–8183. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/nietert21a.html.

Roman Novak, Lechao Xiao, Jiri Hron, Jaehoon Lee, Alexander A. Alemi, Jascha Sohl-Dickstein, and Samuel S. Schoenholz. Neural tangents: Fast and easy infinite neural networks in python. In *International Conference on Learning Representations*, 2020. URL https://github.com/google/neural-tangents.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Karthik A. Sankararaman, Soham De, Zheng Xu, W. Ronny Huang, and Tom Goldstein. The impact of neural network overparameterization on gradient confusion and stochastic gradient descent. In Marina Meila and Tong Zhang, editors, *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, page 8469–8479. PMLR, 2020. URL https://proceedings.mlr.press/v119/sankararaman20a.html.

Dario Trevisan. Wide deep neural networks with gaussian weights are very close to gaussian processes, 2023.

Titouan Vayer and RÃ©mi Gribonval. Controlling wasserstein distances by kernel norms with application to compressive statistical learning. *Journal of Machine Learning Research*, 24(149): 1–51, 2023. URL http://jmlr.org/papers/v24/21-1516.html.

Cédric Villani. *Optimal transport – Old and new*, volume 338, pages xxii+973. 01 2008. doi: 10.1007/978-3-540-71050-9.

Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets v.s. their induced kernel. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/8744cf92c88433f8cb04a02e6db69a0d-Paper.pdf.

Xiaoxia Wu, Simon S. Du, and Rachel Ward. Global convergence of adaptive gradient methods for an over-parameterized neural network, 2019. URL https://arxiv.org/abs/1902.07111.

Greg Yang and Edward J. Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11727–11737. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/yang21c.html.

Greg Yang and Etai Littwin. Tensor programs iib: Architectural universality of neural tangent kernel training dynamics. In *International conference on machine learning*, pages 11762–11772. PMLR, 2021.

Difan Zou and Quanquan Gu. *An improved analysis of training over-parameterized deep neural networks*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine Learning*, 109(3):467–492, Mar 2020. ISSN 1573-0565. doi: 10.1007/s10994-019-05839-6. URL https://doi.org/10.1007/s10994-019-05839-6.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes] .

   Justification: Our main result Theorem 3.4 is a precise statement of the compressed version included in the introduction. Our abstract accurately reflects this result's content and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes] .

   Justification: Limitations of our work are presented together with possible future work directions in Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes] .

Justification: Every new result stated as Theorem, Proposition or Lemma and is accompanied by its proof in the supplementary material. An additional theorem environment was created to present and cross-reference the assumptions in our results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulae, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes] .

Justification: In Section 4 we provide a detailed exposition of how our numerical experiments were programmed and the Python libraries used on them.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes] .

   Justification: The code used to produce the experiments in Figure 1 is available at an anonymized repository.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes] .

   Justification: A detailed description of all the hyperparameters is included in Section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes] .

   Justification: Our plot of the Gaussian process G_t in 1 includes a 95% confidence interval. As for the plot involving the distance $\mathcal{W}_2(G_t, f_t)$, a subsection discussing the approximation error on the computation of the empirical Wasserstein distance has been included.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Our experiments were run on the author's personal computer and do not require any particular hardware specifications.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes] .

   Justification: Our research follows NeurIPS code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA] .

    Justification: The innovations our paper introduces are of a theoretical nature.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA] .

    Justification: No sensitive data is used in our article. The data used in our experiments in Section 4 is artificially generated by our code and consists of points in $\mathbb{R}$.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes] .

    Justification: The python libraries used in our experiments are adequately referenced in Section 4.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: No new assets are introduced in this paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA] .

    Justification: Our paper does not involve human subjects in any way.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA] .

    Justification: Our paper does not involve human subjects in any way.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA] .

    Justification: Our results do not regard LLMs.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Gradient flow of the feature function

In this section we provide a closed analytical solution to (2.2) when $f = f^{\text{lin}}$.

Fix $x \in \mathbb{R}^{n_0}$ a test point. For the sake of clearness, we will use the following notation for each $t \geq 0$: $\overline{y}_t = f^{\text{lin}}(x; \overline{\theta}_t)$, $f_t^{\text{lin}} = f_t(\mathcal{X}; \overline{\theta}_t)$, $k_{\mathcal{X}\mathcal{X}} = k_0(\mathcal{X}, \mathcal{X})$ and $k_{x\mathcal{X}} = k_0(x, \mathcal{X})$. Note that $k_t = k_0$ for each t when $f = f^{\text{lin}}$.

We begin by stating a lemma that shows that our definition of I_t is consistent and commutes with the wide limit:

Lemma A.1. *For any real symmetric matrix $B \in \mathbb{R}^{n\times n}$ we have $I_t(B)B = BI_t(B) = \mathbb{1}_n - e^{-Bt}$; and for real symmetric matrix sequence $(B_n)_{n\in\mathbb{N}}$ with $B_n \to B$ we have $\lim_{n\to\infty} I_t(B_n) = I_t(B)$.*

The proof of this result follows from properties of the matrix exponential and is left to Supplementary Material E.

Consider the system of ODEs in (2.2) given by:

$$\frac{\partial}{\partial t} f_t^{\text{lin}} = -k_{\mathcal{X}\mathcal{X}}(f_t^{\text{lin}} - y), \tag{A.1}$$

$$\frac{\partial}{\partial t} \overline{y}_t = -k_{x\mathcal{X}}(f_t^{\text{lin}} - y). \tag{A.2}$$

Recall that, in general, the solution to the initial value problem $f'(t) = A(t)f(t),\ f(0) = f_0$ can be written as $f(t) = \exp(\int_0^t A(s)ds)f_0$, where $f(t)\colon \mathbb{R} \to \mathbb{R}^m,\ A(t)\colon \mathbb{R} \to \mathbb{R}^{m\times m}$ are integrable functions, and $t > 0$. Therefore in our case, by letting $u_t = f_t^{\text{lin}} - y$:

$$u_t = \exp\left(-\int_0^t k_{\mathcal{X}\mathcal{X}} ds\right) u_0 = \exp(-k_{\mathcal{X}\mathcal{X}} t) u_0. \tag{A.3}$$

Moreover, by letting $v_t = \overline{y}_t - y$ and substituting the solution for u_t, we obtain the following expression for v_t:

$$\frac{\partial}{\partial t} v_t = -k_{x\mathcal{X}} \exp(-k_{\mathcal{X}\mathcal{X}} t)(f_0 - y), \quad v_0 = y_0,$$

which, by integrating and by using Definition 2.2, becomes

$$v_t - v_0 = \overline{y}_t - \overline{y}_0 \tag{A.4}$$

$$= -k_{x\mathcal{X}} \int_0^t \exp\left(-k_{\mathcal{X}\mathcal{X}} s\right) ds (f_0 - y) \tag{A.5}$$

$$= -k_{x\mathcal{X}} I_t(k_{\mathcal{X}\mathcal{X}})(f_0 - y). \tag{A.6}$$

Note that formulae (A.3) and (A.6) agree with the formulae found by the authors of Lee et al. [2020].

When $k_{\mathcal{X}\mathcal{X}}$ is not degenerate, taking the limit when t tends to infinity, we get a prediction for the output of the linearized network at the end of the training:

$$f_\infty^{\text{lin}}(x) = \lim_{t\to\infty} \overline{y}_t = \overline{y}_0 - k_{x\mathcal{X}} k_{\mathcal{X}\mathcal{X}}^{-1}(f_0 - y).$$

A.1 Proof of the characterization of G_t

Here we prove the formulae in (2.5).

Define $B_t = -k_{x\mathcal{X}} I_t(k_{\mathcal{X}\mathcal{X}})$ and $C_t = k_{x\mathcal{X}} I_t(k_{\mathcal{X}\mathcal{X}})$, so that $\overline{y}_t - \overline{y}_0 = B_t f_0 + C_t y$. Also recall that $k_t = k_0$ for all $t \geq 0$ since f is linear on θ. Note that f_0 and $\overline{y}_0$ are centered Gaussian processes and hence $\mathbb{E}[\overline{y}_0 + B_t f_0] = 0$. Therefore, taking the expected value of the wide limit yields:

$$\mathbb{E}[\lim_{n_1\to\infty} \overline{y}_t] = \mathbb{E}[\lim_{n_1\to\infty} C_t y] = k_\infty(x, \mathcal{X}) I_t(k_\infty) y, \tag{A.7}$$

where $k_\infty = k_\infty(\mathcal{X}, \mathcal{X})$. This limit is well defined thanks to Lemma A.1.

Now let $x' \in \mathbb{R}^{n_0}$ and put $y'_t = f_t(x')$ and $B'_t = -k_{x'\mathcal{X}} I_t(k_{\mathcal{X}\mathcal{X}})$ for each $t \geq 0$. Then,

$$\begin{aligned}
\mathrm{Cov}(\lim_{n_1\to\infty} \overline{y}_t, \lim_{n_1\to\infty} y'_t) &= \mathbb{E}[\lim_{n_1\to\infty} (\overline{y}_t - \mathbb{E}[\overline{y}_t])(y'_t - \mathbb{E}[y'_t])] && \text{(A.8)}\\
&= \mathbb{E}[\lim_{n_1\to\infty} (\overline{y}_0 + B_t f_0)(y'_0 + B'_t f_0)] && \text{(A.9)}\\
&= \mathbb{E}[\lim_{n_1\to\infty} \overline{y}_0 y'_0] + \mathbb{E}[\lim_{n_1\to\infty} \overline{y}_0 f_0 B'_t] && \text{(A.10)}\\
&\quad + \mathbb{E}[\lim_{n_1\to\infty} y'_0 f_0 B_t] + \mathbb{E}[\lim_{n_1\to\infty} f_0^2 B_t B'_t] && \text{(A.11)}\\
&= \mathcal{K}(x,x') - \mathcal{K}(x,\mathcal{X}) I_t(k_\infty) k_\infty(\mathcal{X},x') && \text{(A.12)}\\
&\quad - k_\infty(x,\mathcal{X}) I_t(k_\infty) \mathcal{K}(\mathcal{X},x') && \text{(A.13)}\\
&\quad + k_\infty(x,\mathcal{X}) I_t(k_\infty) \mathcal{K}(\mathcal{X},\mathcal{X}) I_t(k_\infty) k_\infty(\mathcal{X},x'). && \text{(A.14)}
\end{aligned}$$

Again, Lemma A.1 ensures the limit exists.

B Auxiliary and related results

In this Supplementary Material we state intermediate results in the proof of our main theorem and recall some useful results. Throughout this section we will use the following notation for each $t \geq 0$: $y_t = f(x;\theta_t)$, $f_t = f(\mathcal{X};\theta_t)$, $k_t = k_t(\mathcal{X},\mathcal{X})$ and $k_\infty = k_\infty(\mathcal{X},\mathcal{X})$. All the proofs are deferred to Supplementary Material E.

In the next lemma we collect some well-known properties of the p-Wasserstein distance:

Lemma B.1. *Let $p \in [1,\infty[$ and let X, Y be random variables with values in $\mathbb{R}^n$ and Z be a random variable with values in $\mathbb{R}^m$. Let $\mathbb{P}_\xi$ denote the law of the random variable ξ for each $\xi \in \{X,Y,Z\}$. Then*

1. *If X, Y are defined on the same probability space, then $\mathcal{W}_p(X,Y) \leq \mathbb{E}[\|X-Y\|^p]^{\frac{1}{p}}$.*
2. *If Z is independent from X and Y then $\mathcal{W}_p(X+Z,Y+Z) \leq \mathcal{W}_p(X,Y)$.*
3. *Convexity of $\mathcal{W}_p^p$: $\mathcal{W}_p^p(X,Y) \leq \int_{\mathbb{R}^m} \mathcal{W}_p^p(\mathbb{P}_{X|Z=z}, \mathbb{P}_Y) d\mathbb{P}_Z(z)$.*
4. *Let $\lambda \in \mathbb{R}^m$ be a constant vector and consider the joint random variables $\tilde{X} = (X,Z), \tilde{Y} = (Y,\lambda)$. Then*
$$\mathcal{W}_p^p(\tilde{X},\tilde{Y}) \leq \mathcal{W}_p^p(X,Y) + \mathcal{W}_p^p(Z,\lambda) = \mathcal{W}_p^p(X,Y) + \left(\int_{\mathbb{R}^m} \|z-\lambda\|^p d\mathbb{P}(z)\right)^{\frac{1}{p}},$$
5. *Let V be a random variable with values in $\mathbb{R}^m$ and consider the joint random variables $\tilde{X} = (X,Z), \tilde{Y} = (Y,V)$. Then, for $p \geq 2$,*
$$\mathcal{W}_p^p(\tilde{X},\tilde{Y}) \leq 2^{\frac{p}{2}-1}\left(\mathcal{W}_{2p}^{2p}(X,Y) + \mathcal{W}_{2p}^{2p}(Z,V)\right).$$
Moreover, for $p = 1$,
$$\mathcal{W}_1(\tilde{X},\tilde{Y}) \leq \mathcal{W}_1(X,Y) + \mathcal{W}_1(Z,V).$$

The following result provides explicit formulae for the components of the Jacobian of f and the NTK:

Lemma B.2 (Gradients f and explicit formulae for $\tilde{k}$ and k)**.** *The following hold for each $x, x' \in \mathbb{R}^{n_0}$:*

$$\nabla_{\theta^{(0)}} f(x,\theta) = \frac{1}{\sqrt{n_1 n_0}} x^\top \Phi'\left(\frac{1}{\sqrt{n_0}} x\theta^{(0)}\right)\theta^{(1)} \in \mathbb{R}^{n_0}, \tag{B.1}$$

$$\nabla_{\theta^{(1)}} f(x,\theta) = \frac{1}{\sqrt{n_1}} \Phi\left(\frac{1}{\sqrt{n_0}} x\theta^{(0)}\right) \in \mathbb{R}^{n_1}. \tag{B.2}$$

Moreover, $\tilde{k}(x,x')$ is a diagonal $n_1 \times n_1$ matrix with

$$\tilde{k}_{ii}(x,x') = \frac{1}{n_0}\sum_{u=1}^{n_0} x_u x'_u, \tag{B.3}$$

and $k(x, x')$ is a real function given by:

$$k(x, x') = \frac{1}{n_1 n_0} \sum_{u=1}^{n_0} x_u x'_u \sum_{v=1}^{n_1} \Phi'(h_v(x))\Phi'(h_v(x'))(\theta_v^{(1)})^2 + \frac{1}{n_1} \sum_{v=1}^{n_1} \Phi(h_v(x))\Phi(h_v(x')). \quad \text{(B.4)}$$

B.1 Results at initialization

Throughout this subsection, we assume $t = 0$ and omit the subindex t unless needed. Our results 3.7 and 3.4 aim to generalize Theorem 4.1 in Trevisan [2023] in different directions. For reference, we reproduce the main result from Trevisan [2023] here in a simplified version:

Theorem B.3 (Trevisan)**.** *Then, for each $p \in \mathbb{N}$ there exists a constant c_p not depending on the network width n_1 such that:*

$$\mathcal{W}_p(f_0(\mathcal{X}), G_0(\mathcal{X})) \leq c_p \frac{1}{\sqrt{n_1}}. \quad \text{(B.5)}$$

Furthermore, if $\varphi \colon \mathbb{R} \to \mathbb{R}$ is a Lipschitz function, then

$$\mathcal{W}_p\left(\varphi((f_0(\mathcal{X})))^{\otimes 2}, \varphi(G_0(\mathcal{X}))^{\otimes 2}\right) \leq c_p \frac{(\text{Lip}\varphi + \varphi(0))^2}{\sqrt{n_1}}. \quad \text{(B.6)}$$

Theorem B.3 provides a quantitative bound for the Gaussian approximation of the neural network f_t. This can be upgraded to a bound for the joint distribution of the empirical kernel and the output of the neural network.

From now to the end of this subsection assume Φ and Φ' are bounded and $x, x' \in \mathbb{R}^{n_0}$ are fixed. Also, we adopt the notation introduced in Supplementary Material A. We state a helpful estimation:

Proposition B.4 (L^p bound for the kernel difference)**.** *Fix x, x' in $\mathbb{R}^{n_0}$ and let $\tilde{k}_{11} = \tilde{k}_{11}(x, x')$, $k = k(x, x')$, $\mathcal{K} = \mathcal{K}(x, x')$ and $k_\infty = k_\infty(x, x')$. There exists a constant $C > 0$ independent of n_1 such that:*

$$\mathbb{E}[|\tilde{k}_{11} - \mathcal{K}|^p] = 0, \quad \text{(B.7)}$$

$$\mathbb{E}[|k - k_\infty|^p] \leq \frac{C}{n_1^{\frac{p}{2}}}. \quad \text{(B.8)}$$

Remark B.5. Note that since k_∞ is deterministic, we have $\mathcal{W}_p^p(k, k_\infty) = \mathbb{E}[\|k - k_\infty\|^p]$. The constant C in B.4 depends on the constants produced by applications of Theorem B.3.

The proof of Proposition B.4 goes by triangle inequality combined with Theorem B.3, and by exploiting the independence between the entries of $\theta^{(1)}$ and those of $\theta^{(0)}$. An auxiliary result to prove B.4 is the following:

Proposition B.6 (L^p bounds for the empirical kernel)**.** *Let $\tilde{k}_{ij} = \tilde{k}_{ij}(x, x')$, for each $1 \leq i, j \leq n_1$, and $k = k(x, x')$. Then, the following inequalities hold:*

$$\mathbb{E}[|k|] = \|\Phi\|_\infty^2, \quad \text{(B.9)}$$

$$\mathbb{E}[|k|^p] \leq 2^{p-1}(2p-1)!!\|\Phi'\|_\infty^{2p}|\tilde{k}_{11}|^p + 2^{p-1}\|\Phi\|_\infty^{2p}. \quad \text{(B.10)}$$

Now we are ready to show the following result:

Proposition B.7 (Joint distribution Basteri-Trevisan)**.** *There exist a positive constant C such that:*

$$\mathcal{W}_p\left((k_0, f_0), (k_\infty, G_0)\right) \leq \frac{C}{\sqrt{n_1}}.$$

with C not depending on the width n_1.

The proof is by using Dudley's lemma (also referred to as the gluing lemma from optimal transport) as outlined in Villani [2008]. This lemma is used to decompose the Wasserstein distance into two terms. The summand regarding the network and its Gaussian approximation is bounded with Theorem B.3, and the other summand is bounded by Proposition B.4.

Lastly, the following lemma is used in the proof of Proposition 3.7.

Lemma B.8. *The following inequalities hold:*

$$\mathbb{E}[\|f_0 - y\|] \leq \sqrt{n}\|\Phi\|_\infty + \|y\|, \quad \text{(B.11)}$$

$$\mathbb{E}[\|f_0 - y\|^4] \leq 32n^2\|\Phi\|_\infty^4 + 8\|y\|^4. \quad \text{(B.12)}$$

B.2 Approximation by linearization at training time $t > 0$

In this subsection we state two results paramount to prove Proposition 3.6, along with all their auxiliary lemmas. The following theorem resembles Theorem 5.4 in Bartlett et al. [2021], or Theorem 2.2 in Chizat et al. [2019], but we would like to remark that those result differ from ours since the first applies to neural networks in which the training is restricted to $\theta^{(0)}$ while keeping $\theta^{(1)}$ frozen, and in both of them the loss function used for training is different than the one considered in our work. The proof is similar to the one in Bartlett et al. [2021] and is carried in full detail in Supplementary Material E.

In this result we prove quenched estimations for the dynamics of the parameters, the linearization error both in the parameters and in the network valued on test points, and the convergence of the network to the labels over the training set.

Assumption 5. The smallest eigenvalue of k_0 is bounded from below by:

$$4L(\mathcal{X})\|f_0 - y\| < \lambda_{\min}(k_0),$$

where $L(\mathcal{X})$ the Lipschitz constant of $\nabla_\theta f_0$ seen as a function of θ.

Theorem B.9. *Let $\lambda_{\min} = \lambda_{\min}(k_0), \sigma_{\min} = \sigma_{\min}(k_0)$ and $\sigma_{\max} = \sigma_{\max}(k_0)$ and let Assumption 5 hold. Then the following hold for $t > 0$ and for any test point $x \in \mathbb{R}^{n_0}$:*

$$\|\theta_t - \theta_0\| \leq \frac{2}{\sigma_{\min}}\|f_0 - y\|, \tag{B.13}$$

$$\|\theta_t - \overline{\theta}_t\| \leq \frac{(8 + 20\sigma_{\max}^2)L(\mathcal{X})}{\sigma_{\min}^3}\|f_0 - y\|^2, \tag{B.14}$$

$$\|f_t - y\|^2 \leq \|f_0 - y\|^2 \exp\left(-\frac{\lambda_{\min}}{2}t\right), \tag{B.15}$$

$$\|f(x;\theta_t) - f^{\text{lin}}(x;\overline{\theta}_t)\| \leq \frac{4L(x)}{\sigma_{\min}^2}\|y - f_0\|^2 \tag{B.16}$$

$$+ \frac{(8 + 20\sigma_{\max}^2)L(\mathcal{X})}{\sigma_{\min}^3}\|f_0 - y\|^2\|\nabla_\theta f_0(x)\|. \tag{B.17}$$

Proposition B.9 relies on a strong assumption involving the Lipschitz constant of the Jacobian of the network, the norm of the network at initialization and the positive-definiteness of the limiting kernel. The following rougher estimations depend on t, but they do not require Assumption 5 and will be key to prove our main theorem.

Theorem B.10. *Assume that Φ and Φ' are bounded. Let $L(\mathcal{X})$ be the Lipschitz constant of $\nabla_\theta f_0$, seen as a function of θ, and let $\psi(\theta_0) = \|f_0 - y\|$. Then, for each $t > 0$ there exist positive constants $A_1, \dots, A_5, B_1, \dots, B_9, C_1, \dots, C_8$ and C_9 not depending on n_1, n_0 nor t such that, $\mathbb{P}$-almost everywhere:*

$$\|y_t - f_t\|^2 \leq \frac{A_0}{n_0}\|\theta_0^{(0)}\theta_0^{(1)}\|^2 + \frac{A_1 t^2}{n_0 n_1}\|\theta_0^{(0)}\|^2\psi(\theta_0)^2 + \frac{A_2\|\theta_0^{(1)}\|^2 t^4}{n_1^2 n_0}\psi(\theta_0)^4 \tag{B.18}$$

$$+ \frac{A_3 t^6}{n_1^2 n_0}\psi(\theta_0)^6 + \frac{A_4 t^2}{n_1 n_0}\|\theta_0^{(1)}\|^4\psi(\theta_0)^2 + \frac{A_5 t^4\|\theta_0^{(1)}\|^2}{n_1^2 n_0}\psi(\theta_0)^4, \tag{B.19}$$

$$\|f^{\text{lin}} - \overline{y}_t\|^2 \leq \frac{B_0}{n_1 n_0}\|\theta_0^{(0)}\theta_0^{(1)}\|^2 + \frac{B_1}{n_1^2 n_0^2}\|\theta_0^{(0)}\theta_0^{(1)}\|^2\psi(\theta_0)^4 t^4 \tag{B.20}$$

$$+ \frac{B_2}{n_1^2 n_0^2}\|\theta_0^{(0)}\|^2\|\theta_0^{(1)}\|^4\psi(\theta_0)^2 t^2 + \frac{B_3}{n_1 n_0^2}\|\theta_0^{(0)}\theta_0^{(1)}\|^2\psi(\theta_0)^2 t^2 \tag{B.21}$$

$$+ \frac{B_4}{n_1^2 n_0}\|\theta_0^{(1)}\|^2\psi(\theta_0)^4 t^4 + \frac{B_5}{n_1^2 n_0}\|\theta_0^{(1)}\|^4\psi(\theta_0)^2 t^2 \tag{B.22}$$

$$+ \frac{B_6}{n_1 n_0}\|\theta_0^{(1)}\|^2\psi(\theta_0)^2 t^2 + \frac{B_7}{n_1^2 n_0}\|\theta_0^{(0)}\|^2\psi(\theta_0)^4 t^4 \tag{B.23}$$

$$+ \frac{B_8}{n_1^2 n_0}\|\theta_0^{(0)}\|^2\|\theta_0^{(1)}\|^2\psi(\theta_0)^2 t^2 + \frac{B_9}{n_1 n_0}\|\theta_0^{(0)}\|^2\psi(\theta_0)^2 t^2, \tag{B.24}$$

$$\|f_t - f^{\text{lin}}\|^2 \leq \frac{L(\mathcal{X})^2}{n_1^2}\left(\frac{C_1\psi(\theta_0)^4\|\theta_0^{(1)}\|^2t^2}{n_1n_0^2} + \frac{C_2\psi(\theta_0)^6t^8}{n_1^2n_0^2} + \frac{C_3\psi(\theta_0)^4t^6}{n_1n_0}\right. \tag{B.25}$$

$$+\frac{C_4\|\theta_0^{(1)}\|^4t^4}{n_0^2} + \frac{C_5\psi(\theta_0)^2\|\theta_0^{(1)}\|^2t^6}{n_1n_0^2} + \frac{C_6\|\theta_0^{(1)}\|^2t^6}{n_0} \tag{B.26}$$

$$\left.+\frac{C_7\psi(\theta_0)^2\|\theta_0^{(1)}\|^2t^4}{n_0} + \frac{C_8\psi(\theta_0)^4t^6}{n_1n_0} + C_9\psi(\theta_0)^2t^4\right), \tag{B.27}$$

where $\theta^{(0)}\theta^{(1)} \in \mathbb{R}^{n_0}$ denotes the usual product of the matrices $\theta^{(0)}$ and $\theta^{(1)}$.

Remark B.11. The dependence on time of the right-hand side of the above formulae comes from Lemma B.12. Indeed, by definition of the operator I_t, the sharpest upper bound for the matrices $I_t(k_t)$ and $I_t(k_0)$ when no lower bound for $\lambda_{\text{min}}(k_t)$ and $\lambda_{\text{min}}(k_0)$ is available is $\mathbb{1}_n t$.

The proof of the first two inequalities in this theorem is by exploiting an expression for the gradients of the network, contained in Lemma B.2; together with an integral result describing the behaviour of the parameters at time t with respect to the parameters at initialization. This is the content of Lemma B.12. The third inequality uses an integral argument together with the semipositive-definiteness of k_t to redirect the problem to studying $\|k_t - k_0\|$. All the constants in Theorem B.10 are multiples of the norms and Lipschitz constants of Φ and Φ', and of the norms of x and $\mathcal{X}$.

Now we state Lemma B.12 along with some concentration inequalities and related auxiliary results.

Lemma B.12 (Inequalities for $\theta_t^{(i)}$)**.** *Fix u and v with $1 \leq u \leq n_0$ and $1 \leq v \leq n_1$ and put $\mathcal{X}_u = ((x_i)_u)_{i=1}^n \in \mathbb{R}^n$. Let λ_{min} and λ_{min}^0 be the smallest eigenvalues of k_t and k_0, respectively. Then the following inequalities hold:*

$$(\theta_v^{(1)})_t \leq (\theta_v^{(1)})_0 + \frac{\|\Phi\|_\infty\|f_0 - y\|}{\sqrt{n_1}} I_t(\lambda_{\text{min}}), \tag{B.28}$$

$$(\theta_{uv}^{(0)})_t \leq (\theta_{uv}^{(0)})_0 + \frac{\|\Phi\|_\infty\|\Phi'\|_\infty\|f_0 - y\|^2\|\mathcal{X}_u\|}{2n_1\sqrt{n_0}} I_t(\lambda_{\text{min}})^2 \tag{B.29}$$

$$+ \frac{\|\Phi'\|_\infty\|f_0 - y\|\,\|\mathcal{X}_u\|}{\sqrt{n_1n_0}}(\theta_v^{(1)})_0 I_t(\lambda_{\text{min}}), \tag{B.30}$$

$$(\overline{\theta}_v^{(1)})_t \leq (\theta_v^{(1)})_0 + \frac{\|\Phi\|_\infty\|f_0 - y\|}{\sqrt{n_1}} I_t(\lambda_{\text{min}}^0), \tag{B.31}$$

$$(\overline{\theta}_{uv}^{(0)})_t \leq (\theta_{uv}^{(0)})_0 + \frac{\|\Phi\|_\infty\|\Phi'\|_\infty\|f_0 - y\|^2\|\mathcal{X}_u\|}{2n_1\sqrt{n_0}} I_t(\lambda_{\text{min}}^0)^2 \tag{B.32}$$

$$+ \frac{\|\Phi'\|_\infty\|f_0 - y\|\,\|\mathcal{X}_u\|}{\sqrt{n_1n_0}}(\theta_v^{(1)})_0 I_t(\lambda_{\text{min}}^0). \tag{B.33}$$

Remark B.13. Recall from the definition of I_t that for $t > 0$, $I_t(a) > 0$, even if $a = 0$. Moreover, $I_t(0) = t$ by definition. Hence $I_t(\lambda_{\text{min}}^0) \leq t$ and $I_t(\lambda_{\text{min}}^0)^2 \leq t^2$.

Recall the well known concentration inequality for χ^2-distributed random variables (see, for example, Laurent and Massart [2000]). For each $\gamma > 0$:

$$\mathbb{P}(\|\theta_0^{(1)}\|^2 \geq 2\gamma + 2\sqrt{\gamma n_1} + n_1) \leq \exp(-\gamma). \tag{B.34}$$

The following is a concentration inequality for the sup-norm of $\theta_0^{(1)}$; and as a consequence we get an estimation of the norm and Lipschitz constant of the Jacobian of f at initialization.

Lemma B.14. *For any $\gamma > 0$:*

$$\|\theta_0^{(1)}\|_\infty \leq \sqrt{r\gamma\log n_1}, \tag{B.35}$$

with probability bigger or equal than $1 - \frac{1}{n_1^{\frac{r\gamma}{2}-1}}$.

This concentration inequality is proven using the fact that $\|\theta_0^{(1)}\|_\infty$ is the supremum of n_1 Gaussian variables in absolute value.

Lemma B.15 (Norm and Lipschitz constant of the Jacobian at $t = 0$)**.** *Fix $r \geq 1$. Then for each $x \in \mathbb{R}^{n_0}$,*

$$\|\nabla_\theta f_0(x)\| \leq \frac{\|x\|\,\|\Phi'\|_\infty \sqrt{\gamma}}{\sqrt{n_0}} + \frac{\|\Phi\|_\infty}{\sqrt{n_1}}, \tag{B.36}$$

$$\mathrm{Lip}\nabla_\theta f_0(x) \leq \frac{\|x\|(\|\Phi'\|_\infty + \mathrm{Lip}\Phi)}{\sqrt{n_1 n_0}} + \frac{\|x\|^2 \mathrm{Lip}\Phi'}{\sqrt{n_1} n_0}\sqrt{r\gamma \log n_1}. \tag{B.37}$$

with probability greater or equal than $1 - \frac{1}{\frac{r\gamma}{n_1^2} - 1} - \exp(-\gamma n_1)$, where $f_0(x) \colon \mathbb{R}^N \to \mathbb{R}$ is understood as a function of θ.

Now we state a concentration inequality controlling the norm of the difference between the NTK and its limit.

Lemma B.16. *Let $k = k_0(\mathcal{X}, \mathcal{X})$ and $k_\infty = k_\infty(\mathcal{X}, \mathcal{X})$ and let $\gamma \in \mathbb{N}$. Put $\lambda_{\min} = \lambda_{\min}(k)$ and $\lambda_{\min}^\infty = \lambda_{\min}(k_\infty)$. Then, for each $p \in \mathbb{N}$,*

$$\|k - k_\infty\| \leq \frac{\gamma \lambda_{\min}^\infty}{2}, \tag{B.38}$$

with probability greater or equal than $1 - \left(\frac{2}{\gamma \lambda_{\min}^\infty}\right)^p \frac{C}{n_1^{\frac{p}{2}}}$, where C is a positive constant not depending on n_1.

Remark B.17. The previous lemma provides useful bounds for the smallest and largest eigenvalues of the empirical kernel at initialization, with arbitrarily high probability when the width diverges. Recall that the smallest eigenvalue of a matrix is a 1-Lipschitz function of the operator norm, which is bounded by the Frobenius norm:

$$|\lambda_{\min}(A) - \lambda_{\min}(B)| \leq \|A - B\|_{op} \leq \|A - B\|. \tag{B.39}$$

In particular, the previous Lemma implies, for $\gamma = 1$:

$$\lambda_{\min} \geq \lambda_{\min}^\infty - \|k - k_\infty\| \geq \frac{\lambda_{\min}^\infty}{2}. \tag{B.40}$$

Conversely, an upper bound for the largest eigenvalue of a matrix using the operator norm is given by:

$$\lambda_{\max}(A) \leq \lambda_{\max}(B) + \|A - B\|_{op} \leq \lambda_{\max}(B) + \|A - B\|. \tag{B.41}$$

Again, taking $\gamma = 1$ in the previous lemma yields:

$$\lambda_{\max} \leq \lambda_{\max}^\infty + \frac{\lambda_{\min}^\infty}{2}. \tag{B.42}$$

Both inequalities hold with probability greater or equal than $1 - \left(\frac{2}{\lambda_{\min}^\infty}\right)^p \frac{C}{n_1^{\frac{p}{2}}}$.

C Proof of Theorems 3.4 and 3.6

Here we prove Theorems 3.4 and 3.6, which share some auxiliary lemmas. Throughout this and the remaining appendices we will use the following notation for each $t \geq 0$: $y_t = f_t(x)$, $f_t = f_t(\mathcal{X})$, $\overline{y}_t = f^{\mathrm{lin}}(x; \overline{\theta}_t)$, $f_t^{\mathrm{lin}} = f^{\mathrm{lin}}(\mathcal{X}; \overline{\theta}_t)$ $k_t = k_t(\mathcal{X}, \mathcal{X})$ and $k_\infty = k_\infty(\mathcal{X}, \mathcal{X})$. The gradient ∇ and the expectation $\mathbb{E}$ will always be taken with respect to the parameters θ, unless otherwise indicated.

Let us begin by Proposition 3.6:

Proof of Proposition 3.6. Fix $p, r \in \mathbb{N}$. Consider the following subset of $\mathbb{R}^N$:

$$S = \{\theta \mid \frac{\|\theta\|^2}{n_1} \leq 5, \|\theta\|_\infty \leq \sqrt{r \log n_1}, \|k - k_\infty\| \leq \frac{\lambda_{\min}^\infty}{2}\}.$$

By the inequality (B.34) and Lemmas B.14 and B.16, the probability of S is bounded from below by $1-\exp(-n_1)-\frac{1}{\sqrt{n_1^{r-2}}}-\frac{\hat{c}}{(\lambda_{\min}^{\infty}\sqrt{n_1})^p}$, for a positive constant $\hat{c}$ not depending on n_1, n_0 nor t. Then,

$$\mathcal{W}_2^2(y_t,\overline{y}_t) \leq \inf_{\theta} \mathbb{E}_{\theta}[\|y_t-\overline{y}_t\|^2] \tag{C.1}$$

$$= \int_S \|y_t-\overline{y}_t\|^2 d\theta + \int_{S^C} \|y_t-\overline{y}_t\|^2 d\theta \tag{C.2}$$

$$\leq \sup_{\theta\in S}\|y_t-\overline{y}_t\|^2 + \left(\exp(-n_1)+\frac{1}{\sqrt{n_1^{r-2}}}+\frac{\hat{c}}{(\lambda_{\min}^{\infty}\sqrt{n_1})^p}\right)\sup_{\theta\in S^C}\|y_t-\overline{y}_t\|^2. \tag{C.3}$$

Let us begin by estimating the supremum over S. Note that by Lemma B.16, $\lambda_{\min}\geq\frac{\lambda_{\min}^{\infty}}{2}>0$ in S. The hypothesis for Proposition B.9 are satisfied when n_1, n_0 are large enough. In particular, thanks to Lemma B.15, Assumption 4

$$\frac{4\|\mathcal{X}\|(\sqrt{5}\|\Phi\|_{\infty}+\|y\|)}{\sqrt{n_1 n_0}}\left(\|\Phi'\|_{\infty}+\text{Lip}\Phi+\frac{\|\mathcal{X}\|\text{Lip}\Phi'\sqrt{r\log n_1}}{\sqrt{n_0}}\right)<\lambda_{\min}^{\infty}$$

implies Assumption 5 for any θ in S. Indeed, by Lemmas B.2 and B.15:

$$4L(\mathcal{X})\|f_0-y\| \tag{C.4}$$

$$\leq \frac{4\|\mathcal{X}\|}{\sqrt{n_1 n_0}}\left(\|\Phi'\|_{\infty}+\text{Lip}\Phi+\frac{\|\mathcal{X}\|\text{Lip}\Phi'\sqrt{r\log n_1}}{\sqrt{n_0}}\right)(\sqrt{5}\|\Phi\|_{\infty}+\|y\|). \tag{C.5}$$

Moreover, Lemma B.16 implies $\lambda_{\min}\geq\frac{\lambda_{\min}^{\infty}}{2}$ in S. These two inequalities together show that Assumption 4, which holds for sufficiently big n_1, is a sufficient condition for Proposition B.9 to hold.

On the other hand, by Lemmas B.16 and B.15, the following inequalities are satisfied in S, for $Z\in\{x,\mathcal{X}\}$:

$$L(Z)\leq\frac{\|Z\|}{\sqrt{n_1 n_0}}\left(\|\Phi'\|_{\infty}+\text{Lip}\Phi+\frac{\|\mathcal{X}\|\text{Lip}\Phi'\sqrt{r\log n_1}}{\sqrt{n_0}}\right), \tag{C.6}$$

$$\|\nabla_{\theta}f_0\|\leq\frac{\|\mathcal{X}\|\|\Phi'\|_{\infty}}{\sqrt{n_0}}+\frac{\|\Phi\|_{\infty}}{\sqrt{n_1}}, \tag{C.7}$$

$$\lambda_{\max}\leq\lambda_{\max}^{\infty}+\frac{\lambda_{\min}^{\infty}}{2}. \tag{C.8}$$

By substituting the estimations of $L(x), L(\mathcal{X}), \nabla f(x;\theta_0), \lambda_{\max}$ and $\lambda_{\min}$ above in Proposition B.9, for each $\theta\in S$:

$$\|y_t-\overline{y}_t\|\leq\frac{8(\sqrt{5}\|\Phi\|_{\infty}+\|y\|)^2}{\lambda_{\min}^{\infty}\sqrt{n_1 n_0}}\left(\|x\|+\frac{2\|\mathcal{X}\|}{\sqrt{\lambda_{\min}^{\infty}}}(\sqrt{2}+15\lambda_{\max}^{\infty})\right. \tag{C.9}$$

$$\left.\cdot\left(\|\Phi'\|_{\infty}+\text{Lip}\Phi+\frac{\|\mathcal{X}\|\text{Lip}\Phi'\sqrt{r\log n_1}}{\sqrt{n_0}}\right)\left(\frac{\|\mathcal{X}\|\|\Phi'\|_{\infty}}{\sqrt{n_0}}+\frac{\|\Phi\|_{\infty}}{\sqrt{n_1}}\right)\right) \tag{C.10}$$

$$\leq c\sqrt{\frac{r\log n_1}{n_1 n_0(\lambda_{\min}^{\infty})^3}}, \tag{C.11}$$

where the constant c is independent of r, t, n_0 and n_1 and can be determined explicitly:
$c=384\|\mathcal{X}\|(\sqrt{5}\|\Phi\|_{\infty}+\|y\|)^2\max\{\|x\|,\sqrt{2},15\lambda_{\max}^{\infty},\|\Phi'\|_{\infty},\text{Lip}\Phi,\|\mathcal{X}\|\text{Lip}\Phi',\|\mathcal{X}\|\|\Phi'\|_{\infty},\|\Phi\|_{\infty}\}$.

Hence,

$$\int_S\|y_t-\overline{y}_t\|^2 d\theta\leq\mathbb{P}(S)\sup_{\theta\in S}\|y_t-\overline{y}_t\|^2 \tag{C.12}$$

$$\leq\sup_{\theta\in S}\|y_t-\overline{y}_t\|^2 \tag{C.13}$$

$$\leq\frac{c^2 r\log n_1}{(\lambda_{\min}^{\infty})^3 n_1 n_0}. \tag{C.14}$$

Put $\alpha_1 = c^2$.

Now it only remains to estimate the second summand in (C.2). For each $\gamma \in \mathbb{N}$ let $\hat{\gamma} = 2\gamma + \sqrt{2\gamma} + 1$ and define the subsets:

$$\Omega_\gamma = \{\theta \mid \frac{\|\theta^{(1)}\|^2}{n_1} > \hat{\gamma}, \|\theta^{(1)}\|_\infty > \sqrt{r\gamma \log n_1}\}. \tag{C.15}$$

Also, define the subset

$$\Omega_* = \{\theta \mid \|k - k_\infty\| > \frac{\lambda_{\min}^\infty}{2}\}, \tag{C.16}$$

and let $\Omega = \Omega_* \setminus \bigcup_{\gamma\in\mathbb{N}} \Omega_\gamma$.

Intuitively, Ω_* and $\bigcup_{\gamma\in\mathbb{N}} \Omega_\gamma$ are the events in which the lower bound for the smallest eigenvalue of the empirical kernel and the upper bound of the Frobenius and sup-norm of the parameters at initialization, respectively, do not hold. Notice that $\mathbb{R}^N = S \sqcup \Omega \sqcup \bigcup_{\gamma\in\mathbb{N}} \Omega_\gamma$. We will use this partition of $\mathbb{R}^N$ to finish the proof.

By Lemma B.14 and by (B.34) we have $\mathbb{P}(\Omega_\gamma) \leq \exp(-\gamma n_1) + \frac{1}{\frac{r\gamma}{n_1^2}-1}$ and $\mathbb{P}(\Omega) \leq \mathbb{P}(\Omega_*) \leq \frac{\hat{c}}{(\lambda_{\min}^\infty \sqrt{n_1})^p}$. Moreover, the family $(\Omega_\gamma)_{\gamma\in\mathbb{N}}$ is a descending filtration of $S^C \setminus \Omega$. Let $D_\gamma = \Omega_\gamma \setminus \Omega_{\gamma+1}$, for each $\gamma \in \mathbb{N}$. This allows us to write:

$$\int_{S^C} \|y_t - \overline{y}_t\|^2 d\theta \leq \int_\Omega \|y_t - \overline{y}_t\|^2 d\theta + \sum_{\gamma\in\mathbb{N}} \int_{D_\gamma} \|y_t - \overline{y}_t\|^2 d\theta \tag{C.17}$$

$$\leq \left(\mathbb{P}(\Omega) + \sum_{\gamma\in\mathbb{N}} \mathbb{P}(D_\gamma)\right) \sup_{\theta\in D_\gamma} \|y_t - \overline{y}_t\|^2 \tag{C.18}$$

$$\leq 3\mathbb{P}(\Omega) \sup_{\theta\in\Omega} \left(\|y_t - f_t\|^2 + \|f_t - f_t^{\text{lin}}\|^2 + \|f_t^{\text{lin}} - \overline{y}_t\|^2\right) \tag{C.19}$$

$$+ 3\sum_{\gamma\in\mathbb{N}} \mathbb{P}(D_\gamma) \sup_{\theta\in D_\gamma} \left(\|y_t - f_t\|^2 + \|f_t - f_t^{\text{lin}}\|^2 + \|f_t^{\text{lin}} - \overline{y}_t\|^2\right) \tag{C.20}$$

$$\leq 3\sum_{\gamma\in\mathbb{N}} \exp(-\gamma n_1) \sup_{\theta\in D_\gamma} \|y_t - f_t\|^2 + 3\sum_{\gamma\in\mathbb{N}} \exp(-\gamma n_1) \sup_{\theta\in D_\gamma} \|f_t - f_t^{\text{lin}}\|^2 \tag{C.21}$$

$$+ 3\sum_{\gamma\in\mathbb{N}} \exp(-\gamma n_1) \sup_{\theta\in D_\gamma} \|f_t^{\text{lin}} - \overline{y}_t\|^2 + 3\sum_{\gamma\in\mathbb{N}} \frac{1}{\frac{r\gamma}{n_1^2}-1} \sup_{\theta\in D_\gamma} \|y_t - f_t\|^2 \tag{C.22}$$

$$+ 3\sum_{\gamma\in\mathbb{N}} \frac{1}{\frac{r\gamma}{n_1^2}-1} \sup_{\theta\in D_\gamma} \|f_t - f_t^{\text{lin}}\|^2 + 3\sum_{\gamma\in\mathbb{N}} \frac{1}{\frac{r\gamma}{n_1^2}-1} \sup_{\theta\in D_\gamma} \|f_t^{\text{lin}} - \overline{y}_t\|^2 \tag{C.23}$$

$$+ 3\frac{\hat{c}}{(\lambda_{\min}^\infty \sqrt{n_1})^p} \sup_{\theta\in\Omega} \|y_t - f_t\|^2 + 3\frac{\hat{c}}{(\lambda_{\min}^\infty \sqrt{n_1})^p} \sup_{\theta\in\Omega} \|f_t - f_t^{\text{lin}}\|^2 \tag{C.24}$$

$$+ 3\frac{\hat{c}}{(\lambda_{\min}^\infty \sqrt{n_1})^p} \sup_{\theta\in\Omega} \|f_t^{\text{lin}} - \overline{y}_t\|^2 \tag{C.25}$$

The 6 series in the previous expression are convergent. We compute an upper bound for each of the 6 series in (C.21) with the aid of Theorem B.10, Lemma B.14 and the inequality (B.34). Let $A = \max\{A_i\}, B = \max\{B_i\}$ and $C = \max\{C_i\}$ the maximums among the constants in Theorem B.10.

1. Let us estimate the first series. Observe that we can bound $\widehat{\gamma+1} \leq 7\gamma$. Moreover, since $A(\sqrt{\gamma+1} + \|y\|) \leq 2A\sqrt{\gamma}$ for γ large enough, up to adding to the constant A a multiple

of $\|y\|$ we can bound:

$$\sum_{\gamma\in\mathbb{N}} \exp(-\gamma n_1) \sup_{\theta\in D_\gamma} \|y_t - f_t\|^2 \tag{C.26}$$

$$\leq A \sum_{\gamma\in\mathbb{N}} \exp(-\gamma n_1) \left((\widehat{\gamma+1})^2 n_1^2 n_0 + \widehat{\gamma+1}(\sqrt{\gamma+1} + \|y\|)^2 t^2 \right. \tag{C.27}$$

$$+ \frac{\widehat{\gamma+1}(\sqrt{\gamma+1} + \|y\|)^4 t^4}{n_1 n_0} + \frac{(\sqrt{\gamma+1} + \|y\|)^6 t^6}{n_1^2 n_0} \tag{C.28}$$

$$\left. + \frac{(\widehat{\gamma+1})^2 (\sqrt{\gamma+1} + \|y\|)^2 n_1 t^2}{n_0} + \frac{\widehat{\gamma+1}(\sqrt{\gamma+1} + \|y\|)^4 t^4}{n_1 n_0} \right) \tag{C.29}$$

$$\leq 7A \sum_{\gamma\in\mathbb{N}} \exp(-\gamma n_1) \left(7\gamma^3 n_1^2 n_0 + 4\gamma^2 t^2 \right. \tag{C.30}$$

$$\left. + \frac{16\gamma^3 t^4}{n_1 n_0} + \frac{64\gamma^3 t^6}{7 n_1^2 n_0} + \frac{28\gamma^3 n_1 t^2}{n_0} + \frac{16\gamma^3 t^4}{n_1 n_0} \right). \tag{C.31}$$

Since the negative exponential function decraeases faster than any polynomial, for each $p \in \mathbb{N}$ there exists a positive constant N_p such that $x^p \leq \exp(-\frac{x}{2})$ for each $x \geq N_p$. Therefore, up to a multiplicative constant not depending on n_1, n_0, t nor γ:

$$\sum_{\gamma\in\mathbb{N}} \exp(-\gamma n_1) \sup_{\theta\in D_\gamma} \|y_t - f_t\|^2 \leq 7 \cdot 64 \cdot 6 A n_0 (1+t^6) \sum_{\gamma\in\mathbb{N}} \exp(-\frac{\gamma n_1}{2}) \tag{C.32}$$

$$\leq \frac{7 \cdot 64 \cdot 6 A n_0 (1+t^6) e^{-\frac{n_1}{2}}}{1 - e^{-\frac{n_1}{2}}}. \tag{C.33}$$

2. Now we estimate the second series. Using the bounds from the first series combined with Theorem B.10:

$$\sum_{\gamma\in\mathbb{N}} \exp(-\gamma n_1) \sup_{\theta\in D_\gamma} \|f_t^{\text{lin}} - \overline{y}_t\|^2 \tag{C.34}$$

$$\leq 7Bn_1 \sum_{\gamma\in\mathbb{N}} \exp(-\gamma n_1) \left(7\gamma^2 n_1 + \frac{112\gamma^4}{n_1} + \frac{196\gamma^4 n_1 t^2}{n_0} + \frac{28\gamma^3 n_1 t^2}{n_0} \right. \tag{C.35}$$

$$\left. + \frac{16\gamma^3 t^4}{n_1 n_0} + \frac{28\gamma^3 t^2}{n_0} + \frac{4\gamma^2 t^2}{n_0} + 16\gamma^3 t^4 n_0 + 28\gamma^3 t^2 + 4\gamma^2 t^2 \right) \tag{C.36}$$

$$\leq 7 \cdot 196 \cdot 9 B n_1 (1+t^4) \sum_{\gamma\in\mathbb{N}} \exp(-\frac{\gamma n_1}{2}) \tag{C.37}$$

$$= \frac{7 \cdot 196 \cdot 9 B n_1 (1+t^4) \exp(-\frac{n_1}{2})}{1 - \exp(-\frac{n_1}{2})}. \tag{C.38}$$

3. Now we estimate the third series in the same fashion as we did with the preceding series. For an upper bound of $L(\mathcal{X})$, recall Lemma B.15, which reads

$$L(\mathcal{X}) \leq \frac{d}{\sqrt{n_1 n_0}} \left(1 + \frac{\sqrt{r(\gamma+1)\log n_1}}{\sqrt{n_0}} \right),$$

for $\theta \in D_\gamma$ and d a constant not depending on n_1, n_0, t nor γ. For n_1 large enough, we can suppose $1 + \sqrt{\frac{r(\gamma+1)\log n_1}{n_0}} \leq 2\sqrt{\frac{r\gamma \log n_1}{n_0}}$. Then,

$$\sum_{\gamma\in\mathbb{N}} \exp(-\gamma n_1) \sup_{\theta\in D_\gamma} \|f_t - f_t^{\text{lin}}\|^2 \tag{C.39}$$

$$\leq \frac{4Cdr\log n_1}{n_1^3 n_0^2} \sum_{\gamma\in\mathbb{N}} \gamma \exp(-\gamma n_1) \left(\frac{112\gamma^3 t^2}{n_0^2} + \frac{64\gamma^3 t^8}{n_1^2 n_0^2} + \frac{16\gamma^2 t^6}{n_1 n_0} \right. \tag{C.40}$$

$$\left. + \frac{49\gamma^2 n_1^2 t^4}{n_0^2} + \frac{28\gamma^2 t^6}{n_0^2} + \frac{7\gamma n_1 t^6}{n_0} + \frac{28\gamma^2 n_1 t^4}{n_0} + \frac{16\gamma^2 t^6}{n_1 n_0} + 4\gamma t^4 \right) \tag{C.41}$$

$$\leq \frac{4\cdot 112\cdot 9Cdr\log n_1(1+t^8)}{n_1^3 n_0^2} \sum_{\gamma\in\mathbb{N}} \exp(-\frac{\gamma n_1}{2}) \tag{C.42}$$

$$= \frac{4\cdot 112\cdot 9Cdr(1+t^8)e^{-\frac{n_1}{2}}}{n_1^2 n_0^2(1-e^{-\frac{n_1}{2}})}. \tag{C.43}$$

4. Now we estimate the fourth series. Following the reasoning from the first series:

$$\sum_{\gamma\in\mathbb{N}} \frac{1}{n_1^{\frac{r\gamma}{2}-1}} \sup_{\theta\in D_\gamma} \|y_t - f_t\|^2 \leq 7A \sum_{\gamma\in\mathbb{N}} \frac{1}{n_1^{\frac{r\gamma}{2}-1}} \left(7\gamma^3 n_1^2 n_0 + 4\gamma^2 t^2 + \frac{16\gamma^3 t^4}{n_1 n_0} \right. \tag{C.44}$$

$$\left. + \frac{64\gamma^3 t^6}{7n_1^2 n_0} + \frac{28\gamma^3 n_1 t^2}{n_0} + \frac{16\gamma^3 t^4}{n_1 n_0} \right). \tag{C.45}$$

Recall that we can choose r large enough so that all the summands have n_1 on the denominator. In particular, it is enough to choose $r \geq 5$ in this case. Moreover, by reasoning like in the proof of the first series, up to a multiplicative constant the terms of the form $\gamma^p n_1^{-\frac{r\gamma}{2}}$ are bounded from above by $n_1^{-\frac{r\gamma}{4}}$. Therefore, up to a multiplicative constant not depending on n_1,n_0,t nor γ:

$$\sum_{\gamma\in\mathbb{N}} \exp(-\gamma n_1) \sup_{\theta\in D_\gamma} \|y_t - f_t\|^2 \leq 7\cdot 64\cdot 6An_0(1+t^6) \sum_{\gamma\in\mathbb{N}} \frac{1}{n_1^{\frac{r\gamma}{4}}} \tag{C.46}$$

$$= \frac{7\cdot 64\cdot 6An_0(1+t^6)n_1^{-\frac{r}{4}}}{1-n_1^{-\frac{r}{4}}}. \tag{C.47}$$

5. Now we estimate the fifth series. Using the bounds from the first series combined with Theorem B.10, up to a multiplicative constant:

$$\sum_{\gamma\in\mathbb{N}} \frac{1}{n_1^{\frac{r\gamma}{2}-1}} \sup_{\theta\in D_\gamma} \|f_t^{\text{lin}} - \overline{y}_t\|^2 \tag{C.48}$$

$$\leq 7Bn_1 \sum_{\gamma\in\mathbb{N}} \frac{1}{n_1^{\frac{r\gamma}{2}}} \left(7\gamma^2 n_1 + \frac{112\gamma^4 t^4}{n_1} + \frac{196\gamma^4 n_1 t^2}{n_0} + \frac{28\gamma^3 n_1 t^2}{n_0} \right. \tag{C.49}$$

$$\left. + \frac{16\gamma^3 t^4}{n_1 n_0} + \frac{28\gamma^3 t^2}{n_0} + \frac{4\gamma^2 t^2}{n_0} + 16\gamma^3 t^4 n_0 + 28\gamma^3 t^2 + 4\gamma^2 t^2 \right). \tag{C.50}$$

As we did for the fourth series, we can choose r large enough so that the previous series is convergent. $r \geq 5$ is sufficient. Up to a multiplicative constant:

$$\sum_{\gamma\in\mathbb{N}} \frac{1}{n_1^{\frac{r\gamma}{2}-1}} \sup_{\theta\in D_\gamma} \|y_t - f_t\|^2 \leq 196\cdot 7\cdot 9Bn_0(1+t^4) \sum_{\gamma\in\mathbb{N}} \frac{1}{n_1^{\frac{r\gamma}{4}}} \tag{C.51}$$

$$\leq \frac{196\cdot 7\cdot 9Bn_0(1+t^4)n_1^{-\frac{r}{4}}}{1-n_1^{-\frac{r}{4}}}. \tag{C.52}$$

6. Now we estimate the sixth and last series. Following the reasoning in the third and fourth series, up to a multiplicative constant:

$$\sum_{\gamma\in\mathbb{N}} \frac{1}{n_1^{\frac{r\gamma}{2}-1}} \sup_{\theta\in D_\gamma} \|f_t - f_t^{\text{lin}}\|^2 \tag{C.53}$$

$$\leq \frac{4Cdr\log n_1}{n_1^2 n_0^2} \sum_{\gamma\in\mathbb{N}} \frac{\gamma}{n_1^{\frac{r\gamma}{2}}} \left(\frac{112\gamma^3 t^2}{n_0^2} + \frac{64\gamma^3 t^8}{n_1^2 n_0^2} + \frac{16\gamma^2 t^7}{n_1 n_0} \right. \tag{C.54}$$

$$\left. + \frac{49\gamma^2 n_1^2 t^4}{n_0^2} + \frac{28\gamma^2 t^6}{n_0^2} + \frac{7\gamma n_1 t^6}{n_0} + \frac{28\gamma^2 n_1 t^4}{n_0} + \frac{16\gamma^2 t^6}{n_1 n_0} + 4\gamma t^4 \right). \tag{C.55}$$

In this case, any $r \geq 3$ makes the series converge.

$$\sum_{\gamma\in\mathbb{N}} \frac{1}{n_1^{\frac{r\gamma}{2}-1}} \sup_{\theta\in D_\gamma} \|f_t - f_t^{\text{lin}}\|^2 \tag{C.56}$$

$$\leq \frac{4\cdot 112\cdot 9Cdr(1+t^8)}{n_1 n_0^2} \sum_{\gamma\in\mathbb{N}} \frac{\gamma}{n_1^{\frac{r\gamma}{4}}} \tag{C.57}$$

$$= \frac{4\cdot 112\cdot 9Cdr(1+t^8) n_1^{-\frac{r}{4}}}{n_1 n_0^2 (1-n_1^{-\frac{r}{4}})}. \tag{C.58}$$

Now we finish the proof by estimating the last 3 terms in (C.21). Recall that, by definition of Ω, the bounds for the Frobenius and the sup-norms of the parameters at initialization used in the estimation of $\int_S \|y_t - \overline{y}_t\| d\theta$ hold also for $\theta \in \Omega$; and recall that $\mathbb{P}(\Omega) \leq \frac{\hat{c}}{(\lambda_{\min}^\infty \sqrt{n_1})^p}$. Let $R_1 = \frac{\hat{c}}{(\lambda_{\min}^\infty)^p}$. Then it suffices to choose p large enough to counterattack the biggest exponent of n_1 in each of the bounds given by Theorem B.10. In particular, as seen while choosing r when computing the six series, it is enough to take $p = r \geq 5$. Then, up to multiplicative constants,

$$\frac{R_1}{n_1^{\frac{p}{2}}} \sup_{\theta\in\Omega} \|y_t - f_t\|^2 \leq AR_1 n_0 (1+t^6) \frac{1}{\sqrt{n_1}}, \tag{C.59}$$

$$\frac{R_1}{n_1^{\frac{p}{2}}} \sup_{\theta\in\Omega} \|f_t\|^2 \leq R_1 B n_0 (1+t^4) \frac{1}{\sqrt{n_1}}, \tag{C.60}$$

$$\frac{R_1}{n_1^{\frac{p}{2}}} \sup_{\theta\in\Omega} \|y_t - f_t\|^2 \leq CdrR_1 (1+t^8) \frac{\log n_1}{n_1\sqrt{n_1} n_0^2} \tag{C.61}$$

$$\leq CdrR_1 (1+t^8) \frac{1}{\sqrt{n_1} n_0^2}. \tag{C.62}$$

Grouping the estimations for the nine summands in (C.21) and taking $\alpha_2 = R_1 \max\{A, B, Cd\}$, up to a multiplicative constant,

$$\int_{S^C} \|y_t - \overline{y}_t\|^2 d\theta \leq \frac{\alpha_2 r n_0}{(\lambda_{\min}^\infty)^r n_1^{\frac{r}{4}}} (1+t^8). \tag{C.63}$$

This concludes the proof. □

Then, our main result, Theorem 3.4, is a direct application of Propositions 3.6 and 3.7:

Proof of Theorem 3.4. By triangle inequality and the elementary inequality $(a+b)^2 \leq 2a^2 + 2b^2$ for $a, b \geq 0$ decompose:

$$\mathcal{W}_2^2(y_t, G_t) \leq 2\mathcal{W}_2^2(y_t, \overline{y}_t) + 2\mathcal{W}_2^2(\overline{y}_t, G_t). \tag{C.64}$$

Then the thesis follows estimating the first summand with Proposition 3.6 and the second one with Proposition 3.7. For large enough n_1 we can take the constants a_1 and a_2 to be a multiple of α_1 and α_2 in Proposition 3.6; since the right-hand side in the statement in that result decreases as $\frac{\log n_1}{n_1} + \frac{1}{n_1^{\frac{r}{4}}}$, which is strictly slower than the right-hand side of the statement in Proposition 3.7 for any $n_1 \geq 2$. □

D Proof of Proposition 3.7

Proof of Proposition 3.7. Fix $x \in \mathbb{R}^{n_0}$ a test point. Let $G_t^x = G_t(x)$ and $G_t^{\mathcal{X}} = G_t(\mathcal{X})$.

First we show the result for the training set. With the aid of Equation (2.2), we derive a closed ODE for $|f_t - G_t|^2$. By Cauchy-Schwarz's inquality and Young's inequality

$$\frac{1}{2}\frac{\partial}{\partial t}\|f_t - G_t^{\mathcal{X}}\|^2 = \langle k_{\mathcal{X}\mathcal{X}}(y - f_t) - k_\infty(y - G_t^{\mathcal{X}}), (f_t - G_t^{\mathcal{X}})\rangle \tag{D.1}$$

$$= (k_{\mathcal{X}\mathcal{X}} - k_\infty)(y - f_t)(f_t - G_t^{\mathcal{X}}) - k_\infty\|f_t - G_t^{\mathcal{X}}\|^2 \tag{D.2}$$

$$\leq \|k_{\mathcal{X}\mathcal{X}} - k_\infty\|\|y - f_t\|\|f_t - G_t^{\mathcal{X}}\| - \lambda_{\min}^\infty\|f_t - G_t^{\mathcal{X}}\|^2 \tag{D.3}$$

$$\leq \frac{1}{2\varepsilon}\|k_{\mathcal{X}\mathcal{X}} - k_\infty\|^2\|y - f_t\|^2 + \frac{\varepsilon}{2}\|f_t - G_t^{\mathcal{X}}\|^2 - \lambda_{\min}^\infty\|f_t - G_t^{\mathcal{X}}\|^2. \tag{D.4}$$

Note that, by gradient flow equation we have $\|f_t - y\| \leq e^{-\frac{\lambda_{\min}^\infty}{2}t}\|f_0 - y\|$ for each $t \geq 0$. Choosing $\varepsilon = \lambda_{\min}^\infty$ and putting $b_t = \frac{e^{-\lambda_{\min}^\infty t}}{\lambda_{\min}^\infty}\|k_{\mathcal{X}\mathcal{X}} - k_\infty\|^2\|y - f_0\|^2$ yields:

$$\frac{\partial}{\partial t}\|f_t - G_t^{\mathcal{X}}\|^2 \leq b_t - \lambda_{\min}^\infty\|f_t - G_t^{\mathcal{X}}\|^2. \tag{D.5}$$

Grönwall's inequality applied to (D.5) implies:

$$\|f_t - G_t\|^2 \leq e^{-\lambda_{\min}^\infty t}\left(\|f_0 - G_0\|^2 + \int_0^t e^{\lambda_{\min}^\infty s} b_s ds\right) \tag{D.6}$$

$$= e^{-\lambda_{\min}^\infty t}\left(\frac{\|k_{\mathcal{X}\mathcal{X}} - k_\infty\|^2\|f_0 - y\|^2 t}{\lambda_{\min}^\infty} + \|f_0 - G_0\|^2\right). \tag{D.7}$$

Recall the definition of 2-Wasserstein distance. Taking the expected value of the previous equation and taking the infimum on all the couplings between f_t and G_t we can bound by Hölder's inequality:

$$\mathcal{W}_2^2(f_t, G_t) \leq \mathbb{E}[\|f_t - G_t\|^2] \tag{D.8}$$

$$\leq e^{-\lambda_{\min}^\infty t}\left(\frac{t}{\lambda_{\min}^\infty}\mathbb{E}[\|k_{\mathcal{X}\mathcal{X}} - k_\infty\|^2\|f_0 - y\|^2] + \mathbb{E}[\|f_0 - G_0\|^2]\right) \tag{D.9}$$

$$\leq e^{-\lambda_{\min}^\infty t}\left(\frac{t}{\lambda_{\min}^\infty}\mathbb{E}[\|k_{\mathcal{X}\mathcal{X}} - k_\infty\|^4]^{\frac{1}{2}}\mathbb{E}[\|f_0 - y\|^4]^{\frac{1}{2}} + \mathbb{E}[\|f_0 - G_0\|^2]\right). \tag{D.10}$$

Now we can take the infimum on the couplings between f_0 and G_0 to apply Theorem B.3, Proposition B.4 and Lemma B.8 to estimate the right-hand side of (D.10). There are positive constants c_1 and c_2 not depending on n_1 such that:

$$\mathcal{W}_2^2(f_t, G_t) \leq e^{-\lambda_{\min}^\infty t}\left(\frac{c_1 t}{\lambda_{\min}^\infty n_1}\sqrt{32n^2\|\Phi\|_\infty^4 + 8\|y\|^4} + \frac{c_2}{n_1}\right) \tag{D.11}$$

$$\leq \frac{Ce^{-\lambda_{\min}^\infty t}}{n_1}(t+1), \tag{D.12}$$

with $C = \max\{\frac{2c_1}{\lambda_{\min}^\infty}\sqrt{8n^2\|\Phi\|_\infty^4 + 2\|y\|^4}, c_2\}$.

Now we show the result for an arbitrary test point $x \in \mathbb{R}^{n_0}$. Let $k_\infty^x = k_\infty(x, \mathcal{X})$. We can bound, by Equation (2.2),

$$\frac{\partial}{\partial t}(y_t - G_t^x)^2 = 2(k_{x\mathcal{X}}(y - f_t) - k_\infty^x(y - G_t^{\mathcal{X}}))(y_t - G_t^x).$$

By the formula for the derivative of the product and Cauchy-Schwarz's inequality, the preceding equation implies:

$$\frac{\partial}{\partial t}(y_t - G_t^x) \leq k_{x\mathcal{X}}(y - f_t) - k_\infty^x(y - G_t^{\mathcal{X}}) = \|k_{x\mathcal{X}} - k_\infty^x\|\|y - f_t\| - k_\infty^x(f_t - G_t^{\mathcal{X}}). \tag{D.13}$$

Put $\overline{\lambda} = \min_{1\leq i\leq n} k_\infty(x, x_i)$. The last summand can be further estimated with the first result in this theorem. There exists a positive constants C not depending on n_1 such that:

$$\frac{\partial}{\partial t}(y_t - G_t^x) \leq \|k_{x\mathcal{X}} - k_\infty^x\|\|y - f_t\| + \frac{\overline{\lambda}Ce^{-\frac{\lambda_{\min}^\infty}{2}t}}{\sqrt{n_1}}\sqrt{t+1}. \tag{D.14}$$

Again, we use $\|f_t - y\| \leq e^{-\frac{\lambda_{\min}^\infty}{2}t}\|f_0 - y\|$. Integrating we obtain:

$$(y_t - G_t^x) \leq (y_0 - G_0^x) + \|k_{x\mathcal{X}} - k_\infty^x\|\|y - f_0\|(1 - e^{-\lambda_{\min}^\infty t}) + \frac{\overline{\lambda}C}{\sqrt{n_1}}\int_0^t \sqrt{s+1}e^{-\frac{\lambda_{\min}^\infty}{2}s}ds \tag{D.15}$$

$$\leq (y_0 - G_0^x) + \|k_{x\mathcal{X}} - k_\infty^x\|\|y - f_0\|(1 - e^{-\lambda_{\min}^\infty t}) + \frac{\overline{\lambda}CD}{\sqrt{n_1}}(2 - \sqrt{t+1}e^{-\frac{\lambda_{\min}^\infty t}{2}}), \tag{D.16}$$

for a positive constant D, which explicit computation we now show separately. First we compute an antiderivative of $\sqrt{t+1}e^{-\frac{\lambda_{\min}^\infty}{2}t}$:

$$\int \sqrt{s+1}e^{-\frac{\lambda_{\min}^\infty}{2}s}ds \tag{D.17}$$

$$= 2e^{\frac{\lambda_{\min}^\infty}{2}}\int u^2 e^{-\frac{\lambda_{\min}^\infty}{2}u^2}du \tag{D.18}$$

$$= 2e^{\frac{\lambda_{\min}^\infty}{2}}\left(-\frac{ue^{-\frac{\lambda_{\min}^\infty}{2}u^2}}{\lambda_{\min}^\infty} + \int \frac{ue^{-\frac{\lambda_{\min}^\infty}{2}u^2}}{\lambda_{\min}^\infty}du\right) + C \tag{D.19}$$

$$= 2e^{\frac{\lambda_{\min}^\infty}{2}}\left(-\frac{ue^{-\frac{\lambda_{\min}^\infty}{2}u^2}}{\lambda_{\min}^\infty} + \frac{\sqrt{\pi}}{\sqrt{2(\lambda_{\min}^\infty)^3}}\int \frac{2e^{-v^2}}{\sqrt{\pi}}dv\right) + C \tag{D.20}$$

$$= 2e^{\frac{\lambda_{\min}^\infty}{2}}\left(-\frac{ue^{-\frac{\lambda_{\min}^\infty}{2}u^2}}{\lambda_{\min}^\infty} + \frac{\sqrt{\pi}\operatorname{erf} v}{\sqrt{2(\lambda_{\min}^\infty)^3}}\right) + C \tag{D.21}$$

$$= 2e^{\frac{\lambda_{\min}^\infty}{2}}\left(-\frac{\sqrt{s+1}e^{-\frac{\lambda_{\min}^\infty}{2}(s+1)}}{\lambda_{\min}^\infty} + \frac{\sqrt{\pi}\operatorname{erf}\left(\frac{\sqrt{\lambda_{\min}^\infty(s+1)}}{\sqrt{2}}\right)}{\sqrt{2(\lambda_{\min}^\infty)^3}}\right) + C \tag{D.22}$$

$$= -\frac{2\sqrt{s+1}e^{-\frac{\lambda_{\min}^\infty s}{2}}}{\lambda_{\min}^\infty} + \frac{\sqrt{2\pi}e^{\frac{\lambda_{\min}^\infty}{2}}\operatorname{erf}\left(\frac{\sqrt{\lambda_{\min}^\infty(s+1)}}{\sqrt{2}}\right)}{(\lambda_{\min}^\infty)^{\frac{3}{2}}} + C. \tag{D.23}$$

where in the first step we substituted $u = \sqrt{s+1}$, in the third step we substituted $v = \sqrt{\frac{\lambda_{\min}^\infty}{2}}u$ and in the fourth step we used the definition of Gauss error function $\operatorname{erf}(x)$; and C denotes the integration constant. Therefore,

$$\int_0^t \sqrt{s+1}e^{-\frac{\lambda_{\min}^\infty}{2}s}ds = \frac{2 - 2\sqrt{t+1}e^{-\frac{\lambda_{\min}^\infty t}{2}}}{\lambda_{\min}^\infty} \tag{D.24}$$

$$+ \frac{\sqrt{2\pi}e^{\frac{\lambda_{\min}^\infty}{2}}\left(\operatorname{erf}\left(\frac{\sqrt{\lambda_{\min}^\infty(t+1)}}{\sqrt{2}}\right) - \operatorname{erf}\left(\frac{\sqrt{\lambda_{\min}^\infty}}{\sqrt{2}}\right)\right)}{(\lambda_{\min}^\infty)^{\frac{3}{2}}}. \tag{D.25}$$

Since $\operatorname{erf}(t) \in]-1, 1[$, we can bound:

$$\int_0^t \sqrt{s+1}e^{-\frac{\lambda_{\min}^\infty}{2}s}ds \leq \frac{2}{\lambda_{\min}^\infty}\left(1 - \sqrt{t+1}e^{-\frac{\lambda_{\min}^\infty t}{2}} + \sqrt{\frac{2\pi}{\lambda_{\min}^\infty}}e^{-\frac{\lambda_{\min}^\infty}{2}}\right) \tag{D.26}$$

$$\leq D(2 - \sqrt{t+1}e^{-\frac{\lambda_{\min}^\infty t}{2}}), \tag{D.27}$$

with $D = \frac{2}{\lambda^{\infty}_{\min}} \max\{1, \sqrt{\frac{2\pi}{\lambda^{\infty}_{\min}}} e^{-\frac{\lambda^{\infty}_{\min}}{2}}\}$.

Turning back to (D.16), we can finish by applying the elementary inequality $(a+b+c)^2 \leq 3a^2+3b^2+3c^2$ for $a, b, c \geq 1$ and Hölder's inequality. After that, Theorem B.3, Proposition B.4 and Lemma B.8 can be applied in the same fashion as in the proof of the training case, yielding positive constants d_1, d_2 and d_3 such that:

$$\mathcal{W}_2^2(y_t, G_t^x) = \mathbb{E}[|y_t - G_t^x|^2] \tag{D.28}$$

$$\leq 3\mathbb{E}[|y_0 - G_0^x|^2] + 3\mathbb{E}[\|k_{x\mathcal{X}} - k^x_{\infty}\|^4]^{\frac{1}{2}} \mathbb{E}[\|y - f_0\|^4]^{\frac{1}{2}} (1 - e^{-\lambda^{\infty}_{\min} t})^2 \tag{D.29}$$

$$+ \frac{3(\overline{\lambda} C D)^2}{n_1} (2 - \sqrt{t+1} e^{-\frac{\lambda^{\infty}_{\min} t}{2}})^2 \tag{D.30}$$

$$\leq \frac{3d_1}{n_1} + \frac{6d_2}{n_1} \sqrt{8n^2 \|\Phi\|^4_{\infty} + 2\|y\|^4} (1 + e^{-2\lambda^{\infty}_{\min} t}) \tag{D.31}$$

$$+ \frac{3(\overline{\lambda} C D)^2}{n_1} (4 + (t+1) e^{-\lambda^{\infty}_{\min} t}). \tag{D.32}$$

By putting $\overline{C} = \max\{3d_1, 12(\overline{\lambda} C D)^2\}$ and $\overline{D} = \max\{6d_2 \sqrt{8n^2 \|\Phi\|^4_{\infty} + 2\|y\|^4}, 3(\overline{\lambda} C D)^2\}$ we obtain the thesis.

Note that $C, \overline{C}$ and $\overline{D}$ do not depend neither on n_1 nor t. □

E Proofs of the auxiliary results

We present here all the remaining proofs. For clearness, we will use the following notation: $X_v = h(x)_v$ and $X'_v = h(x')_v$ for each $1 \leq v \leq n_1$, for any $x, x' \in \mathbb{R}^{n_0}$.

Proof of Lemma A.1. It is trivial when B is nonsingular. Let $\lambda_1, \ldots, \lambda_n$ be the (possible repeated) ordered eigenvalues of B and suppose $\lambda_j = 0$. Then, by using the eigenvalue decomposition of B and elementary properties of the matrix exponential:

$$I_t(B)B = U \begin{pmatrix} \frac{1-e^{-\lambda_1 t}}{\lambda_1} & & & & \\ & \ddots & & & \\ & & t & & \\ & & & \ddots & \\ & & & & \frac{1-e^{-\lambda_n t}}{\lambda_n} \end{pmatrix} U^\top U \begin{pmatrix} \lambda_1 & & & & \\ & \ddots & & & \\ & & 0 & & \\ & & & \ddots & \\ & & & & \lambda_n \end{pmatrix} U^\top \tag{E.1}$$

$$= U \begin{pmatrix} 1-e^{-\lambda_1 t} & & & & \\ & \ddots & & & \\ & & 0 & & \\ & & & \ddots & \\ & & & & 1-e^{-\lambda_n t} \end{pmatrix} U^\top \tag{E.2}$$

$$= UU^\top - U \begin{pmatrix} e^{-\lambda_1 t} & & & & \\ & \ddots & & & \\ & & 0 & & \\ & & & \ddots & \\ & & & & e^{-\lambda_n t} \end{pmatrix} U^\top \tag{E.3}$$

$$= \mathbb{1}_n - e^{-Bt}. \tag{E.4}$$

The converse equality is proven in an analogous way.

As for the limit property, it is enough to show it for real numbers. Let $(a_n)_{n\in\mathbb{N}}$ be a real sequence converging to $a \in \mathbb{R}$. If $a \neq 0$, of if $a = a_n = 0$ for each n the result is trivial. Thus, assume $a = 0$

and, up to taking a subsequence, that $a_n \neq 0$. Then

$$\lim_{n\to\infty} I_t(a_n) = \lim_{n\to\infty} \frac{1-e^{-a_n t}}{a_n} = t = I_t(0).$$

□

Proof of Lemma B.1. For the proof of the first three points we refer to any monograph on the Wasserstein distance such as Villani [2008]. In order to show (4), let π^{XY} be an optimal transport plan between X and Y, let π^{XZ} be the law of $\tilde{X}$ and let $\mu^X = \mathbb{P}_X$ be the marginal law of X. Applying the gluing lemma of optimal transport produces a probability measure π on $\mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n$ given by

$$\pi(x,z,y) = \frac{\pi^{XY}(x,y)}{\mu^X(x)}\mu^X(x)\frac{\pi^{XZ}(x,z)}{\mu^X(x)} = \frac{\pi^{XY}(x,y)\pi^{XZ}(x,z)}{\mu^X(x)}.$$

Note that integrating with respect to y yields π^{XZ} and integrating with respect to z yields π^{XY}. Note also that there is a unique coupling between Y and λ, which is given by $\mathbb{P}_Y \times \delta_\lambda$, hence,

$$\begin{aligned}
\mathcal{W}_p^p\left(\tilde{X},\tilde{Y}\right) &\leq \int_{\mathbb{R}^{n+m+n}} \int_{\mathbb{R}^m} \|\tilde{X}(x,z) - \tilde{Y}(y,w)\|^p \delta_\lambda(dw) d\pi(dx,dz,dy) && \text{(E.5)}\\
&= \int_{\mathbb{R}^{n+m+n}} \|\tilde{X}(x,z) - \tilde{Y}(y,\lambda)\|^p d\pi(dx,dz,dy) && \text{(E.6)}\\
&\leq \int_{\mathbb{R}^{n+n}} \|X-Y\|^p d\pi^{XY}(dx,dy) + \int_{\mathbb{R}^m} \|Z-\lambda\|^p d\mu^Z(dz) && \text{(E.7)}\\
&= \mathcal{W}_p^p(X,Y) + \mathcal{W}_p^p(Z,\lambda), && \text{(E.8)}
\end{aligned}$$

where μ^Z is the marginal probability on Z.

On the other hand, to show (5) fix $\mu \in \mathcal{P}(\mathbb{R}^n)$ and $\nu \in \mathcal{P}(\mathbb{R}^m)$ two probability measures, and denote $\mu \times \nu$ the product measure on $\mathbb{R}^{n+m}$ with the tensor product σ-algebra. Then

$$\begin{aligned}
\mathcal{W}_p^p(\tilde{X},\tilde{Y}) &\leq \mathbb{E}_{\mu\times\nu}[\|\tilde{X}-\tilde{Y}\|^p] && \text{(E.9)}\\
&\leq \mathbb{E}_{\mu\times\nu}[(\|X-Y\|^2 + \|Z-V\|^2)^{\frac{p}{2}}] && \text{(E.10)}\\
&\leq 2^{\frac{p}{2}-1}\left(\mathbb{E}_\mu[\|X-Y\|^{2p}] + \mathbb{E}_\nu[\|Z-V\|^{2p}]\right), && \text{(E.11)}
\end{aligned}$$

where in the last step we applied the elementary inequality $(a+b)^p \leq 2^{p-1}(a^p+b^p)$, for $a,b \geq 0$ and $p \geq 1$. For the 1-Wasserstein distance instead, we apply the elementary inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$:

$$\begin{aligned}
\mathcal{W}_1(\tilde{X},\tilde{Y}) &\leq \mathbb{E}_{\mu\times\nu}[\|\tilde{X}-\tilde{Y}\|] && \text{(E.12)}\\
&\leq \mathbb{E}_{\mu\times\nu}[\left(\|X-Y\|^2 + \|Z-V\|^2\right)^{\frac{1}{2}}] && \text{(E.13)}\\
&\leq \mathbb{E}_\mu[\|X-Y\|] + \mathbb{E}_\nu[\|Z-V\|]. && \text{(E.14)}
\end{aligned}$$

Lastly, taking the infimum over $(\mu,\nu) \in \mathcal{P}(\mathbb{R}^n) \times \mathcal{P}(\mathbb{R}^m)$ finishes the proof. □

Proof of Lemma B.2. The computation of the gradients is by chain rule and definition of f. The claims about the kernels follow from taking the dot product on the gradients we just calculated, for each $1 \leq i,j \leq n_1$:

$$\begin{aligned}
\tilde{k}_{ij}(x,x') &= \left(\nabla_{\theta^{(0)}} X_i\right)\left(\nabla_{\theta^{(0)}} X'_j\right) && \text{(E.15)}\\
&= \frac{1}{n_0} \sum_{\substack{u=1,\dots,n_0\\ v=1,\dots,n_1}} \frac{\partial}{\partial\theta_{uv}^{(0)}}(x\theta_{_i}^{(0)}) \frac{\partial}{\partial\theta_{uv}^{(0)}}(x'\theta_{_j}^{(0)}) && \text{(E.16)}\\
&= \frac{1}{n_0} \sum_{\substack{u=1,\dots,n_0\\ v=1,\dots,n_1}} x_u \delta_{iv} x'_u \delta_{jv} && \text{(E.17)}\\
&= \frac{1}{n_0} \sum_{u=1}^{n_0} x_u x'_u \delta_{ij}. && \text{(E.18)}
\end{aligned}$$

Lastly,

$$k(x,x') = \left(\nabla_{\theta^{(0)}} f^{(2)}(x)\right)\left(\nabla_{\theta^{(0)}} f^{(2)}(x')\right) + \left(\nabla_{\theta^{(1)}} f^{(2)}(x)\right)\left(\nabla_{\theta^{(1)}} f^{(2)}(x')\right) \tag{E.19}$$

$$= \frac{1}{n_1}\sum_{\substack{u=1,\dots,n_0\\ v=1,\dots,n_1}} \frac{\partial}{\partial\theta_{uv}^{(0)}}(\Phi(\frac{1}{\sqrt{n_0}}x\theta^{(0)})\theta^{(1)})\frac{\partial}{\partial\theta_{uv}^{(0)}}(\Phi(\frac{1}{\sqrt{n_0}}x'\theta^{(0)})\theta^{(1)}) \tag{E.20}$$

$$+ \frac{1}{n_1}\sum_{z=1}^{n_1} \frac{\partial}{\partial\theta_{z}^{(1)}}(\Phi(\frac{1}{\sqrt{n_0}}x\theta^{(0)})\theta^{(1)})\frac{\partial}{\partial\theta_{z}^{(1)}}(\Phi(\frac{1}{\sqrt{n_0}}x'\theta^{(0)})\theta^{(1)}) \tag{E.21}$$

$$= \frac{1}{n_1 n_0}\sum_{\substack{u=1,\dots,n_0\\ v=1,\dots,n_1}} x_u x'_u \Phi'(X_v)\Phi'(X_v)(\theta_v^{(1)})^2 + \frac{1}{n_1}\sum_{z=1}^{n_1}\Phi(X_z)\Phi(X'_z). \tag{E.22}$$

□

E.1 Proof of results at initialization

In this subsection we prove the useful Proposition B.4, which generalises the second part of Theorem B.3. This step is paramount for the proof of the rest of our results. From now to the end of this subsection assume Φ and Φ' are bounded and $x, x' \in \mathbb{R}^{n_0}$ are fixed.

Proof of Proposition B.6. Note that, from Lemma B.2, k can be written as:

$$k = \frac{1}{n_1}\tilde{k}_{11}\sum_{v=1}^{n_1}\Phi'(X_v)\Phi'(X'_v)(\theta_v^{(1)})^2 + \frac{1}{n_1}\sum_{v=1}^{n_1}\Phi(X_v)\Phi(X'_v),$$

and recall that $\tilde{k}$ is deterministic, for any fixed inputs x, x'. Hence, by boundedness of Φ and Φ', and independence of $\theta^{(1)}$ and $\theta^{(0)}$,

$$\mathbb{E}[|k|] \le (\|\Phi'\|_\infty^2 \mathbb{E}[|\tilde{k}_{11}|] + \|\Phi\|_\infty^2) = \|\Phi\|_\infty^2, \tag{E.23}$$

$$\mathbb{E}[|k|^p] = \mathbb{E}[|\frac{1}{n_1}\tilde{k}_{11}\sum_{v=1}^{n_1}\Phi'(X_v)\Phi'(X'_v)(\theta_v^{(1)})^2 + \frac{1}{n_1}\sum_{v=1}^{n_1}\Phi(X_v)\Phi(X'_v)|^p] \tag{E.24}$$

$$\le 2^{p-1}\|\Phi'\|_\infty^{2p}\mathbb{E}[|\tilde{k}_{11}|^p]\mathbb{E}[(\theta_1^{(1)})^{2p}] + 2^{p-1}\|\Phi\|_\infty^{2p} \tag{E.25}$$

$$\le 2^{p-1}(2p-1)!!\|\Phi'\|_\infty^{2p}|\tilde{k}_{11}|^p + 2^{p-1}\|\Phi\|_\infty^{2p}, \tag{E.26}$$

where in the last inequality we used that the $2p$-th moment of a standard Gaussian variable is equal to $(2p-1)!!$.

□

Proof of Proposition B.4. We will use the notation instroduced in Proposition B.6. The first claim is trivial since $\tilde{k}$ coincides with $\mathcal{K}$. To show the second claim, we split:

$$\mathbb{E}[|k_{11} - k_\infty|^p] = \mathbb{E}[|\frac{1}{n_1}\tilde{k}_{11}\sum_{v=1}^{n_1}\Phi'(X_v)\Phi'(X'_v)(\theta_v^{(1)})^2 + \frac{1}{n_1}\sum_{v=1}^{n_1}\Phi(X_v)\Phi(X'_v) \tag{E.27}$$

$$- \mathcal{K}\mathbb{E}_G[\Phi'(G(x))\Phi'(G(x'))] - \mathbb{E}_G[\Phi(G(x))\Phi(G(x'))]|^p] \tag{E.28}$$

$$\le 2^{p-1}\mathbb{E}[|\frac{1}{n_1}\tilde{k}_{11}\sum_{v=1}^{n_1}\Phi'(X_v)\Phi'(X'_v)(\theta_v^{(1)})^2 - \mathcal{K}\mathbb{E}_G[\Phi'(G(x))\Phi'(G(x'))]|^p] \tag{E.29}$$

$$+ 2^{p-1}\mathbb{E}[|\frac{1}{n_1}\sum_{v=1}^{n_1}\Phi(X_v)\Phi(X'_v) - \mathbb{E}_G[\Phi(G(x))\Phi(G(x'))]|^p]. \tag{E.30}$$

To bound the first summand in (E.29) we split again by adding and substracting an auxiliary term:

$$\mathbb{E}[|\frac{1}{n_1}\tilde{k}_{11}\sum_{v=1}^{n_1}\Phi'(X_v)\Phi'(X'_v)(\theta_v^{(1)})^2 - \mathcal{K}\mathbb{E}_G[\Phi'(G(x))\Phi'(G(x'))]|^p] \tag{E.31}$$

$$\leq 2^{p-1}\mathbb{E}[|\frac{1}{n_1}\tilde{k}_{11}\sum_{v=1}^{n_1}\Phi'(X_v)\Phi'(X'_v)(\theta_v^{(1)})^2 - \frac{1}{n_1}\mathcal{K}\sum_{v=1}^{n_1}\Phi'(X_v)\Phi'(X'_v)(\theta_v^{(1)})^2|^p] \tag{E.32}$$

$$+ 2^{p-1}\mathbb{E}[|\frac{1}{n_1}\mathcal{K}\sum_{v=1}^{n_1}\Phi'(X_v)\Phi'(X'_v)(\theta_v^{(1)})^2 - \mathcal{K}\mathbb{E}_G[\Phi'(G(x))\Phi'(G(x'))]|^p] \tag{E.33}$$

$$= 2^{p-1}\mathbb{E}[|\frac{1}{n_1}(\tilde{k}_{11}-\mathcal{K})\sum_{v=1}^{n_1}\Phi'(X_v)\Phi'(X'_v)(\theta_v^{(1)})^2|^p] \tag{E.34}$$

$$+ 2^{p-1}(\mathcal{K})^p\mathbb{E}[|\frac{1}{n_1}\sum_{v=1}^{n_1}\Phi'(X_v)\Phi'(X'_v)(\theta_v^{(1)})^2 - \mathbb{E}_G[\Phi'(G(x))\Phi'(G(x'))]|^p]. \tag{E.35}$$

The first summand in (E.34) vanishes since $\tilde{k}$ equals $\mathcal{K}$. We estimate the second summand in (E.34), once again by adding and substracting an auxiliary term:

$$\mathbb{E}[|\frac{1}{n_1}\sum_{v=1}^{n_1}\Phi'(X_v)\Phi'(X'_v)(\theta_v^{(1)})^2 - \mathbb{E}_G[\Phi'(G(x))\Phi'(G(x'))]|^p] \tag{E.36}$$

$$\leq 2^{p-1}\mathbb{E}[|\frac{1}{n_1}\sum_{v=1}^{n_1}\Phi'(X_v)\Phi'(X'_v)((\theta_v^{(1)})^2-1)|^p] \tag{E.37}$$

$$+ 2^{p-1}\mathbb{E}[|\frac{1}{n_1}\sum_{v=1}^{n_1}\Phi'(X_v)\Phi'(X'_v) - \mathbb{E}_G[\Phi'(G(x))\Phi'(G(x'))]|^p]. \tag{E.38}$$

The first summand in (E.37) vanishes, by boundedness of Φ' and independence of the parameters $\theta_v^{(1)}$:

$$\mathbb{E}[|\frac{1}{n_1}\sum_{v=1}^{n_1}\Phi'(X_v)\Phi'(X'_v)((\theta_v^{(1)})^2-1)|^p] \leq \frac{1}{n_1^p}\|\Phi'\|_\infty^{2p}\mathbb{E}[|\sum_{v=1}^{n_1}(\theta_v^{(1)})^2-1|^p] \tag{E.39}$$

$$= \frac{1}{n_1^p}\|\Phi'\|_\infty^{2p}\sum_{\alpha_1,\ldots,\alpha_p=1}^{n_1}\prod_{j=1}^{p}\mathbb{E}[(\theta_{\alpha_j}^{(1)})^2-1] \tag{E.40}$$

$$= 0. \tag{E.41}$$

As for the second summand in (E.37), by Theorem B.3 there exists a constant c_1 not depending on n_1 such that:

$$\mathcal{W}_p^p(\frac{1}{n_1}\sum_{v=1}^{n_1}\Phi'(X_v)\Phi'(X'_v), \mathbb{E}_G[\Phi'(G(x))\Phi'(G(x'))]) \tag{E.42}$$

$$= \mathbb{E}[|\frac{1}{n_1}\sum_{v=1}^{n_1}\Phi'(X_v)\Phi'(X'_v) - \mathbb{E}_G[\Phi'(G(x))\Phi'(G(x'))]|^p] \tag{E.43}$$

$$\leq c_1\left(\frac{\mathrm{Lip}\Phi' + \Phi'(0)}{\sqrt{n_1}}\right)^p. \tag{E.44}$$

It remains only to bound the second summand in (E.29). This is done by using again Theorem B.3. There exists a constant c_2 not depending on n_1 such that:

$$\mathcal{W}_p^p(\frac{1}{n_1}\sum_{v=1}^{n_1}\Phi(X_v)\Phi(X'_v), \mathbb{E}_G[\Phi(G(x))\Phi(G(x'))]) \tag{E.45}$$

$$= \mathbb{E}[\frac{1}{n_1}\sum_{v=1}^{n_1}\Phi(X_v)\Phi(X'_v) - \mathbb{E}_G[\Phi(G(x))\Phi(G(x'))]|^p] \tag{E.46}$$

$$\leq c_2\left(\frac{\text{Lip}\Phi + \Phi(0)}{\sqrt{n_1}}\right)^p. \tag{E.47}$$

Putting together all the preceding estimations we obtain:

$$\mathbb{E}[|k_{11} - k_\infty|^p] \leq C\frac{1}{n_1^{\frac{p}{2}}}, \tag{E.48}$$

with $C = 2^{p-1}\max\{2^{2p-2}c_1(\text{Lip}\Phi' + \Phi'(0)), c_2(\text{Lip}\Phi + \Phi(0))\}$. □

These results suffice to prove Proposition B.7.

Proof of Proposition B.7. Consider the joint random variables $\tilde{X} = (k_{\mathcal{X}\mathcal{X}}, \hat{f}(\mathcal{X}))$ and $\tilde{Y} = (k_\infty(\mathcal{X}, \mathcal{X}), G(\mathcal{X}))$. Then Lemma B.1.4 together with Proposition B.4 and Theorem B.3 yield

$$\mathcal{W}_p(\tilde{X}, \tilde{Y}) \leq \mathcal{W}_p(k_{\mathcal{X}\mathcal{X}}, k_\infty(\mathcal{X}, \mathcal{X})) + \mathcal{W}_p(f(\mathcal{X}), G(\mathcal{X})) \leq \frac{C + D}{\sqrt{n_1}}, \tag{E.49}$$

where C is the constant in Proposition B.4 and D the one in Theorem B.3. Both constants do not depend on n_1. □

Lastly, we prove Lemma B.8.

Proof of Lemma B.8. Let $\theta_{ij}^{(0)}$ denote the ij-th entry of $\theta_0^{(0)} \in \mathbb{R}^{n_0 \times n_1}$, and let $\theta_j^{(1)}$ denote the j-th component of $\theta_0^{(1)} \in \mathbb{R}^{n_1}$. By Jensen's inequality and independence of the parameters and $x_1, \dots x_n$:

$$\mathbb{E}[\|f_0\|] \leq \sqrt{\mathbb{E}[\|\frac{1}{\sqrt{n_1}}\Phi(\mathcal{X}\theta_0^{(0)})\theta_0^{(1)}\|^2]} \tag{E.50}$$

$$\leq \sqrt{n}\sqrt{\mathbb{E}[|\Phi(x_1\theta_{\_1}^{(0)})|^2]\mathbb{E}[|\theta_1^{(1)}|^2]} \tag{E.51}$$

$$\leq \sqrt{n}\|\Phi\|_\infty. \tag{E.52}$$

As for the fourth moment,

$$\mathbb{E}[\|f_0\|^4] = \mathbb{E}\left[\left(\sum_{i=1}^{n}\frac{1}{n_1}\left(\sum_{j=1}^{n_1}\Phi(x_i\theta_{\_j}^{(0)})\theta_j^{(1)}\right)^2\right)^2\right] \tag{E.53}$$

$$\leq \frac{n^2}{n_1^2}\mathbb{E}\left[\left(\sum_{j=1}^{n_1}\Phi(x_1\theta_{\_j}^{(0)})\theta_j^{(1)}\right)^4\right] \tag{E.54}$$

$$\leq \frac{\|\Phi\|_\infty^4 n^2}{n_1^2}(3n_1 + n_1^2) \tag{E.55}$$

$$\leq 4n^2\|\Phi\|_\infty^4. \tag{E.56}$$

Triangular inequality finishes the proof. □

E.2 Approximation of the network by linearization

In this subsection we prove the results involved in the proof of Proposition 3.6.

With a slight abuse of notation, we will denote by $\|x - \mathcal{X}\|$ the positive quantity $\sup_{1\leq i\leq n} \|x - x_i\|$. Also, given any matrix $A = (a_{ij})_{\substack{1\leq i\leq n\\ 1\leq j\leq m}}$ we will denote by $\frac{\partial}{\partial A} f$ the matrix $\nabla_A f = (\frac{\partial}{\partial A_{ij}})_{\substack{1\leq i\leq n\\ 1\leq j\leq m}}$.

We will consider the *linearized gradient flow*, given by

$$\frac{\partial}{\partial t}\overline{\theta}_t = -\nabla_\theta f_0(f^{\text{lin}}(\mathcal{X};\overline{\theta}_t) - y).$$

For this subsection introduce the following notations: $f_t^{\text{lin}} = f^{\text{lin}}(\mathcal{X};\overline{\theta}_t)$ and $\overline{y}_t = f^{\text{lin}}(x;\overline{\theta}_t)$.

Proof of Lemma B.12. Let λ_{min} be the smallest eigenvalue of k_t. By gradient flow equations for the parameters θ_t and $\overline{\theta}_t$:

$$\|f_t - y\| \leq e^{-\lambda_{\text{min}} t}\|f_0 - y\|, \tag{E.57}$$

$$\|f_t^{\text{lin}} - y\| \leq e^{-\lambda_{\text{min}}^0 t}\|f_0 - y\|. \tag{E.58}$$

On the other hand, Lemma B.2 combined with Cauchy-Schwarz's inequality and the gradient flow equations produces the following system of differential inequalities:

$$\frac{\partial}{\partial t}(\theta_v^{(1)})_t \leq \frac{1}{\sqrt{n_1}}\|\Phi\|_\infty\|f_0 - y\|e^{-\lambda_{\text{min}} t}, \tag{E.59}$$

$$\frac{\partial}{\partial t}(\theta_{uv}^{(0)})_t \leq \frac{1}{\sqrt{n_1 n_0}}\|\Phi'\|_\infty\|f_0 - y\|\,\|\mathcal{X}_u\|(\theta_v^{(1)})_t e^{-\lambda_{\text{min}} t}. \tag{E.60}$$

The previous is a triangular system of differential inequalities of the form

$$\begin{cases} \frac{\partial}{\partial t}(\theta_v^{(1)})_t & \leq B_1 e^{-\lambda_{\text{min}} t} \\ \frac{\partial}{\partial t}(\theta_{uv}^{(0)})_t & \leq B_0(\theta_v^{(1)})_t e^{-\lambda_{\text{min}} t}, \end{cases}$$

with $B_1 = \frac{1}{\sqrt{n_1}}\|\Phi\|_\infty\|f_0 - y\|$ and $B_0 = \frac{1}{\sqrt{n_1 n_0}}\|\Phi'\|_\infty\|f_0 - y\|\,\|\mathcal{X}_u\|$.

By integration on $[0, t]$ and substitution we get:,

$$(\theta_v^{(1)})_t \leq (\theta_v^{(1)})_0 + B_1 I_t(\lambda_{\text{min}}) \tag{E.61}$$

$$\leq (\theta_v^{(1)})_0 + \frac{\|\Phi\|_\infty\|f_0 - y\|}{\sqrt{n_1}} I_t(\lambda_{\text{min}}), \tag{E.62}$$

$$(\theta_{uv}^{(0)})_t \leq (\theta_{uv}^{(0)})_0 + B_0 B_1 \int_0^t I_s(\lambda_{\text{min}})ds + B_0\|\theta_0^{(1)}\| I_t(\lambda_{\text{min}}) \tag{E.63}$$

$$\leq (\theta_{uv}^{(0)})_0 + \frac{\|\Phi\|_\infty\|\Phi'\|_\infty\|f_0 - y\|^2\|\mathcal{X}_u\|}{2n_1\sqrt{n_0}} I_t(\lambda_{\text{min}})^2 \tag{E.64}$$

$$+ \frac{\|\Phi'\|_\infty\|f_0 - y\|\,\|\mathcal{X}_u\|}{\sqrt{n_1 n_0}} I_t(\lambda_{\text{min}})(\theta_v^{(1)})_0. \tag{E.65}$$

Note that in the last inequality, we used $\int_0^t I_s(b)ds \leq \frac{I_t(b)^2}{2}$, for any $b \geq 0$.

Thanks to (E.57), the linearised parameters $\overline{\theta}_t$ also satisfy the preceding inequalities, and hence the thesis holds. □

Proof of Lemma B.14. Let Φ denote the CDF of a standard Gaussian variable. For each $a > 0$, since the entries of $\theta_0^{(1)}$ are n_1 i.i.d. standard Gaussian variables,

$$\mathbb{P}(\|\theta_0^{(1)}\|_\infty \leq a) = (1 - 2(1 - \Phi(a)))^{n_1}. \tag{E.66}$$

Bernouilli's inequality and standard estimations for Gaussian tails yield

$$\mathbb{P}(\|\theta_0^{(1)}\|_\infty \leq a) \geq 1 - 2n_1(1-\Phi(a)) \tag{E.67}$$

$$\geq 1 - n_1 \exp\left(-\frac{a^2}{2}\right). \tag{E.68}$$

Let $r \geq 1$ and put $a = \sqrt{r\gamma \log n_1}$. Then:

$$\mathbb{P}(\|\theta_0^{(1)}\|_\infty \leq a) \geq 1 - n_1 \exp\left(-\frac{r\gamma \log n_1}{2}\right) \tag{E.69}$$

$$= 1 - n_1 \exp\left(\log n_1^{-\frac{r\gamma}{2}}\right) \tag{E.70}$$

$$= 1 - \frac{1}{n_1^{\frac{r\gamma}{2}-1}}. \tag{E.71}$$

□

Proof of Lemma B.15. We will write f for short of $f_0(x)(\theta)$. By Lemma B.2 and Cauchy-Schwarz's inequality,

$$\|\nabla_\theta f\|^2 = \|\frac{\partial}{\partial \theta^{(0)}} f\|^2 + \|\frac{\partial}{\partial \theta^{(1)}} f\|^2 \tag{E.72}$$

$$\leq \frac{1}{n_0 n_1}\|x\|^2 \|\Phi'\|_\infty^2 \|\theta^{(1)}\|^2 + \frac{1}{n_1}\|\Phi\|_\infty^2. \tag{E.73}$$

Then the first claim follows by (B.34) and the elementary inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, for $a, b \geq 0$.

Now we prove the second inequality. Let $\theta, \tilde{\theta} \in \mathbb{R}^N$, then,

$$\|\nabla_\theta f(\theta) - \nabla_\theta f(\tilde{\theta})\|^2 = \|\frac{\partial}{\partial \theta^{(0)}} f(\theta) - \frac{\partial}{\partial \theta^{(0)}} f(\tilde{\theta})\|^2 + \|\frac{\partial}{\partial \theta^{(1)}} f(\theta) - \frac{\partial}{\partial \theta^{(1)}} f(\tilde{\theta})\|^2. \tag{E.74}$$

Let us estimate the first summand in the previous expression. By Lemma B.2, for each $1 \leq u \leq n_0, 1 \leq v \leq n_1$,

$$|\frac{\partial}{\partial \theta_{uv}^{(0)}} f(\theta) - \frac{\partial}{\partial \theta_{uv}^{(0)}} f(\tilde{\theta})| \tag{E.75}$$

$$\leq \frac{1}{\sqrt{n_1 n_0}} x_u (\Phi'(\frac{1}{\sqrt{n_0}} \sum_{j=1}^{n_0} x_j \theta_{jv}^{(0)}) \theta_v^{(1)} - \Phi'(\frac{1}{\sqrt{n_0}} \sum_{j=1}^{n_0} x_j \tilde{\theta}_{jv}^{(0)}) \tilde{\theta}_v^{(1)}) \tag{E.76}$$

$$\leq \frac{x_u}{\sqrt{n_1 n_0}} \Phi'(\frac{1}{\sqrt{n_0}} \sum_{j=1}^{n_0} x_j \theta_{jv}^{(0)})(\theta_v^{(1)} - \tilde{\theta}_v^{(1)}) \tag{E.77}$$

$$+ \frac{x_u \tilde{\theta}_v^{(1)}}{\sqrt{n_1 n_0}} x_u (\Phi'(\frac{1}{\sqrt{n_0}} \sum_{j=1}^{n_0} x_j \theta_{jv}^{(0)}) - \Phi'(\frac{1}{\sqrt{n_0}} \sum_{j=1}^{n_0} x_j \tilde{\theta}_{jv}^{(0)})) \tag{E.78}$$

$$\leq \frac{x_u \|\Phi'\|_\infty (\theta_v^{(1)} - \tilde{\theta}_v^{(1)})}{\sqrt{n_1 n_0}} + \frac{x_u \text{Lip}\Phi' \tilde{\theta}_v^{(1)}}{\sqrt{n_1} n_0} \sum_{j=1}^{n_0} x_j (\theta_{jv}^{(0)} - \tilde{\theta}_{jv}^{(0)}). \tag{E.79}$$

Hence,

$$\|\frac{\partial}{\partial \theta^{(0)}} f(\theta) - \frac{\partial}{\partial \theta^{(0)}} f(\tilde{\theta})\|^2 \tag{E.80}$$

$$= \sum_{\substack{u=1,\dots,n_0 \\ v=1,\dots,n_1}} |\frac{\partial}{\partial \theta_{uv}^{(0)}} f(\theta) - \frac{\partial}{\partial \theta_{uv}^{(0)}} f(\tilde{\theta})|^2 \tag{E.81}$$

$$\leq \frac{\|x\|^2 \|\Phi'\|_\infty^2 \|\theta^{(1)} - \tilde{\theta}^{(1)}\|^2}{n_1 n_0} + \frac{\|x\|^4 (\text{Lip}\Phi')^2 \|\tilde{\theta}^{(1)}\|_\infty^2 \|\theta^{(0)} - \tilde{\theta}^{(0)}\|^2}{n_1 n_0^2} \tag{E.82}$$

$$\leq \frac{\|x\|^2 \|\Phi'\|_\infty^2 \|\theta^{(1)} - \tilde{\theta}^{(1)}\|^2}{n_1 n_0} + \frac{\|x\|^4 (\text{Lip}\Phi')^2 \|\theta^{(0)} - \tilde{\theta}^{(0)}\|^2}{n_1 n_0^2} r\gamma \log n_1, \tag{E.83}$$

with probability greater or equal than $1 - \frac{1}{n^{\frac{r\gamma}{2}-1}}$, where in the last step we used Lemma B.14.

Similarly, the second summand can be estimated as follows. First compute the partial derivatives by using Lemma B.2, for each $1 \leq v \leq n_1$:

$$|\frac{\partial}{\partial\theta_v^{(1)}} f(\theta) - \frac{\partial}{\partial\theta_v^{(1)}} f(\tilde{\theta})| \tag{E.84}$$

$$\leq \frac{1}{\sqrt{n_1}}(\Phi(\frac{1}{\sqrt{n_0}}\sum_{j=1}^{n_0} x_j\theta_{jv}^{(0)}) - \Phi(\frac{1}{\sqrt{n_0}}\sum_{j=1}^{n_0} x_j\tilde{\theta}_{jv}^{(0)})) \tag{E.85}$$

$$\leq \frac{\text{Lip}\Phi}{\sqrt{n_1 n_0}}\sum_{j=1}^{n_0} x_j(\theta_{jv}^{(0)} - \tilde{\theta}_{jv}^{(0)}). \tag{E.86}$$

Therefore,

$$\|\frac{\partial}{\partial\theta^{(1)}} f(\theta) - \frac{\partial}{\partial\theta^{(1)}} f(\tilde{\theta})\|^2 \tag{E.87}$$

$$= \sum_{v=1,\dots,n_1} |\frac{\partial}{\partial\theta_v^{(1)}} f(\theta) - \frac{\partial}{\partial\theta_v^{(1)}} f(\tilde{\theta})|^2 \tag{E.88}$$

$$\leq \frac{(\text{Lip}\Phi)^2\|x\|^2}{n_1 n_0}\|\theta^{(0)} - \tilde{\theta}^{(0)}\|^2. \tag{E.89}$$

The preceding estimations, together with $\frac{\|\theta^{(i)}-\tilde{\theta}^{(i)}\|^2}{\|\theta-\tilde{\theta}\|^2} \leq 1$ for $i = 0, 1$, yield:

$$\frac{\|\nabla_\theta f(\theta) - \nabla_\theta f(\tilde{\theta})\|^2}{\|\theta - \tilde{\theta}\|^2} \tag{E.90}$$

$$\leq \frac{\|x\|^2\|\Phi'\|_\infty^2}{n_1 n_0} + \frac{\|x\|^4(\text{Lip}\Phi')^2}{n_1 n_0^2} r\gamma \log n_1 + \frac{(\text{Lip}\Phi)^2\|x\|^2}{n_1 n_0}. \tag{E.91}$$

Taking the square root in the last inequality yields the thisis. □

Proof of Lemma B.16. Fix $\gamma \in \mathbb{N}$. The probability of $Z = \|k - k_\infty\| > \frac{\gamma\lambda_{\min}^\infty}{2}$ can be estimated with Markov's inequality and Proposition B.4. There exists a constant $C > 0$ not depending on n_1 such that:

$$\mathbb{P}(Z > \frac{\gamma\lambda_{\min}^\infty}{2}) = \mathbb{P}\left(Z^p > \left(\frac{\gamma\lambda_{\min}^\infty}{2}\right)^p\right) \tag{E.92}$$

$$\leq \left(\frac{2}{\gamma\lambda_{\min}^\infty}\right)^p \mathbb{E}[\|Z\|^p] \tag{E.93}$$

$$= \left(\frac{2}{\gamma\lambda_{\min}^\infty}\right)^p \mathcal{W}_p^p(k, k_\infty) \tag{E.94}$$

$$\leq \left(\frac{2}{\gamma\lambda_{\min}^\infty}\right)^p \frac{C}{n_1^{\frac{p}{2}}}. \tag{E.95}$$

Note that Proposition B.4 holds for every natural p. This concludes the proof. □

Now are ready to prove Proposition B.9:

Proof of Proposition B.9. For the sake of clearness we introduce the following abbreviations for the remainder of the proof. Let $y_t = f(x;\theta_t), \overline{y}_t = f^{\text{lin}}(x;\overline{\theta}_t), f_t = f(\mathcal{X};\theta_t)$ and $f_t^{\text{lin}} = f^{\text{lin}}(\mathcal{X};\overline{\theta}_t)$. Also, let $k_t = k(\mathcal{X},\mathcal{X};\theta_t)$, $\nabla = \nabla_\theta$ and let $L(\mathcal{X})$ denote the Lipschitz constant of ∇f, seen as a function of θ.

Consider the empirical risk for the quadratic loss $\mathcal{R}_\mathcal{D}(\theta_t) = \frac{1}{2}\sum_{i=1}^n (f^{(L)}(x_i;\theta_t) - y)^2$.

From gradient flow equations we have:

$$\frac{\partial}{\partial t} f_t = -k_t(f_t - y), \tag{E.96}$$

$$\frac{\partial}{\partial t}\|f_t - y\|^2 = -2\langle f_t - y, k_t(f_t - y)\rangle. \tag{E.97}$$

Let $t_* = \inf\{t \mid \|\theta_t - \theta_0\| > \frac{\sigma_{\min}}{2L(\mathcal{X})}\}$ Then for each $t \le t_*$, by 1-Lipschitzianity of the smallest eigenvalue with respect to the operator norm, and by definition of t_*, we obtain an upper bound for $\lambda_{\min}(k_t)$:

$$|\lambda_{\min}(k_t) - \lambda_{\min}| \le \|k_t - k_0\|_{op} \le \|k_t - k_0\| \le L(\mathcal{X})\|\theta_t - \theta_0\| \le \frac{\sigma_{\min}}{2},$$

which implies:

$$\lambda_{\min}(k_t) \ge \lambda_{\min} - \frac{\sigma_{\min}}{2} \ge \frac{\lambda_{\min}}{4}.$$

This estimation combined with Grönwall's inequality applied to (E.97) yield:

$$\|f_t - y\|^2 \le \|f_0 - y\|^2 \exp\left(-\frac{\lambda_{\min}}{2} t\right). \tag{E.98}$$

From (E.97) and Cauchy-Schwarz we deduce:

$$\frac{\partial}{\partial t}\|f_t - y\| = -\frac{\|\nabla f_t(f_t - y)\|^2}{\|f_t - y\|} \tag{E.99}$$

$$\le -\frac{\sigma_{\min}}{2}\|\nabla f_t(f_t - y)\|. \tag{E.100}$$

Hence,

$$\frac{\partial}{\partial t}\left(\|f_t - y\| + \frac{\sigma_{\min}}{2}\|\theta_t - \theta_0\|\right) \le \frac{\partial}{\partial t}\|f_t - y\| + \frac{\sigma_{\min}}{2}\|\frac{\partial}{\partial t}\theta_t\| \le 0. \tag{E.101}$$

for all $t \le t_*$.

Thus, for all $t \le t_*$:

$$\|\theta_t - \theta_0\| \le \frac{2}{\sigma_{\min}}\|f_0 - y\|. \tag{E.102}$$

Let us show that this property holds for all $t > 0$. By contradiction assume $t_* < \infty$. (E.102) with Assumption 5 implies

$$\|\theta_{t_*} - \theta_0\| < \frac{2}{\sigma_{\min}} \frac{\sigma_{\min}^2}{4L(\mathcal{X})} \tag{E.103}$$

$$= \frac{\sigma_{\min}}{2L(\mathcal{X})}. \tag{E.104}$$

In particular the last inequality holds for t_*, which contradicts its definition. Hence $t_* = \infty$.

Let us now prove the rest of the inequalities in the theorem.

The gradient flow equation for the linearised network reads:

$$\frac{\partial}{\partial t} f_t^{\text{lin}} = -k_0(f_t^{\text{lin}} - y). \tag{E.105}$$

Define the difference $r_t = f_t - f_t^{\text{lin}}$. Then

$$\frac{\partial}{\partial t} r_t = -k_t(f_t - y) + k_0(f_t^{\text{lin}} - y) \tag{E.106}$$

$$= -k_t r_t - (k_t - k_0)(f_t^{\text{lin}} - y). \tag{E.107}$$

Then, by Cauchy-Schwarz and (E.98) combined with (E.105),

$$\frac{1}{2}\frac{\partial}{\partial t}\|r_t\|^2 = -\langle r_t, k_t r_t\rangle - \langle r_t, (k_t - k_0)(f_t^{\text{lin}} - y)\rangle \tag{E.108}$$

$$\leq -\lambda_{\min}(k_t)\|r_t\|^2 + \|r_t\|\|k_t - k_0\|\|f_t^{\text{lin}} - y\| \tag{E.109}$$

$$\leq -\frac{\lambda_{\min}}{4}\|r_t\|^2 + \|r_t\|\|k_t - k_0\|\|f_0 - y\| \exp\left(-\frac{\lambda_{\min} t}{4}\right). \tag{E.110}$$

Hence,

$$\frac{\partial}{\partial t}\|r_t\| \leq -\frac{\lambda_{\min}}{4}\|r_t\| + \|k_t - k_0\|\|f_0 - y\| \exp\left(-\frac{\lambda_{\min} t}{4}\right). \tag{E.111}$$

Now let us bound separately the different factors in the previous equation. The norm of the difference between the kernels can be estimated as:

$$\|k_t - k_0\| \leq \|\nabla f_t \nabla f_t^\top - \nabla f_0 \nabla f_0^\top\| \tag{E.112}$$

$$\leq 2\|\nabla f_0\|\|\nabla f_t - \nabla f_0\| + \|\nabla f_t - \nabla f_0\|^2 \tag{E.113}$$

$$\leq 2\sigma_{\max} L(\mathcal{X})\|\theta_t - \theta_0\| + L(\mathcal{X})^2\|\theta_t - \theta_0\|^2 \tag{E.114}$$

$$\leq 2\sigma_{\max} L(\mathcal{X})\|\theta_t - \theta_0\| + L(\mathcal{X})\|\theta_t - \theta_0\|\frac{\sigma_{\min}}{2} \tag{E.115}$$

$$\leq \frac{5}{2}\sigma_{\max} L(\mathcal{X})\|\theta_t - \theta_0\|, \tag{E.116}$$

where in (E.115) we applied the definition of t_*.

Moreover, by Grönwall and Cauchy-Schwarz inequalities we have

$$\|r_t\| \leq \exp\left(-\frac{\lambda_{\min} t}{4}\right)\|f_0 - y\| \int_0^t \|k_s - k_0\| ds \tag{E.117}$$

$$\leq \exp\left(-\frac{\lambda_{\min} t}{4}\right)\|f_0 - y\| \sup_{s\geq 0}\|k_s - k_0\| \tag{E.118}$$

$$\leq \exp\left(-\frac{\lambda_{\min} t}{4}\right)\frac{5}{2}\sigma_{\max} L(\mathcal{X})\|f_0 - y\| \sup_{s\geq 0}\|\theta_s - \theta_0\| \tag{E.119}$$

$$\leq \exp\left(-\frac{\lambda_{\min} t}{4}\right)\frac{5}{2}\sigma_{\max} L(\mathcal{X})\|f_0 - y\| \sup_{s\geq 0}\frac{2}{\sigma_{\min}}\|f_0 - y\| \tag{E.120}$$

$$\leq \exp\left(-\frac{\lambda_{\min} t}{4}\right)\frac{5\sigma_{\max}}{\sigma_{\min}} L(\mathcal{X})\|f_0 - y\|^2 \tag{E.121}$$

$$\tag{E.122}$$

Moreover, by taking the difference of the gradient flow equations for θ_t and $\overline{\theta}_t$ we obtain:

$$\frac{\partial}{\partial t}\|\theta_t - \overline{\theta}_t\| \leq \|\nabla f_t - \nabla f_0\|\|f_t - y\| + \|\nabla f_0\|\|f_t - f_t^{\text{lin}}\| \tag{E.123}$$

$$\leq L(\mathcal{X})\|\theta_t - \theta_0\|\|f_t - y\| + \sigma_{\max}\|f_t - f_t^{\text{lin}}\| \tag{E.124}$$

$$\leq \frac{2L(\mathcal{X})}{\sigma_{\min}}\|f_0 - y\|^2 \exp\left(-\frac{\lambda_{\min} t}{4}\right) \tag{E.125}$$

$$+ \frac{5\sigma_{\max}^2}{\sigma_{\min}} L(\mathcal{X})\|f_0 - y\|^2 \exp\left(-\frac{\lambda_{\min} t}{4}\right) \tag{E.126}$$

$$\leq \frac{(2 + 5\sigma_{\max}^2)L(\mathcal{X})}{\sigma_{\min}}\|f_0 - y\|^2 \exp\left(-\frac{\lambda_{\min} t}{4}\right). \tag{E.127}$$

where in (E.125) we used (E.102), (E.98) and E.121.

Integrating the previous inequality we obtain:

$$\|\theta_t - \overline{\theta}_t\| \leq \frac{(2+5\sigma_{\max}^2)L(\mathcal{X})}{\sigma_{\min}}\|f_0 - y\|^2 \int_0^t \exp\left(-\frac{\lambda_{\min} s}{4}\right) ds \tag{E.128}$$

$$\leq \frac{4(2+5\sigma_{\max}^2)L(\mathcal{X})}{\sigma_{\min}^3}\|f_0 - y\|^2 \left(1 - \exp\left(-\frac{\lambda_{\min} s}{4}\right)\right) \tag{E.129}$$

$$\leq \frac{(8+20\sigma_{\max}^2)L(\mathcal{X})}{\sigma_{\min}^3}\|f_0 - y\|^2. \tag{E.130}$$

Now we are ready to prove the last inequality in the thesis. Decompose by triangle inequality:

$$\|y_t - \overline{y}_t\| \leq \|y_t - f^{\text{lin}}(x;\theta_t)\| + \|f^{\text{lin}}(x;\theta_t) - \overline{y}_t\|. \tag{E.131}$$

First, let us focus on the first summand of (E.131). Denote by $L(x)$ the Lipschitz constant of ∇y_0 seen as a function of θ. Then, by Lemma B.15,

$$\|y_t - f^{\text{lin}}(x;\theta_t)\| = \|\int_0^t (\nabla f(x;\theta_s) - \nabla f(x;\theta_0))\frac{\partial}{\partial t}\theta_s ds\| \tag{E.132}$$

$$\leq L(x)\sup_{t\geq 0}\|\theta_t - \theta_0\| \int_0^t \|\frac{\partial}{\partial t}\theta_s\| ds \tag{E.133}$$

$$\leq L(x)\sup_{t\geq 0}\|\theta_t - \theta_0\| \cdot \frac{2}{\sigma_{\min}}\|y - f_0\| \tag{E.134}$$

$$\leq L(x)\frac{4\|y - f_0\|^2}{\lambda_{\min}}, \tag{E.135}$$

where in the third inequality we used (E.101) and (E.102) on the last one.

As for the second summand of (E.131), by (E.130) and Lemma B.15:

$$\|f^{\text{lin}}(x;\theta_t) - \overline{y}_t\| = \|\nabla f(x;\theta_0)(\theta_t - \overline{\theta}_t)\| \tag{E.136}$$

$$\leq \frac{(8+20\sigma_{\max}^2)L(\mathcal{X})}{\sigma_{\min}^3}\|f_0 - y\|^2\|\nabla f(x;\theta_0)\|. \tag{E.137}$$

Combining the two preceding estimations, we obtain the thesis.

□

Lastly, we prove Theorem B.10.

Proof of Theorem B.10. We prove the three inequalities separately. Let $\lambda_{\min}$ denote the smallest eigenvalue of k_t.

- By Lemma B.2 and Cauchy-Schwarz's inequality,

$$\|y_t - f_t\| \leq \frac{1}{\sqrt{n_1 n_0}}\text{Lip}\Phi\|x - \mathcal{X}\|\|\theta_t^{(0)}\theta_t^{(1)}\|. \tag{E.138}$$

Recall that $I_t(\lambda_{\min}) \leq t$. Then the norm $\|\theta_t^{(0)}\theta_t^{(1)}\|^2$ can be estimated with the aid of Lemma B.12:

$$\|\theta_t^{(0)}\theta_t^{(1)}\|^2 = \sum_{u=1}^{n_0}\left(\sum_{v=1}^{n_1}(\theta_{uv}^{(0)})_t(\theta_v^{(1)})_t\right)^2 \tag{E.139}$$

$$\leq \sum_{u=1}^{n_0}\left(\sum_{v=1}^{n_1}(\theta_v^{(1)})_0(\theta_{uv}^{(0)})_0 + \frac{a_1(\theta_{uv}^{(0)})_0}{\sqrt{n_1}}\psi(\theta_0)t + \frac{a_0(\theta_v^{(1)})_0}{n_1\sqrt{n_0}}\psi(\theta_0)^2t^2\right. \tag{E.140}$$

$$\left. + \frac{a_0a_1}{n_1^{\frac{3}{2}}\sqrt{n_0}}\psi(\theta_0)^3t^3 + \frac{a_0'(\theta_v^{(1)})_0^2}{\sqrt{n_1n_0}}\psi(\theta_0)t + \frac{a_0'a_1(\theta_v^{(1)})_0}{n_1\sqrt{n_0}}\psi(\theta_0)^2t^2\right)^2 \tag{E.141}$$

$$\leq \sum_{u,v} n_1(\theta_v^{(1)})_0^2(\theta_{uv}^{(0)})_0^2 + a_1^2(\theta_{uv}^{(0)})_0^2\psi(\theta_0)^2t^2 + \frac{a_0^2(\theta_v^{(1)})_0^2}{n_1n_0}\psi(\theta_0)^4t^4 \tag{E.142}$$

$$+ \frac{a_0^2a_1^2}{n_1^2n_0}\psi(\theta_0)^6t^6 + \frac{a'^2_0(\theta_v^{(1)})_0^4}{n_0}\psi(\theta_0)^2t^2 + \frac{a'^2_0a_1^2(\theta_v^{(1)})_0^2}{n_1n_0}\psi(\theta_0)^4t^4 \tag{E.143}$$

$$\leq n_1\|\theta_0^{(0)}\theta_0^{(1)}\|^2 + a_1^2\|\theta_0^{(0)}\|^2\psi(\theta_0)^2t^2 + \frac{a_0^2\|\theta_0^{(1)}\|^2}{n_1}\psi(\theta_0)^4t^4 \tag{E.144}$$

$$+ \frac{a_0^2a_1^2}{n_1}\psi(\theta_0)^6t^6 + a'^2_0\|\theta_0^{(1)}\|^4\psi(\theta_0)^2t^2 + \frac{a'^2_0a_1^2\|\theta_0^{(1)}\|^2}{n_1}\psi(\theta_0)^4t^4. \tag{E.145}$$

with $a_0 = \frac{1}{2}\|\Phi\|_\infty\|\Phi'\|_\infty\|\mathcal{X}_u\|$, $a_0' = \|\Phi'\|_\infty\|\mathcal{X}_u\|$ and $a_1 = \|\Phi\|_\infty$.

Hence,

$$\|y_t - f_t\|^2 \leq \frac{(\text{Lip}\Phi)^2\|x-\mathcal{X}\|}{n_0n_1}\|\theta_t^{(0)}\theta_t^{(1)}\|^2 \tag{E.146}$$

$$\leq \frac{A_0}{n_0}\|\theta_0^{(0)}\theta_0^{(1)}\|^2 + \frac{A_1t^2}{n_0n_1}\|\theta_0^{(0)}\|^2\psi(\theta_0)^2 + \frac{A_2\|\theta_0^{(1)}\|^2t^4}{n_1^2n_0}\psi(\theta_0)^4 \tag{E.147}$$

$$+ \frac{A_3t^6}{n_1^2n_0}\psi(\theta_0)^6 + \frac{A_4t^2}{n_1n_0}\|\theta_0^{(1)}\|^4\psi(\theta_0)^2 + \frac{A_5t^4\|\theta_0^{(1)}\|^2}{n_1^2n_0}\psi(\theta_0)^4. \tag{E.148}$$

with $A_0 = (\text{Lip}\Phi)^2\|x-\mathcal{X}\|^2$, $A_1 = a_1^2$, $A_2 = a_0^2$, $A_3 = a_1^2a_0^2$, $A_4 = a'^2_0$ and $A_5 = a'^2_0a_1^2$.

- We follow a similar strategy to prove the second inequality in the Theorem. Put $\overline{w}_t = \overline{\theta}_t - \theta_0$. By the triangle inequality and Cauchy-Schwarz we decompose:

$$\|f^{\text{lin}} - \overline{y}_t\|^2 \leq 2\|f_0 - y_0\|^2 + 2\|\nabla_\theta f_0 - \nabla_\theta y_0\|^2\|\overline{w}_t\|^2. \tag{E.149}$$

The first summand in (E.149) is bounded exactly as the first summand in (E.138) by setting $t = 0$:

$$\|f_0 - y_0\|^2 \leq \frac{(\text{Lip}\Phi)^2\|x-\mathcal{X}\|^2}{n_1n_0}\|\theta_0^{(0)}\theta_0^{(1)}\|^2 \leq \frac{A_0}{n_1n_0}\|\theta_0^{(0)}\theta_0^{(1)}\|^2. \tag{E.150}$$

As for the second summand in (E.149), we decompose by Lemma B.2. Factoring out $\max_i\{(\theta_0^{(1)})_i\} = \|\theta_0^{(1)}\|_\infty$ permits us to write:

$$\|\nabla_\theta f_0 - \nabla_\theta y_0\|^2 \leq \|\frac{\partial}{\partial\theta^{(0)}}(f_0 - y_0)\|^2 + \|\frac{\partial}{\partial\theta^{(1)}}(f_0 - y_0)\|^2 \tag{E.151}$$

$$\leq \frac{1}{n_1 n_0}\|(\mathcal{X}^\top\Phi'(\frac{1}{\sqrt{n_0}}\mathcal{X}\theta_0^{(0)}) - x^\top\Phi'(\frac{1}{\sqrt{n_0}}x\theta_0^{(0)}))\theta_0^{(1)}\|^2 \tag{E.152}$$

$$+ \frac{1}{n_1}\|\Phi(\frac{1}{\sqrt{n_0}}\mathcal{X}\theta_0^{(0)}) - \Phi(\frac{1}{\sqrt{n_0}}x\theta_0^{(0)})\|^2 \tag{E.153}$$

$$\leq \frac{2}{n_1 n_0}\|(\mathcal{X}^\top\Phi'(\frac{1}{\sqrt{n_0}}\mathcal{X}\theta_0^{(0)}) - \mathcal{X}^\top\Phi'(\frac{1}{\sqrt{n_0}}x\theta_0^{(0)}))\theta_0^{(1)}\|^2 \tag{E.154}$$

$$+ \frac{2}{n_1 n_0}\|(\mathcal{X}^\top\Phi'(\frac{1}{\sqrt{n_0}}x\theta_0^{(0)}) - x^\top\Phi'(\frac{1}{\sqrt{n_0}}x\theta_0^{(0)}))\theta_0^{(1)}\|^2 \tag{E.155}$$

$$+ \frac{A_0}{n_1 n_0}\|\theta_0^{(0)}\|^2 \tag{E.156}$$

$$\leq \frac{2}{n_1 n_0^2}(\text{Lip}\Phi')^2\|\mathcal{X}\|^2\|x - \mathcal{X}\|^2\|\theta_0^{(0)}\theta_0^{(1)}\|^2 \tag{E.157}$$

$$+ \frac{2}{n_1 n_0}\|\Phi'\|_\infty^2\|x - \mathcal{X}\|^2\|\theta_0^{(1)}\|^2 + \frac{A_0}{n_1 n_0}\|\theta_0^{(0)}\|^2. \tag{E.158}$$

Moreover we can bound the norm of $\overline{w}_t$ with Lemma B.12:

$$\|\overline{w}_t\|^2 \leq \|\overline{\theta}_t^{(0)} - \theta_0^{(0)}\|^2 + \|\overline{\theta}_t^{(1)} - \theta_0^{(1)}\|^2 \tag{E.159}$$

$$\leq \sum_{u,v}\frac{2a_0^2}{n_1^2 n_0}\psi(\theta_0)^4 t^4 + \frac{2a_0'^2(\theta_v^{(1)})_0^2}{n_1 n_0}\psi(\theta_0)^2 t^2 + \sum_v \frac{a_1^2}{n_1}\psi(\theta_0)^2 t^2 \tag{E.160}$$

$$\leq \frac{2a_0^2}{n_1}\psi(\theta_0)^4 t^4 + \frac{2a_0'^2\|\theta^{(1)}\|^2}{n_1}\psi(\theta_0)^2 t^2 + a_1^2\psi(\theta_0)^2 t^2. \tag{E.161}$$

Hence, (E.149) can be written as:

$$\|f^{\text{lin}} - \overline{y}_t\|^2 \leq \frac{B_0}{n_1 n_0}\|\theta_0^{(0)}\theta_0^{(1)}\|^2 + \frac{B_1}{n_1^2 n_0^2}\|\theta_0^{(0)}\theta_0^{(1)}\|^2\psi(\theta_0)^4 t^4 \tag{E.162}$$

$$+ \frac{B_2}{n_1^2 n_0^2}\|\theta_0^{(0)}\|^2\|\theta_0^{(1)}\|^4\psi(\theta_0)^2 t^2 + \frac{B_3}{n_1 n_0^2}\|\theta_0^{(0)}\theta_0^{(1)}\|^2\psi(\theta_0)^2 t^2 \tag{E.163}$$

$$+ \frac{B_4}{n_1^2 n_0}\|\theta_0^{(1)}\|^2\psi(\theta_0)^4 t^4 + \frac{B_5}{n_1^2 n_0}\|\theta_0^{(1)}\|^4\psi(\theta_0)^2 t^2 \tag{E.164}$$

$$+ \frac{B_6}{n_1 n_0}\|\theta_0^{(1)}\|^2\psi(\theta_0)^2 t^2 + \frac{B_7}{n_1^2 n_0}\|\theta_0^{(0)}\|^2\psi(\theta_0)^4 t^4 \tag{E.165}$$

$$+ \frac{B_8}{n_1^2 n_0}\|\theta_0^{(0)}\|^2\|\theta_0^{(1)}\|^2\psi(\theta_0)^2 t^2 + \frac{B_9}{n_1 n_0}\|\theta_0^{(0)}\|^2\psi(\theta_0)^2 t^2, \tag{E.166}$$

where the constants in the last inequality are, explicitly, $B_0 = 2A_0$, $B_1 = 8(\text{Lip}\Phi'\|\mathcal{X}\|\|x - \mathcal{X}\|a_0)^2$, $B_2 = 8(\text{Lip}\Phi'\|\mathcal{X}\|\|x - \mathcal{X}\|a_0')^2$, $B_3 = 4(\text{Lip}\Phi'\|\mathcal{X}\|\|x - \mathcal{X}\|a_1)^2$, $B_4 = 8(\|\Phi'\|_\infty\|x - \mathcal{X}\|a_0)^2$, $B_5 = 8(\|\Phi'\|_\infty\|x - \mathcal{X}\|a_0')^2$, $B_6 = 4(\|\Phi'\|_\infty\|x - \mathcal{X}\|a_1)^2$, $B_7 = 4A_0 a_0^2$, $B_8 = 4A_0 a_0'^2$, and $B_9 = 2A_0 a_1^2$.

- It remains to estimate the last inequality. Consider $\Delta(t) = \|f_t - f_t^{\text{lin}}\|$ Then by gradient flow equations and Cauchy-Schwarz,

$$\frac{\partial}{\partial t}(\Delta(t)^2) = \langle k_t(f_t - y) - k_0(f_t^{\text{lin}} - y), f_t - f_t^{\text{lin}} \rangle \tag{E.167}$$

$$= \sum_{i=1}^{n} (k_t(x_i, \mathcal{X})(f_t - y) - k_0(x_i, \mathcal{X})(f_t^{\text{lin}} - y))(f_t(x_i) - f_t^{\text{lin}}(x_i)) \tag{E.168}$$

$$= \sum_{i=1}^{n} (k_t(x_i, \mathcal{X}) - k_0(x_i, \mathcal{X}))(f_t - y)(f_t(x_i) - f_t^{\text{lin}}(x_i)) \tag{E.169}$$

$$- k_0(x_i, \mathcal{X})(f_t - f_t^{\text{lin}})(f_t(x_i) - f_t^{\text{lin}}(x_i)) \tag{E.170}$$

$$= \|(k_t - k_0)(f_t - y)(f_t - f_t^{\text{lin}})^\top\|_1 - \|k_0(f_t - f_t^{\text{lin}})(f_t - f_t^{\text{lin}})^\top\|_1 \tag{E.171}$$

By equivalence of the 1-norm and the euclidean norm for $v \in \mathbb{R}^d$ we have $\|v\| \leq \|v\|_1 \leq \sqrt{d}\|v\|$. Then, by Cauchy-Schwarz's inequality,

$$\frac{\partial}{\partial t}\Delta(t) = n\|k_t - k_0\|\|f_t - y\| - \lambda_{\text{min}}^0 \Delta(t) \tag{E.172}$$

$$\leq n\|k_t - k_0\|\psi(\theta_0)e^{-\lambda_{\text{min}} t} - \lambda_{\text{min}}^0 \Delta(t) \tag{E.173}$$

Let us bound the norm of $k_t - k_0$:

$$\|k_t - k_0\| = \|\nabla_\theta f_t(\mathcal{X})\nabla_\theta f_t(\mathcal{X})^\top - \nabla_\theta f_0(\mathcal{X})\nabla_\theta f_0(\mathcal{X})^\top\| \tag{E.174}$$

$$\leq \|\nabla_\theta f_t(\mathcal{X}) + \nabla_\theta f_0(\mathcal{X})\|L(\mathcal{X})\|\theta_t - \theta_0\| \tag{E.175}$$

$$\tag{E.176}$$

From Lemmas B.2 and B.12, we have:

$$\|\nabla_\theta f_t(\mathcal{X})\|^2 = \|\nabla_{\theta^{(0)}} f_t(\mathcal{X})\|^2 + \|\nabla_{\theta^{(1)}} f_t(\mathcal{X})\|^2 \tag{E.177}$$

$$\leq \frac{\|\mathcal{X}\|^2\|\Phi'\|_\infty^2\|\theta_t^{(1)}\|^2}{n_1 n_0} + \frac{\|\Phi\|_\infty^2}{n_1} \tag{E.178}$$

$$\leq \frac{\|\mathcal{X}\|^2\|\Phi'\|_\infty^2}{n_1 n_0}\left(\|\theta_0^{(1)}\| + \frac{\|\Phi\|_\infty \psi(\theta^0)}{\sqrt{n_1}}t\right)^2 + \frac{\|\Phi\|_\infty^2}{n_1}. \tag{E.179}$$

Analogously,

$$\|\nabla_\theta f_0(\mathcal{X})\|^2 = \|\nabla_{\theta^{(0)}} f_0(\mathcal{X})\|^2 + \|\nabla_{\theta^{(1)}} f_0(\mathcal{X})\|^2 \tag{E.180}$$

$$\leq \frac{\|\mathcal{X}\|^2\|\Phi'\|_\infty^2\|\theta_0^{(1)}\|^2}{n_1 n_0} + \frac{\|\Phi\|_\infty^2}{n_1}. \tag{E.181}$$

Moreover, again by Lemma B.12,

$$\|\theta_t - \theta_0\|^2 = \|\theta_t^{(0)} - \theta_0^{(0)}\|^2 + \|\theta_t^{(1)} - \theta_0^{(1)}\|^2 \tag{E.182}$$

$$\leq \frac{2a_0^2}{n_1^2 n_0}\psi(\theta_0)^4 t^4 + \frac{2{a_0'}^2}{n_1 n_0}\|\theta_0^{(1)}\|^2 t^2 + \frac{a_1^2 \psi(\theta^0)^2}{n_1}t^2 \tag{E.183}$$

$$\tag{E.184}$$

Inequalities (E.179),(E.181) and (E.183) allow us to estimate:

$$\|k_t - k_0\| \leq \frac{L(\mathcal{X})}{n_1}\left(\frac{c_1\psi(\theta_0)^2\|\theta_0^{(1)}\|}{\sqrt{n_1}n_0} + \frac{c_2\psi(\theta_0)^3 t^3}{n_1 n_0} + \frac{c_3\psi(\theta_0)^2 t^2}{\sqrt{n_1 n_0}}\right. \tag{E.185}$$

$$+\frac{c_4\|\theta_0^{(1)}\|^2 t}{n_0} + \frac{c_5\psi(\theta_0)\|\theta_0^{(1)}\|t^2}{\sqrt{n_1}n_0} + \frac{c_6\|\theta_0^{(1)}\|t}{\sqrt{n_0}} \tag{E.186}$$

$$\left.+\frac{c_7\psi(\theta_0)\|\theta_0^{(1)}\|t}{\sqrt{n_0}} + \frac{c_8\psi(\theta_0)^2 t^2}{\sqrt{n_1 n_0}} + c_9\psi(\theta_0)t\right), \tag{E.187}$$

with $c_1 = 2\sqrt{2}a_0\|\mathcal{X}\|\|\Phi'\|_\infty$, $c_2 = \sqrt{2}a_0\|\Phi\|_\infty$, $c_3 = 2\sqrt{2}a_0\|\Phi\|_\infty$, $c_4 = 2\sqrt{2}a_0'\|\mathcal{X}\|\|\Phi'\|_\infty$, $c_5 = \sqrt{2}a_0'\|\Phi\|_\infty$, $c_6 = 2\sqrt{2}a_0'\|\Phi\|_\infty$, $c_7 = 2a_1\|\mathcal{X}\|\|\Phi'\|_\infty$, $c_8 = a_1\|\Phi\|_\infty$ and $c_9 = 2a_1\|\Phi\|_\infty$. Let $C(n_1, n_0, t, \theta_0)$ be the right hand side of (E.185). Then, the reight-hand side of (E.173) can be $\mathbb{P}$-almost surely bounded from above with:

$$\frac{\partial}{\partial t}\Delta(t) \leq n\|k_t - k_0\|\psi(\theta_0)e^{-\lambda_{\min}t} - \lambda_{\min}^0\Delta(t) \tag{E.188}$$

$$\leq nC(n_1, n_0, t, \theta_0). \tag{E.189}$$

In the previous inequality we used that the event $\lambda_{\min} = 0$ has null measure. Integrating, and using that $\Delta(0) = 0$:

$$\Delta(t) \leq \frac{nL(\mathcal{X})}{n_1}\left(\frac{c_1\psi(\theta_0)^2\|\theta_0^{(1)}\|t}{\sqrt{n_1}n_0} + \frac{c_2\psi(\theta_0)^3t^4}{2n_1n_0} + \frac{c_3\psi(\theta_0)^2t^3}{\sqrt{n_1n_0}}\right. \tag{E.190}$$

$$+\frac{c_4\|\theta_0^{(1)}\|^2t^2}{2n_0} + \frac{c_5\psi(\theta_0)\|\theta_0^{(1)}\|t^3}{3\sqrt{n_1}n_0} + \frac{c_6\|\theta_0^{(1)}\|t^3}{2\sqrt{n_0}} \tag{E.191}$$

$$\left.+\frac{c_7\psi(\theta_0)\|\theta_0^{(1)}\|t^2}{2\sqrt{n_0}} + \frac{c_8\psi(\theta_0)^2t^3}{3\sqrt{n_1n_0}} + \frac{c_9\psi(\theta_0)t^2}{2}\right) \tag{E.192}$$

Taking the square, applying the elementary inequality $(\sum_{i=1}^n a_i)^2 \leq n\sum_{i=1}^n a_i^2$, for $a_i \geq 0$, and adjusting the constants yields the desired result.

□