

## 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  $t \geq 0$ :  $\bar{y}_t = f_t^{\text{lin}}(x)$ ,  $f_t^{\text{lin}} = f_t(\mathcal{X})$ ,  $k_{\mathcal{X}\mathcal{X}} = k_0(\mathcal{X}, \mathcal{X})$  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  $B_n \rightarrow B$  we have  $\lim_{n \rightarrow \infty} I_t(B_n) = I_t(B)$ .*

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:

$$\dot{f}_t^{\text{lin}} = -k_{\mathcal{X}\mathcal{X}}(f_t^{\text{lin}} - y), \quad (\text{A.1})$$

$$\dot{\bar{y}}_t = -k_{x\mathcal{X}}(f_t^{\text{lin}} - y). \quad (\text{A.2})$$

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  $f(t): \mathbb{R} \rightarrow \mathbb{R}^m$ ,  $A(t): \mathbb{R} \rightarrow \mathbb{R}^{m \times m}$  are integrable functions, and  $t > 0$ . Therefore in our case, by letting  $u_t = f_t^{\text{lin}} - y$ :

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

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

$$\dot{v}_t = -k_{x\mathcal{X}} \exp(-k_{\mathcal{X}\mathcal{X}} t) (f_0 - y), \quad v_0 = y_0,$$

which, by integrating and by using Definition 2.2, becomes

$$v_t - v_0 = \bar{y}_t - \bar{y}_0 \quad (\text{A.4})$$

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

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

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:

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

### 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  $\bar{y}_t - \bar{y}_0 = B_t f_0 + C_t y$ . Also recall that  $k_t = k_0$  for all  $t \geq 0$  since  $f$  is linear on  $\theta$ . Note that  $f_0$  and  $\bar{y}_0$  are centered Gaussian processes and hence  $\mathbb{E}[\bar{y}_0 + B_t f_0] = 0$ . Therefore, taking the expected value of the wide limit yields:

$$\mathbb{E}[\lim_{n_1 \rightarrow \infty} \bar{y}_t] = \mathbb{E}[\lim_{n_1 \rightarrow \infty} C_t y] = k_\infty(x, \mathcal{X}) I_t(k_\infty) y, \quad (\text{A.7})$$

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,

$$\text{Cov}\left(\lim_{n_1 \rightarrow \infty} \bar{y}_t, \lim_{n_1 \rightarrow \infty} y'_t\right) = \mathbb{E}\left[\lim_{n_1 \rightarrow \infty} (\bar{y}_t - \mathbb{E}[\bar{y}_t])(y'_t - \mathbb{E}[y'_t])\right] \quad (\text{A.8})$$

$$= \mathbb{E}\left[\lim_{n_1 \rightarrow \infty} (\bar{y}_0 + B_t f_0)(y'_0 + B'_t f_0)\right] \quad (\text{A.9})$$

$$= \mathbb{E}\left[\lim_{n_1 \rightarrow \infty} \bar{y}_0 y'_0\right] + \mathbb{E}\left[\lim_{n_1 \rightarrow \infty} \bar{y}_0 f_0 B'_t\right] \quad (\text{A.10})$$

$$+ \mathbb{E}\left[\lim_{n_1 \rightarrow \infty} y'_0 f_0 B_t\right] + \mathbb{E}\left[\lim_{n_1 \rightarrow \infty} f_0^2 B_t B'_t\right] \quad (\text{A.11})$$

$$= \mathcal{K}(x, x') - \mathcal{K}(x, \mathcal{X}) I_t(k_\infty) k_\infty(\mathcal{X}, x') \quad (\text{A.12})$$

$$- k_\infty(x, \mathcal{X}) I_t(k_\infty) \mathcal{K}(\mathcal{X}, x') \quad (\text{A.13})$$

$$+ k_\infty(x, \mathcal{X}) I_t(k_\infty) \mathcal{K}(\mathcal{X}, \mathcal{X}) I_t(k_\infty) k_\infty(\mathcal{X}, x'). \quad (\text{A.14})$$

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  $t \geq 0$ :  $y_t = f_t(x)$ ,  $f_t = f_t(\mathcal{X})$ ,  $k_t = k_t(\mathcal{X}, \mathcal{X})$  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  $\xi$  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  $x, x' \in \mathbb{R}^{n_0}$ :*

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

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

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

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

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

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

## 788 B.1 Results at initialization

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

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

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

794 Furthermore, if  $\varphi: \mathbb{R} \rightarrow \mathbb{R}$  is a Lipschitz function, then

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

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

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

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

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

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

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

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

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

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

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

810 Now we are ready to show the following result:

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

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

812 with  $C$  not depending on the width  $n_1$ .

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

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

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

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

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

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

820 In this subsection we state two results paramount to prove Proposition 3.6, along with all their  
 821 auxiliary lemmas. The following theorem resembles Theorem 5.4 in Bartlett et al. [2021], or Theorem  
 822 2.2 in Chizat et al. [2019], but we would like to remark that those result differ from ours since the first  
 823 applies to neural networks in which the training is restricted to  $\theta^{(0)}$  while keeping  $\theta^{(1)}$  frozen, and  
 824 in both of them the loss function used for training is different than the one considered in our work.  
 825 The proof is similar to the one in Bartlett et al. [2021] and is carried in full detail in Supplementary  
 826 Material E.

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

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

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

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

832 **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  
 833 hold. Then the following hold for  $t > 0$  and for any test point  $x \in \mathbb{R}^{n_0}$ :

$$\|\theta_t - \theta_0\| \leq \frac{2}{\sigma_{\min}} \|f_0 - y\|, \quad (\text{B.13})$$

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

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

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

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

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

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

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

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

842

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

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

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

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

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

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

$$\left. + \frac{C_4 \|\theta_0^{(1)}\|^4 t^4}{n_0^2} + \frac{C_5 \psi(\theta_0)^2 \|\theta_0^{(1)}\|^2 t^6}{n_1 n_0^2} + \frac{C_6 \|\theta_0^{(1)}\|^2 t^6}{n_0} \right. \quad (\text{B.26})$$

$$\left. + \frac{C_7 \psi(\theta_0)^2 \|\theta_0^{(1)}\|^2 t^4}{n_0} + \frac{C_8 \psi(\theta_0)^4 t^6}{n_1 n_0} + C_9 \psi(\theta_0)^2 t^4 \right), \quad (\text{B.27})$$

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

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

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

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

855 **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  
856  $\mathcal{X}_u = ((x_i)_u)_{i=1}^n \in \mathbb{R}^n$ . Let  $\lambda_{\min}$  and  $\lambda_{\min}^0$  be the smallest eigenvalues of  $k_t$  and  $k_0$ , respectively.  
857 Then the following inequalities hold:

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

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

$$+ \frac{\|\Phi'\|_{\infty} \|f_0 - y\| \|\mathcal{X}_u\|}{\sqrt{n_1 n_0}} (\theta_v^{(1)})_0 I_t(\lambda_{\min}), \quad (\text{B.30})$$

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

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

$$+ \frac{\|\Phi'\|_{\infty} \|f_0 - y\| \|\mathcal{X}_u\|}{\sqrt{n_1 n_0}} (\bar{\theta}_v^{(1)})_0 I_t(\lambda_{\min}^0). \quad (\text{B.33})$$

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

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

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

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

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

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

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

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

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

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

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

870 *with probability greater or equal than  $1 - \frac{1}{\frac{r}{n_1^{\frac{r}{2}}} - 1} - \exp(-\gamma n_1)$ , where  $f_0(x): \mathbb{R}^N \rightarrow \mathbb{R}$  is understood*  
 871 *as a function of  $\theta$ .*

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

874 **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*  
 875  *$\lambda_{\min}^\infty = \lambda_{\min}(k_\infty)$ . Then, for each  $p \in \mathbb{N}$ ,*

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

876 *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*  
 877 *on  $n_1$ .*

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

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

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

$$\lambda_{\min} \geq \lambda_{\min}^\infty - \|k - k_\infty\| \geq \frac{\lambda_{\min}^\infty}{2}. \quad (\text{B.40})$$

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

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

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

$$\lambda_{\max} \leq \lambda_{\max}^\infty + \frac{\lambda_{\min}^\infty}{2}. \quad (\text{B.42})$$

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

## 886 C Proof of Theorems 3.4 and 3.6

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

891 Let us begin by Proposition 3.6:

892 *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,

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

$$= \int_S \|y_t - \bar{y}_t\|^2 d\theta + \int_{S^c} \|y_t - \bar{y}_t\|^2 d\theta \quad (\text{C.2})$$

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

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  $\theta$  in  $S$ . Indeed, by Lemmas B.2 and B.15:

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

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

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  $Z \in \{x, \mathcal{X}\}$ :

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

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

$$\lambda_{\max} \leq \lambda_{\max}^\infty + \frac{\lambda_{\min}^\infty}{2}. \quad (\text{C.8})$$

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  $\theta \in S$ :

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

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

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

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,

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

$$\leq \sup_{\theta \in S} \|y_t - \bar{y}_t\|^2 \quad (\text{C.13})$$

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

908 Put  $\alpha_1 = c^2$ .

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

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

911 Also, define the subset

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

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

913 Intuitively,  $\Omega_*$  and  $\bigcup_{\gamma \in \mathbb{N}} \Omega_\gamma$  are the events in which the lower bound for the smallest eigenvalue  
 914 of the empirical kernel and the upper bound of the Frobenius and sup-norm of the parameters at  
 915 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  
 916 partition of  $\mathbb{R}^N$  to finish the proof.

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

$$\int_{S^C} \|y_t - \bar{y}_t\|^2 d\theta \leq \int_{\Omega} \|y_t - \bar{y}_t\|^2 d\theta + \sum_{\gamma \in \mathbb{N}} \int_{D_\gamma} \|y_t - \bar{y}_t\|^2 d\theta \quad (\text{C.17})$$

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

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

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

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

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

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

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

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

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

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



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

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

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

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

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

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

$$+ \frac{16\gamma^3 t^4}{n_1 n_0} + \frac{64\gamma^3 t^6}{7n_1^2 n_0} + \frac{28\gamma^3 n_1 t^2}{n_0} + \frac{16\gamma^3 t^4}{n_1 n_0}). \quad (\text{C.31})$$

Since the negative exponential function decreases 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  $\gamma$ :

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

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

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

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

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

$$+ \frac{16\gamma^3 t^4}{n_1 n_0} + \frac{28\gamma^3 t^2}{n_0} + \frac{4\gamma^2 t^2}{n_0} + 16\gamma^3 t^4 n_0 + 28\gamma^3 t^2 + 4\gamma^2 t^2 \quad (\text{C.36})$$

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

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

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

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

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

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

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

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

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

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

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

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

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

937 Recall that we can choose  $r$  large enough so that all the summands have  $n_1$  on the denomi-  
 938 nator. In particular, it is enough to choose  $r \geq 5$  in this case. Moreover, by reasoning like in  
 939 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  
 940 bounded from above by  $n_1^{-\frac{r\gamma}{4}}$ . Therefore, up to a multiplicative constant not depending on  
 941  $n_1, n_0, t$  nor  $\gamma$ :

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

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

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

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

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

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

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

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

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

946  
947

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

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

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

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

948

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

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

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

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

949 Now we finish the proof by estimating the last 3 terms in (C.21). Recall that, by definition of  $\Omega$ , the  
950 bounds for the Frobenius and the sup-norms of the parameters at initialization used in the estimation  
951 of  $\int_S \|y_t - \bar{y}_t\| d\theta$  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  
952 suffices to choose  $p$  large enough to counterattack the biggest exponent of  $n_1$  in each of the bounds  
953 given by Theorem B.10. In particular, as seen while choosing  $r$  when computing the six series, it is  
954 enough to take  $p = r \geq 5$ . Then, up to multiplicative constants,

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

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

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

$$\leq Cdr R_1 (1+t^8) \frac{1}{\sqrt{n_1} n_0^2}. \quad (\text{C.62})$$

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

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

957

This concludes the proof.  $\square$

958

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

959

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

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

961

Then the thesis follows estimating the first summand with Proposition 3.6 and the second one with  
962 Proposition 3.7. For large enough  $n_1$  we can take the constants  $a_1$  and  $a_2$  to be a multiple of  $\alpha_1$   
963 and  $\alpha_2$  in Proposition 3.6; since the right-hand side in the statement in that result decreases as

964

$\frac{\log n_1}{n_1} + \frac{1}{n_1^{\frac{r}{4}}}$ , which is strictly slower than the right-hand side of the statement in Proposition 3.7 for

965

any  $n_1 \geq 2$ .  $\square$

## 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 inequality and Young's inequality

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

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

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

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

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  $b_t = \frac{e^{-\frac{\lambda_{\min}^\infty}{2}t}}{\lambda_{\min}^\infty} \|k_{\mathcal{X}\mathcal{X}} - k_\infty\|^2 \|y - f_0\|^2$  yields:

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

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

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

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

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:

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

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

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

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:

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

$$\leq \frac{C e^{-\lambda_{\min}^\infty t}}{n_1} (t + 1), \quad (\text{D.12})$$

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:

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

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

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

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

$$\begin{aligned} (y_t - G_t^x) &\leq (y_0 - G_0^x) + \|k_{x\mathcal{X}} - k_{\infty}^x\| \|y - f_0\| (1 - e^{-\lambda_{\min}^{\infty} t}) + \frac{\bar{\lambda} C}{\sqrt{n_1}} \int_0^t \sqrt{s+1} e^{-\frac{\lambda_{\min}^{\infty}}{2} s} ds \\ &\leq (y_0 - G_0^x) + \|k_{x\mathcal{X}} - k_{\infty}^x\| \|y - f_0\| (1 - e^{-\lambda_{\min}^{\infty} t}) + \frac{\bar{\lambda} C D}{\sqrt{n_1}} (2 - \sqrt{t+1} e^{-\frac{\lambda_{\min}^{\infty}}{2} t}), \end{aligned} \quad (\text{D.15})$$

(D.16)

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

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

$$= 2e^{\frac{\lambda_{\min}^{\infty}}{2}} \int u^2 e^{-\frac{\lambda_{\min}^{\infty}}{2} u^2} du \quad (\text{D.18})$$

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

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

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

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

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

988 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  
 989 the fourth step we used the definition of Gauss error function  $\operatorname{erf}(x)$ ; and  $C$  denotes the integration  
 990 constant. Therefore,

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

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

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

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

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

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

993 Turning back to (D.16), we can finish by applying the elementary inequality  $(a + b + c)^2 \leq$   
 994  $3a^2 + 3b^2 + 3c^2$  for  $a, b, c \geq 1$  and Hölder's inequality. After that, Theorem B.3, Proposition B.4 and  
 995 Lemma B.8 can be applied in the same fashion as in the proof of the training case, yielding positive  
 996 constants  $d_1, d_2$  and  $d_3$  such that:

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

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

$$+ \frac{3(\bar{\lambda}CD)^2}{n_1} (2 - \sqrt{t+1} e^{-\frac{\lambda_{\min}^{\infty} t}{2}})^2 \quad (\text{D.30})$$

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

$$+ \frac{3(\bar{\lambda}CD)^2}{n_1} (4 + (t+1)e^{-\lambda_{\min}^{\infty} t}). \quad (\text{D.32})$$

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

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

## 1000 E Proofs of the auxiliary results

1001 We present here all the remaining proofs. For clearness, we will use the following notation:  $X_v =$   
 1002  $h(x)_v$  and  $X'_v = h(x')_v$  for each  $1 \leq v \leq n_1$ , for any  $x, x' \in \mathbb{R}^{n_0}$ .

1003 *Proof of Lemma A.1.* It is trivial when  $B$  is nonsingular. Let  $\lambda_1, \dots, \lambda_n$  be the (possible repeated)  
 1004 ordered eigenvalues of  $B$  and suppose  $\lambda_j = 0$ . Then, by using the eigenvalue decomposition of  $B$   
 1005 and elementary properties of the matrix exponential:

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

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

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

$$= \mathbb{I}_n - e^{-Bt}. \quad (\text{E.4})$$

1006 The converse equality is proven in an analogous way.

1007 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  
 1008 converging to  $a \in \mathbb{R}$ . If  $a \neq 0$ , or 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

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

□

*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  $\pi^{XY}$  be an optimal transport plan between  $X$  and  $Y$ , let  $\pi^{XZ}$  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  $\pi$  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  $\pi^{XZ}$  and integrating with respect to  $z$  yields  $\pi^{XY}$ . Note also that there is a unique coupling between  $Y$  and  $\lambda$ , which is given by  $\mathbb{P}_Y \times \delta_\lambda$ , hence,

$$\mathcal{W}_p^p(\tilde{X}, \tilde{Y}) \leq \int_{\mathbb{R}^{n+m+n}} \int_{\mathbb{R}^m} \|\tilde{X}(x, z) - \tilde{Y}(y, w)\|^p \delta_\lambda(dw) d\pi(dx, dz, dy) \quad (\text{E.5})$$

$$= \int_{\mathbb{R}^{n+m+n}} \|\tilde{X}(x, z) - \tilde{Y}(y, \lambda)\|^p d\pi(dx, dz, dy) \quad (\text{E.6})$$

$$\leq \int_{\mathbb{R}^{n+n}} \|X - Y\|^p d\pi^{XY}(dx, dy) + \int_{\mathbb{R}^m} \|Z - \lambda\|^p d\mu^Z(dz) \quad (\text{E.7})$$

$$= \mathcal{W}_p^p(X, Y) + \mathcal{W}_p^p(Z, \lambda), \quad (\text{E.8})$$

where  $\mu^Z$  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  $\sigma$ -algebra. Then

$$\mathcal{W}_p^p(\tilde{X}, \tilde{Y}) \leq \mathbb{E}_{\mu \times \nu}[\|\tilde{X} - \tilde{Y}\|^p] \quad (\text{E.9})$$

$$\leq \mathbb{E}_{\mu \times \nu}[(\|X - Y\|^2 + \|Z - V\|^2)^{\frac{p}{2}}] \quad (\text{E.10})$$

$$\leq 2^{\frac{p}{2}-1} (\mathbb{E}_\mu[\|X - Y\|^{2p}] + \mathbb{E}_\nu[\|Z - V\|^{2p}]), \quad (\text{E.11})$$

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  $\sqrt{a + b} \leq \sqrt{a} + \sqrt{b}$ :

$$\mathcal{W}_1(\tilde{X}, \tilde{Y}) \leq \mathbb{E}_{\mu \times \nu}[\|\tilde{X} - \tilde{Y}\|] \quad (\text{E.12})$$

$$\leq \mathbb{E}_{\mu \times \nu}[(\|X - Y\|^2 + \|Z - V\|^2)^{\frac{1}{2}}] \quad (\text{E.13})$$

$$\leq \mathbb{E}_\mu[\|X - Y\|] + \mathbb{E}_\nu[\|Z - V\|]. \quad (\text{E.14})$$

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  $1 \leq i, j \leq n_1$ :

$$\tilde{k}_{ij}(x, x') = (\nabla_{\theta^{(0)}} X_i) (\nabla_{\theta^{(0)}} X'_j) \quad (\text{E.15})$$

$$= \frac{1}{n_0} \sum_{\substack{u=1, \dots, n_0 \\ v=1, \dots, n_1}} \frac{\partial}{\partial \theta_{uv}^{(0)}} (x \theta_{-i}^{(0)}) \frac{\partial}{\partial \theta_{uv}^{(0)}} (x' \theta_{-j}^{(0)}) \quad (\text{E.16})$$

$$= \frac{1}{n_0} \sum_{\substack{u=1, \dots, n_0 \\ v=1, \dots, n_1}} x_u \delta_{iv} x'_u \delta_{jv} \quad (\text{E.17})$$

$$= \frac{1}{n_0} \sum_{u=1}^{n_0} x_u x'_u \delta_{ij}. \quad (\text{E.18})$$

1028 Lastly,

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

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

$$+ \frac{1}{n_1} \sum_{z=1}^{n_1} \frac{\partial}{\partial \theta_z^{(1)}} \left( \Phi \left( \frac{1}{\sqrt{n_0}} x \theta^{(0)} \right) \theta^{(1)} \right) \frac{\partial}{\partial \theta_z^{(1)}} \left( \Phi \left( \frac{1}{\sqrt{n_0}} x' \theta^{(0)} \right) \theta^{(1)} \right) \quad (\text{E.21})$$

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

1029

□

### 1030 E.1 Proof of results at initialization

1031 In this subsection we prove the useful Proposition B.4, which generalises the second part of Theorem  
1032 B.3. This step is paramount for the proof of the rest of our results. From now to the end of this  
1033 subsection assume  $\Phi$  and  $\Phi'$  are bounded and  $x, x' \in \mathbb{R}^{n_0}$  are fixed.

1034 *Proof of Proposition B.6.* Note that, from Lemma B.2,  $k$  can be written as:

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

1035 and recall that  $\tilde{k}$  is deterministic, for any fixed inputs  $x, x'$ . Hence, by boundedness of  $\Phi$  and  $\Phi'$ , and  
1036 independence of  $\theta^{(1)}$  and  $\theta^{(0)}$ ,

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

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

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

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

1037 where in the last inequality we used that the  $2p$ -th moment of a standard Gaussian variable is equal to  
1038  $(2p-1)!!$ .

1039

□

1040 *Proof of Proposition B.4.* We will use the notation introduced in Proposition B.6. The first claim is  
1041 trivial since  $\tilde{k}$  coincides with  $\mathcal{K}$ . To show the second claim, we split:

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

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

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

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



1042 To bound the first summand in (E.29) we split again by adding and subtracting an auxiliary term:

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

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

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

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

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

1043 The first summand in (E.34) vanishes since  $\tilde{k}$  equals  $\mathcal{K}$ . We estimate the second summand in (E.34),  
1044 once again by adding and subtracting an auxiliary term:

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

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

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

1045 The first summand in (E.37) vanishes, by boundedness of  $\Phi'$  and independence of the parameters  
1046  $\theta_v^{(1)}$ :

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

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

$$= 0. \quad (\text{E.41})$$

1047 As for the second summand in (E.37), by Theorem B.3 there exists a constant  $c_1$  not depending on  $n_1$   
1048 such that:

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

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

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

1049 It remains only to bound the second summand in (E.29). This is done by using again Theorem B.3.  
 1050 There exists a constant  $c_2$  not depending on  $n_1$  such that:

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

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

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

1051 Putting together all the preceding estimations we obtain:

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

1052 with  $C = 2^{p-1} \max\{2^{2p-2} c_1 (\text{Lip}\Phi' + \Phi'(0)), c_2 (\text{Lip}\Phi + \Phi(0))\}$ .  $\square$

1053 These results suffice to prove Proposition B.7.

1054 *Proof of Proposition B.7.* Consider the joint random variables  $\tilde{X} = (k_{\mathcal{X}\mathcal{X}}, \hat{f}(\mathcal{X}))$  and  $\tilde{Y} =$   
 1055  $(k_\infty(\mathcal{X}, \mathcal{X}), G(\mathcal{X}))$ . Then Lemma B.1.4 together with Proposition B.4 and Theorem B.3 yield

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

1056 where  $C$  is the constant in Proposition B.4 and  $D$  the one in Theorem B.3. Both constants do not  
 1057 depend on  $n_1$ .  $\square$

1058 Lastly, we prove Lemma B.8.

1059 *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  
 1060 component of  $\theta_0^{(1)} \in \mathbb{R}^{n_1}$ . By Jensen's inequality and independence of the parameters and  $x_1, \dots, x_n$ :

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

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

$$\leq \sqrt{n} \|\Phi\|_\infty. \quad (\text{E.52})$$

1061 As for the fourth moment,

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

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

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

$$\leq 4n^2 \|\Phi\|_\infty^4. \quad (\text{E.56})$$

1062 Triangular inequality finishes the proof.  $\square$

## 1063 E.2 Approximation of the network by linearization

1064 In this subsection we prove the results involved in the proof of Proposition 3.6.

1065 With a slight abuse of notation, we will denote by  $\|x - \mathcal{X}\|$  the positive quantity  $\sup_{1 \leq i \leq n} \|x - x_i\|$ .  
 1066 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  $\nabla_A f = (\frac{\partial}{\partial A_{ij}})_{\substack{1 \leq i \leq n \\ 1 \leq j \leq m}}$ .

1067 We will consider the *linearized gradient flow*, given by

$$\dot{\bar{\theta}}_t = -\nabla_{\theta} f_0(f^{\text{lin}}(\mathcal{X}; \bar{\theta}_t) - y).$$

1068 For this subsection introduce the following notations:  $f_t^{\text{lin}} = f^{\text{lin}}(\mathcal{X}; \bar{\theta}_t)$  and  $\bar{y}_t = f^{\text{lin}}(x; \bar{\theta}_t)$ .

1069 *Proof of Lemma B.12.* Let  $\lambda_{\min}$  be the smallest eigenvalue of  $k_t$ . By gradient flow equations for the  
 1070 parameters  $\theta_t$  and  $\bar{\theta}_t$ :

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

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

1071 On the other hand, Lemma B.2 combined with Cauchy-Schwarz's inequality and the gradient flow  
 1072 equations produces the following system of differential inequalities:

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

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

1073 The previous is a triangular system of differential inequalities of the form

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

1074 with  $B_1 = \frac{1}{\sqrt{n_1}} \|\Phi\|_{\infty} \|f_0 - y\|$  and  $B_0 = \frac{1}{\sqrt{n_1 n_0}} \|\Phi'\|_{\infty} \|f_0 - y\| \|\mathcal{X}_u\|$ .

1075 By integration on  $[0, t]$  and substitution we get,;

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

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

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

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

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

1076 Note that in the last inequality, we used  $\int_0^t I_s(b) ds \leq \frac{I_t(b)^2}{2}$ , for any  $b \geq 0$ .

1077 Thanks to (E.57), the linearised parameters  $\bar{\theta}_t$  also satisfy the preceding inequalities, and hence the  
 1078 thesis holds.  $\square$

1079 *Proof of Lemma B.14.* Let  $\Phi$  denote the CDF of a standard Gaussian variable. For each  $a > 0$ , since  
 1080 the entries of  $\theta_0^{(1)}$  are  $n_1$  i.i.d. standard Gaussian variables,

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

1081 Bernouilli's inequality and standard estimations for Gaussian tails yield

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

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

1082 Let  $r \geq 1$  and put  $a = \sqrt{r\gamma \log n_1}$ . Then:

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

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

$$= 1 - \frac{1}{n_1^{\frac{r\gamma}{2}-1}}. \quad (\text{E.71})$$

1083

□

1084 *Proof of Lemma B.15.* We will write  $f$  for short of  $f_0(x)(\theta)$ . By Lemma B.2 and Cauchy-Schwarz's  
1085 inequality,

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

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

1086 Then the first claim follows by (B.34) and the elementary inequality  $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ , for  $a, b \geq 0$ .

1087 Now we prove the second inequality. Let  $\theta, \tilde{\theta} \in \mathbb{R}^N$ , then,

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

1088 Let us estimate the first summand in the previous expression. By Lemma B.2, for each  $1 \leq u \leq$   
1089  $n_0, 1 \leq v \leq n_1$ ,

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

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

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

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

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

1090 Hence,

$$\left\|\frac{\partial}{\partial \theta^{(0)}} f(\theta) - \frac{\partial}{\partial \theta^{(0)}} f(\tilde{\theta})\right\|^2 \quad (\text{E.80})$$

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

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

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

1091 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.

1092 Similarly, the second summand can be estimated as follows. First compute the partial derivatives by  
1093 using Lemma B.2, for each  $1 \leq v \leq n_1$ :

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

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

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

1094 Therefore,

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

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

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

1095 The preceding estimations, together with  $\frac{\|\theta^{(i)} - \tilde{\theta}^{(i)}\|^2}{\|\theta - \tilde{\theta}\|^2} \leq 1$  for  $i = 0, 1$ , yield:

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

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

1096 Taking the square root in the last inequality yields the thisis.  $\square$

1097 *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  
1098 Markov's inequality and Proposition B.4. There exists a constant  $C > 0$  not depending on  $n_1$  such  
1099 that:

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

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

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

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

1100 Note that Proposition B.4 holds for every natural  $p$ . This concludes the proof.  $\square$

1101 Now are ready to prove Proposition B.9:

1102 *Proof of Proposition B.9.* For the sake of clearness we introduce the following abbreviations for the  
1103 remainder of the proof. Let  $y_t = f(x; \theta_t)$ ,  $\bar{y}_t = f^{\text{lin}}(x; \bar{\theta}_t)$ ,  $f_t = f(\mathcal{X}; \theta_t)$  and  $f_t^{\text{lin}} = f^{\text{lin}}(\mathcal{X}; \bar{\theta}_t)$ .  
1104 Also, let  $k_t = k(\mathcal{X}, \mathcal{X}; \theta_t)$ ,  $\nabla = \nabla_{\theta}$  and let  $L(\mathcal{X})$  denote the Lipschitz constant of  $\nabla f$ , seen as a  
1105 function of  $\theta$ .

1106 Consider the empirical risk for the quadratic loss  $\mathcal{R}_{\mathcal{D}}(\theta_t) = \frac{1}{2} \sum_{i=1}^n (f^{(L)}(x_i; \theta_t) - y)^2$ .

1107 From gradient flow equations we have:

$$\dot{f}_t = -k_t(f_t - y), \quad (\text{E.96})$$

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

1108 Let  $t_* = \inf\{t \mid \|\theta_t - \theta_0\| > \frac{\sigma_{\min}}{2L(\mathcal{X})}\}$  Then for each  $t \leq t_*$ , by 1-Lipschitzianity of the smallest  
 1109 eigenvalue with respect to the operator norm, and by definition of  $t_*$ , we obtain an upper bound for  
 1110  $\lambda_{\min}(k_t)$ :

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

1111 which implies:

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

1112 This estimation combined with Grönwall's inequality applied to (E.97) yield:

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

1113 From (E.97) and Cauchy-Schwarz we deduce:

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

$$\leq -\frac{\sigma_{\min}}{2} \|\nabla f_t(f_t - y)\|. \quad (\text{E.100})$$

1114 Hence,

$$\frac{\partial}{\partial t} \left( \|f_t - y\| + \frac{\sigma_{\min}}{2} \|\theta_t - \theta_0\| \right) \leq \frac{\partial}{\partial t} \|f_t - y\| + \frac{\sigma_{\min}}{2} \|\dot{\theta}_t\| \leq 0. \quad (\text{E.101})$$

1115 for all  $t \leq t_*$ .

1116 Thus, for all  $t \leq t_*$ :

$$\|\theta_t - \theta_0\| \leq \frac{2}{\sigma_{\min}} \|f_0 - y\|. \quad (\text{E.102})$$

1117 Let us show that this property holds for all  $t > 0$ . By contradiction assume  $t_* < \infty$ . (E.102) with  
 1118 Assumption 5 implies

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

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

1119 In particular the last inequality holds for  $t_*$ , which contradicts its definition. Hence  $t_* = \infty$ .

1120 Let us now prove the rest of the inequalities in the theorem.

1121 The gradient flow equation for the linearised network reads:

$$\dot{f}_t^{\text{lin}} = -k_0(f_t^{\text{lin}} - y). \quad (\text{E.105})$$

1122 Define the difference  $r_t = f_t - f_t^{\text{lin}}$ . Then

$$\dot{r}_t = -k_t(f_t - y) + k_0(f_t^{\text{lin}} - y) \quad (\text{E.106})$$

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

1123 Then, by Cauchy-Schwarz and (E.98) combined with (E.105),

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

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

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

1124 Hence,

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

1125 Now let us bound separately the different factors in the previous equation. The norm of the difference  
1126 between the kernels can be estimated as:

$$\|k_t - k_0\| \leq \|\nabla f_t \nabla f_t^\top - \nabla f_0 \nabla f_0^\top\| \quad (\text{E.112})$$

$$\leq 2\|\nabla f_0\| \|\nabla f_t - \nabla f_0\| + \|\nabla f_t - \nabla f_0\|^2 \quad (\text{E.113})$$

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

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

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

1127 where in (E.115) we applied the definition of  $t_*$ .

1128 Moreover, by Grönwall and Cauchy-Schwarz inequalities we have

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

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

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

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

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

$$(\text{E.122})$$

1129 Moreover, by taking the difference of the gradient flow equations for  $\theta_t$  and  $\bar{\theta}_t$  we obtain:

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

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

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

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

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

1130 where in (E.125) we used (E.102), (E.98) and E.121.

1131 Integrating the previous inequality we obtain:

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

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

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

1132 Now we are ready to prove the last inequality in the thesis. Decompose by triangle inequality:

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

1133 First, let us focus on the first summand of (E.131). Denote by  $L(x)$  the Lipschitz constant of  $\nabla y_0$   
 1134 seen as a function of  $\theta$ . Then, by Lemma B.15,

$$\|y_t - f^{\text{lin}}(x; \theta_t)\| = \left\| \int_0^t (\nabla f(x; \theta_s) - \nabla f(x; \theta_0)) \dot{\theta}_s ds \right\| \quad (\text{E.132})$$

$$\leq L(x) \sup_{t \geq 0} \|\theta_t - \theta_0\| \int_0^t \|\dot{\theta}_s\| ds \quad (\text{E.133})$$

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

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

1135 where in the third inequality we used (E.101) and (E.102) on the last one.

1136 As for the second summand of (E.131), by (E.130) and Lemma B.15:

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

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

1137 Combining the two preceding estimations, we obtain the thesis.

1138 □

1139 Lastly, we prove Theorem B.10.

1140 *Proof of Theorem B.10.* We prove the three inequalities separately. Let  $\lambda_{\min}$  denote the smallest  
 1141 eigenvalue of  $k_t$ .

1142 • By Lemma B.2 and Cauchy-Schwarz's inequality,

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



1143  
1144

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:

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

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

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

$$\leq \sum_{u,v} n_1 (\theta_v^{(1)})_0^2 (\theta_{uv}^{(0)})_0^2 + a_1^2 (\theta_{uv}^{(0)})_0^2 \psi(\theta_0)^2 t^2 + \frac{a_0^2 (\theta_v^{(1)})_0^2}{n_1 n_0} \psi(\theta_0)^4 t^4 \quad (\text{E.142})$$

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

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

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

1145  
1146

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,

$$\|y_t - f_t\|^2 \leq \frac{(\text{Lip}\Phi)^2 \|x - \mathcal{X}\|}{n_0 n_1} \|\theta_t^{(0)}\theta_t^{(1)}\|^2 \quad (\text{E.146})$$

$$\leq \frac{A_0}{n_0} \|\theta_0^{(0)}\theta_0^{(1)}\|^2 + \frac{A_1 t^2}{n_0 n_1} \|\theta_0^{(0)}\|^2 \psi(\theta_0)^2 + \frac{A_2 \|\theta_0^{(1)}\|^2 t^4}{n_1^2 n_0} \psi(\theta_0)^4 \quad (\text{E.147})$$

$$+ \frac{A_3 t^6}{n_1^2 n_0} \psi(\theta_0)^6 + \frac{A_4 t^2}{n_1 n_0} \|\theta_0^{(1)}\|^4 \psi(\theta_0)^2 + \frac{A_5 t^4 \|\theta_0^{(1)}\|^2}{n_1^2 n_0} \psi(\theta_0)^4. \quad (\text{E.148})$$

1147

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

1148  
1149

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

$$\|f^{\text{lin}} - \bar{y}_t\|^2 \leq 2\|f_0 - y_0\|^2 + 2\|\nabla_\theta f_0 - \nabla_\theta y_0\|^2 \|\bar{w}_t\|^2. \quad (\text{E.149})$$

1150  
1151

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

$$\|f_0 - y_0\|^2 \leq \frac{(\text{Lip}\Phi)^2 \|x - \mathcal{X}\|^2}{n_1 n_0} \|\theta_0^{(0)}\theta_0^{(1)}\|^2 \leq \frac{A_0}{n_1 n_0} \|\theta_0^{(0)}\theta_0^{(1)}\|^2. \quad (\text{E.150})$$

1152  
1153

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:

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

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

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

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

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

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

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

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

1154

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

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

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

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

1155

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

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

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

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

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

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

1156  
1157  
1158  
1159

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  $B_9 = 2A_0 a_1^2$ .

1160  
1161

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

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

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

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

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

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

1162  
1163

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,

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

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

1164

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

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

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

$$(\text{E.176})$$

1165

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

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

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

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

1166

Analogously,

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

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

1167

Moreover, again by Lemma B.12,

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

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

$$(\text{E.184})$$

1168

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

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

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

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

1169 with  $c_1 = 2\sqrt{2}a_0\|\mathcal{X}\|\|\Phi'\|_\infty$ ,  $c_2 = \sqrt{2}a_0\|\Phi\|_\infty$ ,  $c_3 = 2\sqrt{2}a_0\|\Phi\|_\infty$ ,  $c_4 =$   
 1170  $2\sqrt{2}a'_0\|\mathcal{X}\|\|\Phi'\|_\infty$ ,  $c_5 = \sqrt{2}a'_0\|\Phi\|_\infty$ ,  $c_6 = 2\sqrt{2}a'_0\|\Phi\|_\infty$ ,  $c_7 = 2a_1\|\mathcal{X}\|\|\Phi'\|_\infty$ ,  $c_8 =$   
 1171  $a_1\|\Phi\|_\infty$  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,  
 1172 the reight-hand side of (E.173) can be  $\mathbb{P}$ -almost surely bounded from above with:

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

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

1173 In the previous inequality we used that the event  $\lambda_{\min} = 0$  has null measure. Integrating,  
 1174 and using that  $\Delta(0) = 0$ :

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

$$+ \frac{c_4\|\theta_0^{(1)}\|^2t^2}{2n_0} + \frac{c_5\psi(\theta_0)\|\theta_0^{(1)}\|t^3}{3\sqrt{n_1n_0}} + \frac{c_6\|\theta_0^{(1)}\|t^3}{2\sqrt{n_0}} \quad (\text{E.191})$$

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

1175 Taking the square, applying the elementary inequality  $(\sum_{i=1}^n a_i)^2 \leq n \sum_{i=1}^n a_i^2$ , for  $a_i \geq 0$ ,  
 1176 and adjusting the constants yields the desired result.

1177

□