On Generalization in Bilevel Least Squares with Overparameterized Neural Networks

Fares El Khoury*
Université Grenoble Alpes, Inria,
CNRS, Grenoble INP, LJK,
38000 Grenoble, France

Edouard Pauwels
Toulouse School of Economics,
Université Toulouse Capitole,
31080 Toulouse, France

Samuel Vaiter
CNRS & Université Côte d'Azur,
Laboratoire J. A. Dieudonné,
06108 Nice, France

Michael Arbel
Université Grenoble Alpes, Inria,
CNRS, Grenoble INP, LJK,
38000 Grenoble, France

Abstract

Bilevel optimization addresses machine learning problems in which the outer objective is a *value function* implicitly defined through the solution of an inner problem. While its algorithmic aspects are well developed, its generalization properties remain much less understood. In this paper, we study the generalization of inexact gradient methods when both objectives are least-squares regression problems and when the estimator of the value function's gradient, or *hypergradient*, is obtained by implicit differentiation in function space. This approach allows us to derive stationarity bounds for these methods in terms of the uniform generalization errors of the two regression functions needed to estimate the hypergradient. When these regression functions are defined as overparameterized two-layer ReLU networks trained by gradient flow with *early stopping*, we obtain explicit finite-sample bounds on these generalization errors in the Neural Tangent Kernel (NTK) regime. Combining these with our general inexact gradient analysis shows the effect of sample size on the overall convergence of the bilevel method. Numerical experiments on synthetic data illustrate our results. Beyond the NTK regime, further experiments reveal an *initialization-dependent implicit bias* induced by the bilevel structure, which can lead to *feature learning* at the inner level and overall performance improvements.

1 Introduction

Bilevel optimization is a class of *hierarchical* optimization problems in which the solution to an *outer* problem depends on the solution of an *inner* problem [28]. These methods naturally arise in numerous machine learning applications, including hyperparameter optimization [68, 37, 40], meta-learning [41, 38], neural architecture search [65], and reinforcement learning [51, 74]. The main challenge in bilevel problems is that the outer objective is defined *implicitly* through the solution of the inner problem, making both gradient computation and *generalization* analysis nontrivial [77].

Bilevel optimization poses two intertwined challenges: numerically, one must approximate not only the inner-level solution but also its sensitivity to outer variables in order to compute hypergradients [31, 13, 42]; and statistically, one must understand how the outer solution generalizes as the sample size grows [33, 54, 6]. A large body of work in the well-specified setting with a unique inner level solution addresses the numerical difficulty. Crucially, they typically require the inner objective to be

*Correspondence to: fares.el-khoury@inria.fr.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

strongly convex, which ensures a unique, smoothly varying inner solution and is therefore amenable to stability and convergence analyses [43, 1, 20]. Modern machine learning applications, however, usually employ neural networks as inner models [11, 92, 65]; the resulting inner problems are nonconvex and often overparameterized, so multiple inner solutions can exist and their dependence on the outer variables becomes ambiguous [90, 4]. This dependence can lead to instability, both because inner solutions may be non-unique and because small perturbations in the lower-level problem can induce large changes in the hypergradient [66, 7, 21]. As a consequence, strong generalization guarantees in the nonconvex setting are still largely limited, and the behavior of the resulting solutions on unseen data is still poorly understood. On the statistical side, most generalization results adopt a fixed finite-dimensional parametric viewpoint [29, 52]. This perspective becomes restrictive for overparameterized networks and nonparametric methods, where the effective parameter space (or feature map) can grow with the sample size and the parameters obtained at different sample sizes are not directly comparable. Bridging this gap therefore requires analytical tools that can handle *(i)* nonconvex inner landscapes with possibly multiple solutions arising from overparameterization, and *(ii)* nonparametric function spaces, such as a Reproducing Kernel Hilbert Space (RKHS), in order to obtain meaningful stability and generalization guarantees for bilevel learning in modern settings.

To address these challenges, the *functional bilevel optimization* framework, introduced in [78], provides a principled framework that shifts focus from parameter-centric approaches to a function-space perspective. By formulating the inner problem as an optimization over an infinite-dimensional *Hilbert* space of functions, such as an RKHS [33], this approach naturally accommodates multiple inner solutions, which frequently arise in overparameterized neural networks. A crucial advantage of this functional perspective is that it only requires strong convexity in the function space, rather than with respect to the parameter vector, the latter being rarely satisfied by modern neural networks [64]. This ensures the *uniqueness* of the inner solution for any outer variable, providing a well-defined dependency of the inner problem on the outer parameters and making stability and generalization analyses tractable in practice. In a related line of work, El Khoury et al. [33] provided generalization guarantees for functional bilevel optimization where the considered space is an RKHS. However, these results remain restrictive in the deep learning setting, as they do not accommodate highly expressive models such as overparameterized networks, leaving the question of generalization in bilevel optimization in such settings largely open.

In the present paper, we address this gap by studying *Overparameterized Bilevel Optimization* (OBO) problems in which the lower-level model is *overparameterized*, *i.e.*, it contains more parameters than training samples and thus may have multiple global minimizers for the inner problem [90]. Formally, we consider problems of the form:

$$\min_{\omega \in \Omega} \mathcal{F}(\omega) \coloneqq L_{out}(\omega, h^\star_\omega) \quad \text{such that} \quad h^\star_\omega = \operatorname*{argmin}_{h \in \mathcal{H}} L_{in}(\omega, h), \tag{OBO}$$

where $\Omega \subset \mathbb{R}^d$ is nonempty, compact, and convex, $\mathcal{H}$ is a Hilbert space of functions on $\mathcal{X} \subset \mathbb{R}^p$ taking values in $\mathcal{Y} \subset \mathbb{R}$, and $L_{in}, L_{out} : \Omega \times \mathcal{H} \to \mathbb{R}$ denote the inner and outer objective functions, respectively. Throughout this work, both L_{in} and L_{out} take the form of expected squared losses, a choice motivated by its practical relevance in machine learning [41, 77] and its ubiquity as a theoretical framework [23, 9].

Contributions. We study generalization in *bilevel least squares*, *i.e.*, bilevel optimization problems where both levels are least-squares regressions. This setting is advantageous, as it expresses the hypergradient $\nabla \mathcal{F}$ in terms of two regression functions, the inner solution and the outer regression function. The latter does not depend on the outer variable and can thus be estimated once and for all, which simplifies both the analysis and the algorithmic complexity. Our contributions are as follows.

- **(i) Convergence under inexact hypergradients.** We bound the near-stationarity measure of a projected mini-batch stochastic gradient descent (SGD) method driven by a plug-in hypergradient estimator, in terms of the two estimators' generalization errors (Theorem 2.4).
- **(ii) Generalization rates in the NTK regime.** When these two estimators are two-layer ReLU networks with bias in both layers trained by gradient flow with early stopping, we derive finite-sample generalization bounds in the neural tangent kernel (NTK) regime (Proposition 3.1). These bounds hold in expectation, and uniformly over the outer variable for the inner estimator. To obtain them, we truncate the trained networks, *i.e.*, set them to zero, on a rare event outside of which they provably remain in the NTK regime.

- **(iii) End-to-end sample complexity.** Plugging the rates of **(ii)** into the bound of **(i)** yields the sample complexity of the bilevel method in the NTK regime (Corollary 3.2): the statistical error is governed by the slower of the two regression rates, so the inner and outer sample sizes should be balanced to prevent either level from becoming a bottleneck.
- **(iv) Implicit bias beyond the NTK regime.** Synthetic experiments suggest an initialization-dependent implicit bias of the bilevel structure, which promotes inner-level feature learning and improves performance over the NTK regime.

2 Overparameterized bilevel optimization with squared losses

In this section, we introduce the problem under consideration, derive the expression of its hypergradient, and present the gradient-based algorithm we analyze along with its convergence guarantee.

2.1 Problem formulation

In many machine learning problems with a bilevel structure, the inner and outer objectives L_{in} and L_{out} are typically *expectations of pointwise losses* over some data distributions. In this work, we instantiate both as *squared losses*, a choice that arises widely in regression [46] and causal inference [93, 56], and that allows us to develop our theoretical analysis in a simplified setting. Specifically, we consider:

$$L_{in}(\omega, h) = \frac{1}{2}\mathbb{E}_{(x,z)\sim\mathbb{P}_{XZ}}\left[|h(x) - g_\omega(z)|^2\right], \quad L_{out}(h) = \frac{1}{2}\mathbb{E}_{(x,y)\sim\mathbb{P}_{XY}}\left[|h(x) - y|^2\right], \quad (1)$$

where $g_\omega : \mathcal{Z} \subset \mathbb{R}^r \to \mathbb{R}$ is a model parameterized by ω. Note that we omit the explicit dependence of L_{out} on ω, as it occurs only implicitly through $h^\star_\omega$. The distributions $\mathbb{P}_{XZ}$ and $\mathbb{P}_{XY}$ are the marginals of a joint distribution $\mathbb{P}$ over triplets (x, y, z) in $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, sharing the same marginal $\mathbb{P}_X$ over elements $x \in \mathcal{X}$. A natural space of functions $\mathcal{H}$ in this context is $L^2(\mathcal{X})$, the space of square-integrable functions with respect to $\mathbb{P}_X$, defined on $\mathcal{X}$ and taking values in $\mathcal{Y}$. Provided that $z \mapsto g_\omega(z)$ is square integrable, this problem is guaranteed to admit a *unique* solution to the inner problem, which takes the simple form of a conditional expectation, *i.e.*, $h^\star_\omega(x) = \mathbb{E}_\mathbb{P}[g_\omega(z) \mid x]$, for any $\omega \in \Omega$. In practice, such conditional expectation is only estimated, typically using a neural network, and thus requires careful analysis of the generalization error, which is a central challenge in this study. Typical applications of this problem in machine learning include instrumental variable regression and value function fitting in reinforcement learning, as further discussed in Appendix A. In what follows, $\|\cdot\|$ denotes the Euclidean norm, $\|\cdot\|_{L^2(\mathcal{X})}$ denotes the L^2 norm with respect to $\mathbb{P}_X$, and $\langle\cdot,\cdot\rangle_{L^2(\mathcal{X})}$ denotes the corresponding inner product. For notational simplicity, we will write $\mathbb{P}$ without subscript in the remainder of the paper whenever the meaning can be inferred from context.

Throughout the paper, we assume that the model g_ω is uniformly smooth with bounded values and gradients, that the target space is bounded, that the input space is compact, and that $\mathbb{P}_X$ admits a density $\mathsf{p}_\mathbb{P}$ with respect to the Lebesgue measure on $\mathcal{X}$ that is bounded above and bounded away from zero on some open ball. These assumptions are formalized below.

Assumption A. (Parameterization properties of g_ω) There exists an open convex set $U \supset \Omega$ such that, for all $z \in \mathcal{Z}$, the map $(\omega, z) \mapsto g_\omega(z)$ is differentiable on U with respect to ω and its partial derivative $\partial_\omega g_\omega(z) \in \mathbb{R}^d$ is L-Lipschitz on U. Moreover, there exist constants $G, D > 0$ such that:

$$\sup_{\omega\in U,\, z\in\mathcal{Z}} \|\partial_\omega g_\omega(z)\| \le G \quad \text{and} \quad \sup_{\omega\in U,\, z\in\mathcal{Z}} |g_\omega(z)| \le D.$$

Assumption B. (Boundedness of $\mathcal{Y}$) The target space $\mathcal{Y} \subset \mathbb{R}$ is bounded, *i.e.*, there exists a constant $B > 0$ such that $|y| \le B$ for any $y \in \mathcal{Y}$.
Assumption C. (Compactness of $\mathcal{X}$) The input space $\mathcal{X} \subset \mathbb{R}^p$ is compact with a constant $A > 0$ such that $\|x\| \le A$ for any $x \in \mathcal{X}$.
Assumption D. (Bounded density) There exists a constant $M_\mathbb{P} > 0$ such that $\mathsf{p}_\mathbb{P} \le M_\mathbb{P}$. Moreover, $\mathsf{p}_\mathbb{P}$ is bounded below by a positive constant on some open ball contained in $\mathcal{X}$.

These are relatively mild conditions introduced mainly to simplify the analysis as further discussed in Appendix B. Additional assumptions will be introduced in later sections whenever needed.

2.2 Optimization algorithm for solving (OBO)

We aim to solve (OBO) with a gradient-based algorithm, driven by the *hypergradient* $\nabla\mathcal{F}$. The following result establishes the differentiability of $\mathcal{F}$ and provides an explicit expression for $\nabla\mathcal{F}$, by applying the functional implicit differentiation theorem [78, Proposition 2.2].

Proposition 2.1 (Hypergradient of $\mathcal{F}$)**.** *Under Assumptions A and B, $\mathcal{F}$ is differentiable on U, and its hypergradient is given, for any $\omega \in U$, by:*

$$\nabla\mathcal{F}(\omega) = -\mathbb{E}_{\mathbb{P}}\left[\partial_\omega g_\omega(z) a^\star_\omega(x)\right] \in \mathbb{R}^d, \tag{2}$$

where $a^\star_\omega = -h^\star_\omega + b^\star \in L^2(\mathcal{X})$ is the adjoint function, $h^\star_\omega : x \mapsto \mathbb{E}_{\mathbb{P}}[g_\omega(z) \mid x]$ is the inner solution, and $b^\star : x \mapsto \mathbb{E}_{\mathbb{P}}[y \mid x]$ is the regression function, defined as the minimizer of L_{out} over $L^2(\mathcal{X})$. Moreover, $\mathcal{F}$ is $\mathcal{L}$-smooth on U, in particular on Ω, with $\mathcal{L} := G^2 + L(D+B)$ where the constants G, L, D and B are defined in Assumptions A and B.

While the expression in Equation (2) is exact, it requires full knowledge of $\mathbb{P}$, $h^\star_\omega$ and $b^\star$, which are unavailable in practice, making it necessary to estimate $\nabla\mathcal{F}(\omega)$ from data. Its structure is particularly appealing: since $\partial_\omega g_\omega(z)$ is known, the statistical difficulty reduces entirely to approximating the adjoint function $a^\star_\omega$, which itself decomposes into two regression functions $b^\star$ and $h^\star_\omega$. This structure is key to our analysis, as it allows us to approximate $h^\star_\omega$ and $b^\star$ independently and combine their estimates to obtain an approximation of the hypergradient using *finite-sample* estimates.

Empirical losses. Estimating $h^\star_\omega$ and $b^\star$ requires a *finite-sample approximation* of (OBO). To this end, we assume access to two datasets $\mathcal{D}_{in} = \{(x_i, z_i)\}_{i=1}^{n_{in}}$ and $\mathcal{D}_{out} = \{(\tilde{x}_j, \tilde{y}_j)\}_{j=1}^{n_{out}}$ consisting of *i.i.d.* samples drawn from the distributions $\mathbb{P}_{XZ}$ and $\mathbb{P}_{XY}$, respectively. This is a standard setting in machine learning, since it leads to *tractable* approximations of (OBO). We therefore consider the corresponding empirical problem, obtained by replacing the population losses L_{in} and L_{out} with their empirical counterparts:

$$\widehat{L}_{in}(\omega, h) = \frac{1}{2n_{in}} \sum_{i=1}^{n_{in}} \left(h(x_i) - g_\omega(z_i)\right)^2, \quad \widehat{L}_{out}(h) = \frac{1}{2n_{out}} \sum_{j=1}^{n_{out}} \left(h(\tilde{x}_j) - \tilde{y}_j\right)^2. \tag{3}$$

These serve as training objectives for the estimators $\hat{h}_\omega$ and $\hat{b}$ of $h^\star_\omega$ and $b^\star$: the former is trained by minimizing $\widehat{L}_{in}(\omega, \cdot)$ for a given ω, while the latter is estimated by minimizing $\widehat{L}_{out}$.

Hypergradient estimator. With $\hat{h}_\omega$ and $\hat{b}$ at hand, it remains to approximate the expectation in Equation (2). Accordingly, we assume access, at each use of the estimator, to a *fresh* mini-batch $\mathcal{B}_g = \{(\underline{x}_l, \underline{z}_l)\}_{l=1}^{n_g}$ of *i.i.d.* samples drawn from $\mathbb{P}_{XZ}$, independent of $\mathcal{D}_{in}$, $\mathcal{D}_{out}$ and of the previously drawn batches. We then employ a *plug-in* estimator for $\nabla\mathcal{F}(\omega)$ obtained by replacing the population expectation in Equation (2) with the empirical average over $\mathcal{B}_g$:

$$\widehat{\nabla\mathcal{F}}(\omega) = -\frac{1}{n_g} \sum_{l=1}^{n_g} \partial_\omega g_\omega(\underline{z}_l)(-\hat{h}_\omega(\underline{x}_l) + \hat{b}(\underline{x}_l)), \tag{4}$$

where we approximate $a^\star_\omega$ by its finite-sample estimate of the form $-\hat{h}_\omega + \hat{b}$ since $a^\star_\omega = -h^\star_\omega + b^\star$.

Algorithm. We use the plug-in estimator $\widehat{\nabla\mathcal{F}}$ defined in Equation (4) to drive a gradient-based bilevel algorithm that approximately solves (OBO). Starting from $\omega_0 \in \Omega$, we first compute $\hat{b}$ from $\mathcal{D}_{out}$ once and for all, since it does not depend on the outer variable. At each outer iteration $k = 0, \dots, K-1$, we then re-estimate $\hat{h}_{\omega_k}$ from $\mathcal{D}_{in}$ at the current outer variable, draw a *fresh* mini-batch $\mathcal{B}_g^k$ of *fixed* size n_g, independent of the previous ones, and update $\omega_{k+1} = \Pi_\Omega(\omega_k - \eta\widehat{\nabla\mathcal{F}}(\omega_k))$, where $\eta > 0$ is the stepsize and Π_Ω is the Euclidean projection onto the feasible set Ω. Only the mini-batches are redrawn along the iterations: $\mathcal{D}_{in}$ and $\mathcal{D}_{out}$ are sampled *once* and reused at every ω_k, so that $\hat{h}_\omega$ is required to be accurate *uniformly* over Ω. Since $\hat{h}_\omega$ and $\hat{b}$ are finite-sample approximations

Algorithm 1 Projected mini-batch biased SGD

Require: $\omega_0 \in \Omega$, η, K, n_g, $\mathcal{D}_{out}$, $\mathcal{D}_{in}$

Estimate $\hat{b}$ on $\mathcal{D}_{out}$
for $k = 0, \dots, K-1$ **do**
  Estimate $\hat{h}_{\omega_k}$ on $\mathcal{D}_{in}$
  Draw a fresh $\mathcal{B}_g^k$ of size n_g
  Form $\widehat{\nabla\mathcal{F}}(\omega_k)$ via Equation (4)
  $\omega_{k+1} \leftarrow \Pi_\Omega\left(\omega_k - \eta\widehat{\nabla\mathcal{F}}(\omega_k)\right)$
end for
return $\omega_{k^\star}$, $k^\star$ uniform in $\{0, \dots, K-1\}$

of $h^\star_\omega$ and $b^\star$, the resulting $\widehat{\nabla \mathcal{F}}$ is a *biased* estimator of $\nabla \mathcal{F}$. As the iterates are constrained to Ω, we refer to this scheme as *projected mini-batch biased stochastic gradient descent (SGD)*.

Remark 2.2 (Full-batch variant). Instead of drawing a fresh mini-batch at each outer iteration, one can draw a single dataset once and use all of it at every iteration. Since the iterates then depend on this dataset, the analysis requires the hypergradient estimate to be accurate simultaneously for all $\omega \in \Omega$, and not only at the current iterate. Such a uniform control can be obtained through a covering argument over Ω, but it requires a dataset substantially larger than a single mini-batch, and thus full-batch gradient evaluations, which are costly. As both approaches require a comparable number of outer iterations, and each iteration processes its own batch, the full-batch variant is computationally *more expensive* than Algorithm 1.

Remark 2.3 (Computational benefit of the squared loss setting). In general functional bilevel problems, the adjoint $a^\star_\omega$ is defined as the minimizer of a quadratic problem depending on ω. Hence, both $a^\star_\omega$ and $h^\star_\omega$ must be re-estimated at each outer iteration [78, 33]. Our setting simplifies this step considerably. Indeed, in the estimator $-\hat{h}_\omega + \hat{b}$ of $a^\star_\omega$, only $\hat{h}_\omega$ depends on ω, while $\hat{b}$ can be estimated *once* and kept fixed throughout the optimization. This reduces the computational cost of the algorithm: each outer iteration involves a single re-estimation of $\hat{h}_{\omega_k}$, n_g evaluations of the summands in Equation (4) and one projection onto Ω, so that K outer iterations require $K+1$ functional estimations instead of $2K$ in the general case. Among these operations, the dominant one is the re-estimation of $\hat{h}_{\omega_k}$, the projection being inexpensive for simple sets such as Euclidean balls or boxes.

2.3 General convergence result for the projected mini-batch biased SGD

Since $\mathcal{F}$ is nonconvex and the iterates are constrained to Ω, we quantify the convergence of the algorithm through the *gradient mapping* [14, Definition 10.5], defined for any $\lambda > 0$ by:

$$\mathcal{G}_\lambda(\omega) \coloneqq \lambda^{-1}\left(\omega - \Pi_\Omega\left(\omega - \lambda \nabla \mathcal{F}(\omega)\right)\right),$$

and adopt $\|\mathcal{G}_\lambda(\omega)\|$ as our near-stationarity measure for (OBO). This quantity measures the displacement produced by a single *exact* projected gradient step at stepsize λ, and vanishes precisely at the stationary points of (OBO). We now state our main convergence result for Algorithm 1 when using general estimators $\hat{h}_\omega$ and $\hat{b}$.

Theorem 2.4 (Convergence guarantee of projected mini-batch biased SGD)**.** *Assume A and B, and fix $\omega_0 \in \Omega$. Let $\Delta_{\mathcal{F}} \coloneqq \mathcal{F}(\omega_0) - \mathcal{F}^\star$ with $\mathcal{F}^\star \coloneqq \min_{\omega\in\Omega} \mathcal{F}(\omega)$ and $\{\omega_k\}_{k\geq 0}$ be generated by $\omega_{k+1} = \Pi_\Omega(\omega_k - \eta \widehat{\nabla \mathcal{F}}(\omega_k))$ with stepsize $0 < \eta \leq \frac{1}{2\mathcal{L}}$, where $\mathcal{L}$ is the smoothness constant of $\mathcal{F}$ provided in Proposition 2.1. For any $K \geq 1$, the iterates $\{\omega_k\}_{k\geq 0}$ satisfy:*

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\left[\left\|\mathcal{G}_{1/\mathcal{L}}(\omega_k)\right\|^2\right] \leq c\left[\frac{\Delta_{\mathcal{F}}}{\eta K} + 4G^2(e_h + e_b) + \frac{4\mathcal{L}G^2\eta}{n_g}\left((D+B)^2 + 2(e_h+e_b)\right)\right],$$

where $c = 18 + 12\sqrt{2}$, $e_h = \mathbb{E}\left[\sup_{\omega\in\Omega} \|\hat{h}_\omega - h^\star_\omega\|^2_{L^2(\mathcal{X})}\right]$, $e_b = \mathbb{E}\left[\|\hat{b} - b^\star\|^2_{L^2(\mathcal{X})}\right]$, and B, D, G are the constants of Assumptions A and B. The expectation is over all sources of randomness, namely the samples $\mathcal{D}_{in}$ and $\mathcal{D}_{out}$ used to construct $\hat{h}_\omega$ and $\hat{b}$, and the mini-batches $\{\mathcal{B}^k_g\}_{k\geq 0}$.

The proof of Theorem 2.4 is given in Appendix C. It proceeds by controlling the bias and the variance of the hypergradient estimator in terms of e_h and e_b (Lemmas C.2 and C.3), combining these bounds with a one-step descent inequality along the projected update (Lemma C.4), and telescoping over the iterations. The descent inequality is established for the *Moreau envelope* of $\mathcal{F}$, whose gradient is equivalent to $\mathcal{G}_{1/\mathcal{L}}$ up to absolute constants in our case [30, Section 2.2]. Note that the same bound holds for $\mathbb{E}[\|\mathcal{G}_{1/\mathcal{L}}(\omega_{k^\star})\|^2]$, where $k^\star$ is drawn uniformly at random in $\{0, \dots, K-1\}$, so that the guarantee applies to an iterate Algorithm 1 can return.

Error decomposition. The bound decomposes into three terms of distinct origins. The first is the usual optimization error, and vanishes as the number of iterations K grows. The second is a *bias* term: it quantifies the two-stage error incurred by replacing the adjoint $a^\star_\omega$ with $-\hat{h}_\omega + \hat{b}$, and is unaffected by K, η and n_g. The third is a *variance* term, reflecting the sampling noise of the empirical average over the mini-batch. For a *fixed* batch size n_g, it is proportional to η and can therefore be reduced by decreasing the stepsize. The stepsize thus trades off the first and third terms, while leaving the bias

unchanged. Choosing η achieving the best trade-off makes the algorithmic part of the error decay at the rate $\sqrt{\mathcal{L}\Delta_{\mathcal{F}}/(Kn_g)}$ (Corollary C.6).

Towards explicit rates for e_h and e_b. Theorem 2.4 is stated in terms of the generalization errors e_h and e_b alone, and is therefore agnostic to how $\hat{h}_\omega$ and $\hat{b}$ are obtained. These errors are *statistical* in nature and decrease only as the two regression estimates improve, through larger sample sizes n_{in} and n_{out} or hypothesis spaces better suited to $h^\star_\omega$ and $b^\star$. We make them explicit for overparameterized two-layer ReLU neural networks trained by gradient flow, whose dynamics we describe in Equation (6) and whose generalization bounds we derive in the NTK regime in Proposition 3.1.

3 Explicit statistical rate for Algorithm 1 in the NTK regime

In this section, we build estimators for both targets as truncated neural networks trained by gradient flow, derive their finite-sample generalization rates in the NTK regime, and deduce the sample complexity of Algorithm 1.

3.1 Neural network estimators: training dynamics and truncation

Architecture. We construct estimators $\hat{h}_\omega$ and $\hat{b}$ of the target functions $h^\star_\omega$ and $b^\star$ using *overparameterized* neural networks with a common architecture: two-layer ReLU networks of the form:

$$f_\theta(x) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} v_i \, \sigma(\boldsymbol{u}_i^\top x + c_i) + a, \tag{5}$$

where $\sigma(z) = \max(0, z)$ is the ReLU activation function, m is the network's (even) width, and $\theta = (\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{c}, a)$ collects the parameters. The first-layer weights are $\boldsymbol{u} \in \mathbb{R}^{m \times p}$ with corresponding biases $\boldsymbol{c} \in \mathbb{R}^m$, the second-layer weights are $\boldsymbol{v} \in \mathbb{R}^m$, and $a \in \mathbb{R}$ is the output bias. The width m and the parameters θ above are generic: we write m_h and θ^h for the width and parameters of the network approximating $h^\star_\omega$, and m_b and θ^b for those of the network approximating $b^\star$. The scaling $1/\sqrt{m}$ allows reaching the NTK regime [53] as m grows and under the initialization discussed next.

Initialization. We set the initial output bias at $a = 0$, while the rest of the parameters of both networks are drawn, independently of each other and of the data, from a standard Gaussian distribution using a *mirrored initialization* [58, 62], to ensure the output is exactly zero initially. Specifically, for $1 \le i \le m/2$, we draw $\boldsymbol{u}_i$, c_i and v_i independently from a standard Gaussian and set $\boldsymbol{u}_{i+m/2} = \boldsymbol{u}_i$, $c_{i+m/2} = c_i$ and $v_{i+m/2} = -v_i$, so that the contributions of neurons i and $i + m/2$ cancel. For the network approximating $h^\star_\omega$, the parameters are drawn once and shared by all $\omega \in \Omega$, so that their dependence on ω arises only through training.

Gradient flow training. While neural networks are commonly trained with discrete-time gradient-based algorithms, we consider here their continuous-time limit, *gradient flow*, for theoretical tractability [35]. Starting from the mirrored initialization, and for a given $\omega \in \Omega$, the parameters of both networks evolve in time according to the following ordinary differential equations:

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta^h_{t,\omega} = -\partial_\theta \widehat{\mathcal{L}}_{in}(\omega, \theta^h_t), \qquad \frac{\mathrm{d}}{\mathrm{d}t}\theta^b_t = -\partial_\theta \widehat{\mathcal{L}}_{out}(\theta^b_t), \tag{6}$$

where we defined the training objectives in terms of the network parameters: $\widehat{\mathcal{L}}_{in}(\omega, \theta^h) := \widehat{L}_{in}(\omega, f_{\theta^h})$ and $\widehat{\mathcal{L}}_{out}(\theta^b) := \widehat{L}_{out}(f_{\theta^b})$, while ∂_θ denotes the gradient with respect to the network parameters. We next introduce the NTK, which is needed later to define our final estimators. As shown in [53], the evolution of the predictor functions $f_{\theta^h_t}$ and $f_{\theta^b_t}$ is governed by the *finite-width* NTKs, defined for any $x, x' \in \mathcal{X}$ by:

$$K_t^{m_h}(x, x') = \left(\partial_\theta f_{\theta^h_t}(x)\right)^\top \partial_\theta f_{\theta^h_t}(x'), \qquad K_t^{m_b}(x, x') = \left(\partial_\theta f_{\theta^b_t}(x)\right)^\top \partial_\theta f_{\theta^b_t}(x').$$

As the widths m_h and m_b increase, the network's scaling and initialization choices ensure that both kernels $K_t^{m_h}$ and $K_t^{m_b}$ converge to a common deterministic, time-invariant *infinite-width* limit [53, 4, 59], denoted here by K^{NTK}, whose explicit expression is given in Equation (38) of Appendix E.2. Consequently, the predictor's dynamics converge to those of a kernel regression trained by gradient flow, which we refer to as *NTK regression*.

Early stopping. Since we consider an overparameterized setting in which the network's width can greatly exceed the number of training samples, the predictors $f_{\theta_t^h}$ and $f_{\theta_t^b}$ may interpolate the training data as $t \to \infty$. Although interpolation can still yield good generalization performance [47, 16], precisely characterizing generalization in this regime is beyond the scope of this work. Instead, we adopt an *early stopping* strategy, halting the gradient flows at carefully chosen stopping times t_{op_h} and t_{op_b}, which allows us to leverage techniques similar to those developed in [62]. Early stopping acts as an *implicit regularizer* [82, 95], with the stopping time playing the role of a regularization parameter that controls the bias-variance trade-off and thereby improves generalization.

Truncated estimators. We could, in principle, use the early-stopped predictors as our final estimators. However, obtaining an explicit finite-sample convergence guarantee from Theorem 2.4 requires bounds on their expected generalization errors. Existing results, such as [62, Proposition 13], are not directly applicable: they provide high-probability bounds, holding with probability at least $1-\delta$, under width requirements that depend on the training samples and on the confidence level δ, preventing the usual integration over δ to obtain an expectation bound for a fixed-width network. Instead, one would need to *(i)* address the dependence of the required width on the training data and *(ii)* control rare events on which the early-stopped predictors fall outside the high-probability guarantee and may potentially take arbitrarily large values. The former is addressed through a finer high-probability bound in Proposition E.22, which makes the required width a function of the sample size and δ only, while the latter is addressed by instead considering truncated estimators:

$$\hat{h}_\omega = f_{\theta_{t_{\mathrm{op}_h}}} \mathbb{1}_{\mathcal{E}^h}, \qquad \hat{b} = f_{\theta_{t_{\mathrm{op}_b}}} \mathbb{1}_{\mathcal{E}^b}, \tag{7}$$

where $\mathcal{E}^h$ and $\mathcal{E}^b$ depend only on the training data and initialization. These events are constructed to occur with high probability and ensure that, whenever they occur, the early-stopped predictors satisfy the required probabilistic bounds. The truncated estimators therefore coincide with the early-stopped networks on truncation events, inheriting their high-probability guarantees, and are set to zero on the low-probability complements, where no precise control over the early-stopped networks is available.

Truncation events. We construct each event $\mathcal{E}^h$ and $\mathcal{E}^b$ as the intersection of four explicit high-probability events depending on the data and initialization used to estimate $h_\omega^\star$ and $b^\star$. Since these events have the same structure for both estimators, we drop the superscript h/b for simplicity and state them using the notation of Equation (5), for a width-m network trained on inputs $\{x_i\}_{i=1}^n$. The first three events depend only on the random initialization θ_0, whereas the fourth depends only on the training inputs. For conciseness, we provide only a high-level description here and refer the reader to Appendix E.8.2 for the detailed definitions and discussion.

- **Small initial weights event** $(\mathcal{E}_{\theta_0})$**:** the event when initial weights satisfy: $\|\theta_0\|_\infty \le \sqrt{3 \ln m}$.
- **Initial pre-activation margin event** $(\mathcal{R})$**:** the event when, uniformly over the input domain, only a small fraction $\mathcal{O}(m^\kappa)$ of initial pre-activations are within a prescribed margin of 0, for $\kappa < 1$.
- **Initial NTK concentration event** $(\mathcal{C})$**:** the event when, uniformly over a grid on the input domain, the finite-width NTK at initialization is within order $\sqrt{\ln m / m}$ of its infinite-width limit.
- **NTK non-degeneracy event** $(\mathcal{E}_{\lambda_0})$**:** the event when the Gram matrix of the infinite-width NTK on training data has its smallest eigenvalue λ_0 bounded below by a prescribed quantity $\lambda_\star(n,\delta) > 0$.

Under $(\mathcal{E}_{\theta_0})$, it is possible to show that the gradients remain bounded along training, which helps keep the parameters in a controlled neighborhood of θ_0. Under $(\mathcal{R})$, it can be shown that the signs of their pre-activations cannot change unless the parameters move far. This property is essential in proving that wide neural networks trained in the NTK regime remain close to their initialization [53, 4]. $(\mathcal{C})$ is expected for large m, since the finite-width NTK is an average over the m neurons of terms that are random at initialization and whose mean is K^{NTK}. Finally, $(\mathcal{E}_{\lambda_0})$ guarantees that the gradient flow of the NTK regression converges on the training data.

3.2 Finite-sample generalization bounds

To state our generalization rates, we need a *source condition*, *i.e.*, a notion of regularity for the target functions $h_\omega^\star$ and $b^\star$. A natural way to quantify this regularity is through how well these targets can be approximated by elements of the RKHS $\mathcal{H}^{\mathrm{NTK}}$ associated with the limiting kernel K^{NTK}. Indeed, in the infinite-width limit, the predictors used to approximate the targets are elements of $\mathcal{H}^{\mathrm{NTK}}$. Rather than requiring these targets to also belong to $\mathcal{H}^{\mathrm{NTK}}$, we will instead assume they lie in an *interpolation space* between $L^2(\mathcal{X})$ and $\mathcal{H}^{\mathrm{NTK}}$, as this allows capturing intermediate regularity

properties. To formally define these spaces, we introduce the integral operator $\mathcal{T} : L^2(\mathcal{X}) \to L^2(\mathcal{X})$ associated with K^{NTK}, given for every $f \in L^2(\mathcal{X})$ and $x \in \mathcal{X}$ by:

$$(\mathcal{T}f)(x) = \mathbb{E}_{X\sim\mathbb{P}_X}\left[K^{\mathrm{NTK}}(x, X)f(X)\right].$$

Since K^{NTK} is a Mercer kernel, $\mathcal{T}$ is compact, self-adjoint, and positive semi-definite. Hence, by the spectral theorem [86, Theorem A.5.13], $\mathcal{T}$ admits a spectral decomposition with non-increasing eigenvalues $(\lambda_i)_{i\geq 1}$ and corresponding orthonormal eigenfunctions $(e_i)_{i\geq 1}$ forming a basis of $L^2(\mathcal{X})$. We then define the interpolation space $[\mathcal{H}^{\mathrm{NTK}}]^s$ with parameter $s \geq 0$ as the subspace of functions $f \in L^2(\mathcal{X})$ such that $\|f\|_{[\mathcal{H}^{\mathrm{NTK}}]^s} < \infty$, where $\|\cdot\|_{[\mathcal{H}^{\mathrm{NTK}}]^s}$ is the interpolation norm defined as the sum of the squared projections onto each e_i, weighted by λ_i^{-s}, *i.e.*, $\|f\|^2_{[\mathcal{H}^{\mathrm{NTK}}]^s} = \sum_{i\geq 1} \lambda_i^{-s}\langle f, e_i\rangle^2_{L^2(\mathcal{X})}$[2]. We now give the source conditions needed to establish generalization bounds.

Assumption E. (Source condition for $h^\star_\omega$) There exist $s_h, R_h > 0$ such that $\|h^\star_\omega\|_{[\mathcal{H}^{\mathrm{NTK}}]^{s_h}} \leq R_h$ for every $\omega \in \Omega$, hence in particular at every iterate of Algorithm 1.

Assumption F. (Source condition for $b^\star$) There exist $s_b, R_b > 0$ such that $\|b^\star\|_{[\mathcal{H}^{\mathrm{NTK}}]^{s_b}} \leq R_b$.

Assumptions E and F are standard source conditions in kernel regression [63, 23, 39], with s_h and s_b controlling the regularity of $h^\star_\omega$ and $b^\star$ relative to $\mathcal{H}^{\mathrm{NTK}}$: larger values correspond to smoother, easier to learn targets. In addition, Assumption E requires this regularity bound to hold uniformly over ω, ensuring that the difficulty of estimating $h^\star_\omega$ remains controlled throughout the optimization. When $\mathcal{X}$ is a bounded Lipschitz domain, such as a ball or a cube, we show in Appendix G that $[\mathcal{H}^{\mathrm{NTK}}]^s$ coincides with the Sobolev space $H^{\frac{s(p+1)}{2}}(\mathcal{X})$ for $0 \leq s \leq 1$. Thus, when $s_h, s_b \leq 1$, Assumptions E and F can be expressed as Sobolev regularity conditions, uniformly over $\omega \in \Omega$ for $h^\star_\omega$.

We can now derive generalization bounds for the estimators $\hat{h}_\omega$ and $\hat{b}$ from Section 3.1.

Proposition 3.1 (Expected generalization bounds)**.** *Assume $p \geq 2$ and let $\Omega \subset \mathbb{R}^d$ be compact. Let Assumptions A to F hold with $s_h, s_b \in (\frac{1}{p+1}, 2]$, and choose the stopping times $t_{\mathrm{op}_h} = n_{in}^{\gamma_h}$ and $t_{\mathrm{op}_b} = n_{out}^{\gamma_b}$, with $\gamma_h = \frac{p+1}{s_h(p+1)+p}$ and $\gamma_b = \frac{p+1}{s_b(p+1)+p}$. There exist two polynomials poly_h and poly_b such that, when n_{in} and n_{out} are sufficiently large and the widths satisfy $m_h \geq \mathrm{poly}_h(n_{in}, \ln n_{in})$ and $m_b \geq \mathrm{poly}_b(n_{out}, \ln n_{out})$, the truncated predictors of Equation (7) satisfy:*

$$\mathbb{E}\left[\sup_{\omega\in\Omega}\left\|\hat{h}_\omega - h^\star_\omega\right\|^2_{L^2(\mathcal{X})}\right] \lesssim d^4(\ln n_{in})^4 n_{in}^{-s_h\gamma_h}, \qquad \mathbb{E}\left[\left\|\hat{b} - b^\star\right\|^2_{L^2(\mathcal{X})}\right] \lesssim (\ln n_{out})^4 n_{out}^{-s_b\gamma_b},$$

where $\lesssim$ hides constants independent of n_{in}, n_{out} and d, and the expectations are taken over the training samples and the random initialization.

Proposition 3.1 shows that the truncated predictors achieve the minimax-optimal rates of the corresponding regression problems, up to logarithmic factors [19, 96]. We defer the detailed proof, including the explicit dependence of the constants and polynomial coefficients on the problem parameters, to Appendices D.2 and E.8.2. The proof first shows that, with high probability, sufficiently wide networks remain uniformly close to their NTK regression counterparts, and then relies on truncation to turn these high-probability guarantees into bounds in expectation. Compared with existing results in the NTK literature [62, 58, 24], the main challenge is to obtain a bound that holds uniformly over $\omega \in \Omega$. This requires deriving bounds with constants independent of ω and combining a covering argument over Ω with the Lipschitz continuity of the NTK regression error in ω (Lemma D.3).

The above bounds are precisely the quantities e_h and e_b arising in Theorem 2.4. Thus combining both the convergence result in Theorem 2.4 with the generalization bounds in Proposition 3.1 yields the following explicit convergence result for the iterates produced by Algorithm 1 when using the truncated estimators in Equation (7).

Corollary 3.2 (Sample complexity of Algorithm 1)**.** *Let Assumptions A to F hold and consider the iterates ω_k obtained from Algorithm 1 using stepsize $0 < \eta \leq \frac{1}{2\mathcal{L}}$, gradient estimator in Equation (4) and truncated predictors in Equation (7) with width parameters, sample sizes and stopping times as in Proposition 3.1. Then, for any $K \geq 1$, the iterates $\{\omega_k\}_{k\geq 0}$ satisfy:*

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\mathcal{G}_{1/\mathcal{L}}(\omega_k)\right\|^2\right] \lesssim \frac{\Delta_{\mathcal{F}}}{\eta K} + \frac{\mathcal{L}G^2\eta(D+B)^2}{n_g} + G^2\left(d^4(\ln n_{in})^4 n_{in}^{-s_h\gamma_h} + (\ln n_{out})^4 n_{out}^{-s_b\gamma_b}\right),$$

[2] In our setting, the norm $\|f\|_{[\mathcal{H}^{\mathrm{NTK}}]^s}$ is equivalent to $\left(\sum_{i\geq 1} i^{s(p+1)/p}\langle f, e_i\rangle^2_{L^2(\mathcal{X})}\right)^{1/2}$, since the eigenvalues decay at a rate exactly equal to $\frac{p+1}{p}$ (Lemma E.12).

where $\lesssim$ hides constants which are independent of K, η, n_g, n_{in}, n_{out} and d.

In Corollary 3.2, the statistical error is dominated by the slower of the two rates $n_{in}^{-s_h\gamma_h}$ and $n_{out}^{-s_b\gamma_b}$. This suggests balancing the two sample sizes so that $n_{in}^{-s_h\gamma_h} \approx n_{out}^{-s_b\gamma_b}$ up to constant and logarithmic factors, ensuring that neither level becomes a bottleneck. Moreover, since the exponents $s_h\gamma_h$ and $s_b\gamma_b$ increase with s_h and s_b, respectively, the level whose target has the larger smoothness parameter needs fewer samples to reach the same accuracy.

4 Related work

El Khoury et al. [33] and Chen et al. [25] study generalization in bilevel nonparametric settings: the former uses *full-batch gradient descent* to optimize the inner level over an RKHS, while the latter relies on *mean-field Langevin dynamics* for two-stage least squares. Both works assume a *fixed* regularization parameter to control the bias-variance trade-off, leaving open the question of how the optimal choice of the regularization, as a function of the sample sizes, would affect sample complexity and whether it could lead to faster rates. In contrast, in our setting, this trade-off is achieved by the early stopping time, which we select as a function of the sample sizes, yielding explicit finite-sample bounds that quantify the sample complexity of the problem. Relatedly, Patel and Vlatakis-Gkaragkounis [76] provide convergence guarantees for training overparameterized neural networks in *min-max (zero-sum) games* in the NTK regime, which can be viewed as a particular bilevel structure. Our work differs in that we study bilevel learning and generalization, and quantify how stochastic, approximate NTK training affects hypergradient accuracy and excess risk rather than game-theoretic optimization dynamics. Also, Kim et al. [56] analyze the deep feature instrumental variable regression, which shares structural similarities with our bilevel setup. They establish minimax optimal convergence rates when the target function belongs to a Besov space. However, their analysis focuses on statistical aspects, whereas we also account for optimization. Finally, El Khoury et al. [34] study the same problem as ours and *debias* the hypergradient to attain *parametric* rates, whereas we derive generalization bounds for the plug-in estimator with neural networks in the NTK regime.

5 Numerical experiments

5.1 Empirical validation of the theory in the NTK regime on synthetic data

We empirically validate the convergence bound of Corollary 3.2 on targets whose smoothness is known in closed form. We consider inputs $x \in [0,1]^2$, outer parameters $\omega \in \mathbb{R}^4$, and cusp targets $h^\star_\omega(x) = \psi(x)^\top\omega$ and $b^\star(x) = \|x-x_0\|^{q_b}$, where $\psi(x) := [\|x-c_1\|^{q_h}, \dots, \|x-c_4\|^{q_h}]^\top$, the centers $c_1, \dots, c_4$ form the grid $\{\frac{1}{4}, \frac{3}{4}\}^2$, and $x_0 = (\frac{1}{2}, \frac{1}{2})$. By [84, Section 2.3.1, Lemma 1], $h^\star_\omega$ and $b^\star$ satisfy the source conditions uniformly over ω for every $s_h < \bar{s}_h := (2q_h+2)/3$ and $s_b < \bar{s}_b := (2q_b+2)/3$, with $q_h, q_b \in \{0.05, 0.25, 0.45\}$, giving $\bar{s}_h, \bar{s}_b \in \{0.70, 0.83, 0.97\}$ and decay rates of 0.51, 0.56, and 0.59, respectively. Both targets are approximated by networks of the form in Equation (5) trained by full-batch gradient descent. We follow Algorithm 1 with hypergradients evaluated on fresh batches of size 200. The sample sizes n_{in} and n_{out} vary independently from 100 to 25,600. Further details about the setup are given in Appendix J.1. Figure 1 plots the near-stationarity measure averaged over the outer iterations against n_{in} or n_{out}, minus a constant C collecting the terms of the bound in Corollary 3.2 that are *independent* of the varied sample size. In panel (a), n_{in} varies for $s_h \in \{0.70, 0.83, 0.97\}$, with $n_{out} = 25{,}600$ and $s_b = 0.70$; panel (b) swaps the inner and outer roles. The curves follow the slopes predicted by Corollary 3.2, thus validating our theory.

5.2 Insights into feature learning beyond the NTK regime

Our analysis relies on the NTK regime, which assumes fixed features and thus precludes *feature learning*. For a single network, the initialization scale governs the transition between the NTK and feature learning regimes [91]. In the bilevel setting, the inner and outer initializations may *jointly* determine the regime. We investigate this insight, which lies beyond the scope of our theory, on a synthetic problem in which $h^\star_\omega$ and $b^\star$ are estimated by two-layer networks of the form in Equation (5) normalized by m instead of $\sqrt{m}$ to allow feature learning [26], with first-layer initialization variances σ_0^2 and τ_0^2, respectively. Figure 2 shows that both the test inner and outer losses are lowest when both initializations are small (feature learning) and degrade as either grows (NTK), indicating a *joint effect*

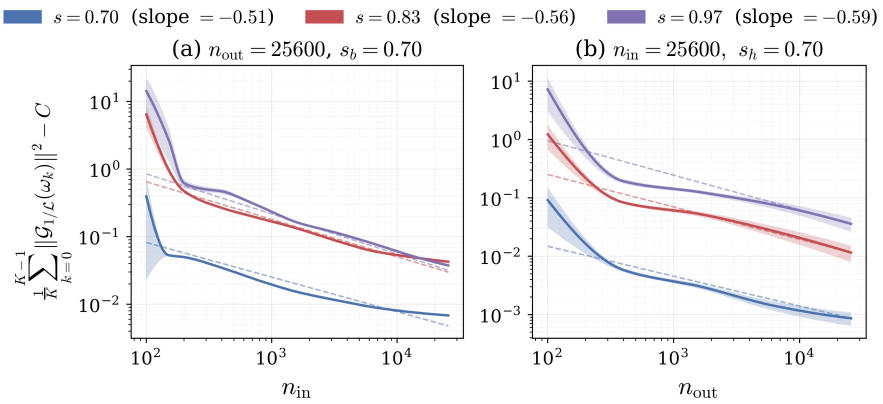


Figure 1: Illustration of Algorithm 1 on synthetic data with targets of known smoothness. Dashed lines show the theoretical slopes. The curves are averaged over 5 runs and displayed on a log scale. Solid lines represent the mean across runs, and shaded regions indicate the 95% confidence intervals.

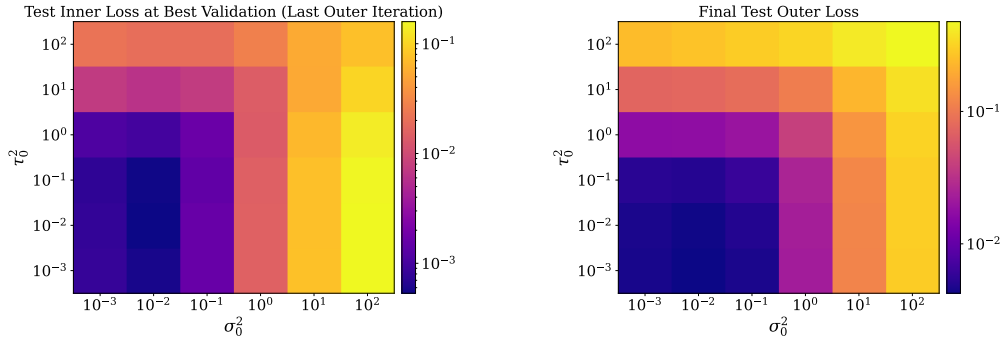


Figure 2: Test inner (left) and outer (right) losses at the last outer iteration as a function of (σ_0^2, τ_0^2): both losses are lowest when both initializations are small (feature learning) and degrade as either grows large (NTK). The heatmaps are averaged over 5 runs and displayed on a log scale.

of the two initializations on the effective regime. The full setup and additional results, showing that this ranking holds along the outer iterations and across sample sizes, are given in Appendix J.2.

6 Conclusion and perspectives

Summary. In this paper, we provided a general convergence guarantee for inexact bilevel gradient methods when both objectives are least-squares regressions. The resulting bound depends on the estimators of $h^\star_\omega$ and $b^\star$ only through their uniform generalization errors. In the NTK regime, we derived finite-sample bounds on these errors when both are estimated by overparameterized two-layer ReLU networks trained by gradient flow with early stopping. Plugging these rates into the general bound yields an explicit convergence guarantee for Algorithm 1. Our experimental results suggest that, beyond the NTK regime, the bilevel formulation introduces an implicit bias that promotes feature learning and leads to improved performance.

Limitations and future work. Our theoretical analysis does not address minimax optimality, and establishing minimax rates for (OBO) remains an open problem. While the transition between the NTK and feature learning regimes is well studied for single networks [91, 26, 94], its theoretical characterization in the bilevel setting is still unexplored, which could perhaps be approached through the lens of *grokking* [79]. Indeed, this phenomenon has been interpreted as a transition between the NTK and feature learning regimes [67, 57, 69], potentially offering an intermediate trade-off. Further directions include extending our analysis to deeper architectures, relaxing the polynomial width requirement typical of the NTK regime, and considering general losses, for which the adjoint function no longer reduces to two independently estimable regression functions.

Acknowledgments and Disclosure of Funding

FEK would like to thank Elie Louis for helpful discussions related to this paper. This work was supported by the ANR project BONSAI (grant ANR-23-CE23-0012-01). EP acknowledges funding from the AI Interdisciplinary Institute ANITI through the French Investments for the Future – PIA3 program under grant agreement ANR-19-PI3A-0004; the Air Force Office of Scientific Research, Air Force Materiel Command, USAF under grant number FA8655-22-1-7012; TSE-P; the ANR project Chess (grant ANR-17-EURE-0010); the ANR project Regulia; and the Institut Universitaire de France (IUF). EP and SV acknowledge support from the ANR project MAD. SV also thanks the PEPR PDE-AI program (ANR-23-PEIA-0004) and the 3IA Côte d'Azur Chair BOGL.

References

[1] P. Ablin, G. Peyré, and T. Moreau. Super-efficiency of automatic differentiation for functions defined as a minimum. In *International Conference on Machine Learning (ICML)*, 2020.

[2] R. A. Adams and J. J. Fournier. *Sobolev Spaces*, volume 140. Elsevier, 2003.

[3] M. S. Agranovich. *Sobolev Spaces, Their Generalizations, and Elliptic Problems in Smooth and Lipschitz Domains*. Springer Monographs in Mathematics. Springer, Cham, 2015.

[4] Z. Allen-Zhu, Y. Li, and Z. Song. A Convergence Theory for Deep Learning via Over-Parameterization. In *International Conference on Machine Learning (ICML)*, 2019.

[5] J. D. Angrist and G. W. Imbens. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, 90 (430):431–442, 1995.

[6] M. Arbel and J. Mairal. Amortized Implicit Differentiation for Stochastic Bilevel Optimization. In *International Conference on Learning Representations (ICLR)*, 2022.

[7] M. Arbel and J. Mairal. Non-convex bilevel games with critical point selection maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[8] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

[9] S. Arora, S. S. Du, W. Hu, Z. Li, R. R. Salakhutdinov, and R. Wang. On Exact Computation with an Infinitely Wide Neural Net. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[10] K. Atkinson and W. Han. *Spherical harmonics and approximations on the unit sphere: an introduction. Lecture Notes in Mathematics*, volume 2044. Springer Heidelberg, 2012.

[11] F. Bao, C. Li, K. Xu, H. Su, J. Zhu, and B. Zhang. Bi-level score matching for learning energy-based latent variable models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[12] J. A. Barceló, T. Luque, and S. Pérez-Esteva. Characterization of Sobolev spaces on the sphere. *Journal of Mathematical Analysis and Applications*, 491(1):124240, 2020.

[13] A. G. Baydin, B. A. Pearlmutter, A. A. Radul, and J. M. Siskind. Automatic differentiation in machine learning: a survey. *Journal of Machine Learning Research (JMLR)*, 18(153):1–43, 2018.

[14] A. Beck. *First-order methods in optimization*. SIAM, 2017.

[15] A. Behzadan and M. Holst. Sobolev-Slobodeckij Spaces on Compact Manifolds, Revisited. *Mathematics*, 10(3):522, 2022.

[16] M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116 (32):15849–15854, 2019.

[17] J. Bergh and J. Löfström. *Interpolation Spaces: An Introduction*, volume 223 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, Berlin, 1976.

[18] A. Bietti and J. Mairal. On the inductive bias of neural tangent kernels. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[19] G. Blanchard and N. Mücke. Optimal Rates for Regularization of Statistical Inverse Learning Problems. *Foundations of Computational Mathematics*, 18(4):971–1013, 2018.

[20] M. Blondel, Q. Berthet, M. Cuturi, R. Frostig, S. Hoyer, F. Llinares-López, F. Pedregosa, and J.-P. Vert. Efficient and modular implicit differentiation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[21] J. Bolte, Q.-T. Lê, E. Pauwels, and S. Vaiter. Geometric and computational hardness of bilevel programming. *Mathematical Programming*, 215(1):539–574, 2026.

[22] S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

[23] A. Caponnetto and E. De Vito. Optimal Rates for the Regularized Least-Squares Algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007. doi: 10.1007/s10208-006-0196-8.

[24] G. Chen, Y. Li, and Q. Lin. On the Impacts of the Random Initialization in the Neural Tangent Kernel Theory. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[25] Z. Chen, A. Nitanda, A. Gretton, and T. Suzuki. Towards a Unified Analysis of Neural Networks in Nonparametric Instrumental Variable Regression: Optimization and Generalization. *arXiv preprint arXiv:2511.14710*, 2025.

[26] L. Chizat, E. Oyallon, and F. Bach. On Lazy Training in Differentiable Programming. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[27] Y. Cho and L. Saul. Kernel Methods for Deep Learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

[28] B. Colson, P. Marcotte, and G. Savard. An overview of bilevel optimization. *Annals of Operations Research*, 153:235–256, 2007.

[29] M. Dagréou, P. Ablin, S. Vaiter, and T. Moreau. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[30] D. Davis and D. Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[31] J. Domke. Generic methods for optimization-based modeling. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

[32] D. E. Edmunds and H. Triebel. *Function spaces, entropy numbers, differential operators*, volume 120. Cambridge University Press Cambridge, 1996.

[33] F. El Khoury, É. Pauwels, S. Vaiter, and M. Arbel. Learning Theory for Kernel Bilevel Optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[34] F. El Khoury, H. Zenati, N. Kallus, M. Arbel, and A. Bibaut. Parametric Rates for Nonparametric Conditional-Moment Bilevel Optimization. *arXiv preprint arXiv:2605.21341*, 2026.

[35] O. Elkabetz and N. Cohen. Continuous vs. Discrete Optimization of Deep Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[36] L. C. Evans and R. F. Gariepy. *Measure Theory and Fine Properties of Functions*. CRC Press, Boca Raton, FL, revised edition, 2015. ISBN 9781482242393.

[37] M. Feurer and F. Hutter. Hyperparameter Optimization. In *Automated Machine Learning: Methods, Systems, Challenges*, The Springer Series on Challenges in Machine Learning, pages 3–33. Springer International Publishing, 2019.

[38] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning (ICML)*, 2017.

[39] S. Fischer and I. Steinwart. Sobolev Norm Learning Rates for Regularized Least-Squares Algorithms. *Journal of Machine Learning Research (JMLR)*, 21(205):1–38, 2020.

[40] L. Franceschi, M. Donini, P. Frasconi, and M. Pontil. Forward and reverse gradient-based hyperparameter optimization. In *International Conference on Machine Learning (ICML)*, 2017.

[41] L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi, and M. Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International Conference on Machine Learning (ICML)*, 2018.

[42] S. Ghadimi and M. Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.

[43] S. Gould, B. Fernando, A. Cherian, P. Anderson, R. S. Cruz, and E. Guo. On differentiating parameterized argmin and argmax problems with application to bi-level optimization. *arXiv preprint arXiv:1607.05447*, 2016.

[44] P. Grisvard. *Elliptic Problems in Nonsmooth Domains*, volume 24 of *Monographs and Studies in Mathematics*. Pitman (Advanced Publishing Program), Boston, 1985.

[45] M. Haas, D. Holzmüller, U. von Luxburg, and I. Steinwart. Mind the spikes: Benign overfitting of kernels and neural networks in fixed dimension. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[46] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2009.

[47] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in High-Dimensional Ridgeless Least Squares Interpolation. *Annals of Statistics*, 50(2):949–986, 2022.

[48] D. Hendrycks and K. Gimpel. Gaussian Error Linear Units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

[49] D. P. Hewett and A. Moiola. A note on properties of the restriction operator on Sobolev spaces. *Journal of Applied Analysis*, 23(1):1–8, 2017.

[50] S. Hofmann, M. Mitrea, and M. Taylor. Geometric and transformational properties of Lipschitz domains, Semmes-Kenig-Toro domains, and other classes of finite perimeter domains. *The Journal of Geometric Analysis*, 17(4):593–647, 2007.

[51] M. Hong, H.-T. Wai, Z. Wang, and Z. Yang. A two-timescale stochastic algorithm framework for bilevel optimization: Complexity analysis and application to actor-critic. *SIAM Journal on Optimization*, 33(1):147–180, 2023.

[52] Y. Hu, J. Wang, Y. Xie, A. Krause, and D. Kuhn. Contextual Stochastic Bilevel Optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[53] A. Jacot, F. Gabriel, and C. Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[54] K. Ji, J. Yang, and Y. Liang. Bilevel optimization: Convergence analysis and enhanced design. In *International Conference on Machine Learning (ICML)*, 2021.

[55] K. Karhadkar, M. Murray, and G. Montúfar. Bounds for the smallest eigenvalue of the NTK for arbitrary spherical data of arbitrary dimension. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[56] J. Kim, D. Meunier, A. Gretton, T. Suzuki, and Z. Li. Optimality and Adaptivity of Deep Neural Features for Instrumental Variable Regression. In *International Conference on Learning Representations (ICLR)*, 2025.

[57] T. Kumar, B. Bordelon, S. J. Gershman, and C. Pehlevan. Grokking as the transition from lazy to rich training dynamics. In *International Conference on Learning Representations (ICLR)*, 2024.

[58] J. Lai, M. Xu, R. Chen, and Q. Lin. Generalization Ability of Wide Neural Networks on $\mathbb{R}$. *arXiv preprint arXiv:2302.05933*, 2023.

[59] J. Lee, L. Xiao, S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington. Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[60] J. M. Lee. *Introduction to Smooth Manifolds*. Springer, 2 edition, 2013.

[61] S. Li. Concise Formulas for the Area and Volume of a Hyperspherical Cap. *Asian Journal of Mathematics and Statistics*, pages 66–70, 2011.

[62] Y. Li, Z. Yu, G. Chen, and Q. Lin. On the Eigenvalue Decay Rates of a Class of Neural-Network Related Kernel Functions Defined on General Domains. *Journal of Machine Learning Research (JMLR)*, 25(82):1–47, 2024.

[63] J. Lin, A. Rudi, L. Rosasco, and V. Cevher. Optimal rates for spectral algorithms with least-squares regression over Hilbert spaces. *Applied and Computational Harmonic Analysis*, 48(3): 868–890, 2020.

[64] C. Liu, L. Zhu, and M. Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022.

[65] H. Liu, K. Simonyan, and Y. Yang. DARTS: Differentiable architecture search. In *International Conference on Learning Representations (ICLR)*, 2019.

[66] R. Liu, Y. Liu, S. Zeng, and J. Zhang. Towards gradient-based bilevel optimization with non-convex followers and beyond. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[67] K. Lyu, J. Jin, Z. Li, S. S. Du, J. D. Lee, and W. Hu. Dichotomy of Early and Late Phase Implicit Biases Can Provably Induce Grokking. In *International Conference on Learning Representations (ICLR)*, 2024.

[68] D. Maclaurin, D. Duvenaud, and R. Adams. Gradient-based hyperparameter optimization through reversible learning. In *International Conference on Machine Learning (ICML)*, 2015.

[69] M. A. Mohamadi, Z. Li, L. Wu, and D. J. Sutherland. Why Do You Grok? A Theoretical Analysis on Grokking Modular Addition. In *International Conference on Machine Learning (ICML)*, 2024.

[70] J.-J. Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France*, 93:273–299, 1965.

[71] R. Munos and C. Szepesvári. Finite-Time Bounds for Fitted Value Iteration. *Journal of Machine Learning Research (JMLR)*, 9(27):815–857, 2008.

[72] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Science & Business Media, 2013.

[73] W. K. Newey and J. L. Powell. Instrumental Variable Estimation of Nonparametric Models. *Econometrica*, 71(5):1565–1578, 2003. doi: 10.1111/1468-0262.00459.

[74] E. Nikishin, R. Abachi, R. Agarwal, and P.-L. Bacon. Control-oriented model-based reinforcement learning with implicit differentiation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7886–7894, 2022.

[75] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[76] D. Patel and E.-V. Vlatakis-Gkaragkounis. Solving Neural Min-Max Games: The Role of Architecture, Initialization & Dynamics. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[77] F. Pedregosa. Hyperparameter optimization with approximate gradient. In *International Conference on Machine Learning (ICML)*, 2016.

[78] I. Petrulionytė, J. Mairal, and M. Arbel. Functional Bilevel Optimization for Machine Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[79] A. Power, Y. Burda, H. Edwards, I. Babuschkin, and V. Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

[80] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

[81] P. Ramachandran, B. Zoph, and Q. V. Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

[82] G. Raskutti, M. J. Wainwright, and B. Yu. Early Stopping and Non-Parametric Regression: An Optimal Data-Dependent Stopping Rule. *Journal of Machine Learning Research (JMLR)*, 15 (1):335–366, 2014.

[83] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*, volume 317 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, 1998.

[84] T. Runst and W. Sickel. *Sobolev Spaces of Fractional Order, Nemytskij Operators, and Nonlinear Partial Differential Equations*. De Gruyter, Berlin, New York, 1996. ISBN 9783110812411. doi: 10.1515/9783110812411.

[85] E. M. Stein. *Singular Integrals and Differentiability Properties of Functions*, volume 2. Princeton University Press, 1970.

[86] I. Steinwart and A. Christmann. *Support Vector Machines*. Information Science and Statistics. Springer New York, New York, NY, USA, 2008.

[87] I. Steinwart and C. Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, 35(3):363–417, 2012.

[88] H. Triebel. *Theory of Function Spaces II*, volume 84 of *Monographs in Mathematics*. Birkhäuser, Basel, 1992. ISBN 978-3-7643-2639-5. doi: 10.1007/978-3-0346-0419-2.

[89] R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2 edition, 2026.

[90] P. Vicol, J. P. Lorraine, F. Pedregosa, D. Duvenaud, and R. B. Grosse. On Implicit Bias in Overparameterized Bilevel Optimization. In *International Conference on Machine Learning (ICML)*, 2022.

[91] B. Woodworth, S. Gunasekar, J. D. Lee, E. Moroshko, P. Savarese, I. Golan, D. Soudry, and N. Srebro. Kernel and Rich Regimes in Overparametrized Models. In *Conference on Learning Theory (COLT)*, 2020.

[92] Q. Xiao, H. Yuan, A. Saif, G. Liu, R. Kompella, M. Wang, and T. Chen. A First-order Generative Bilevel Optimization Framework for Diffusion Models. In *International Conference on Machine Learning (ICML)*, 2025.

[93] L. Xu, Y. Chen, S. Srinivasan, N. De Freitas, A. Doucet, and A. Gretton. Learning Deep Features in Instrumental Variable Regression. In *International Conference on Learning Representations (ICLR)*, 2021.

[94] G. Yang and E. J. Hu. Tensor Programs IV: Feature Learning in Infinite-Width Neural Networks. In *International Conference on Machine Learning (ICML)*, 2021.

[95] Y. Yao, L. Rosasco, and A. Caponnetto. On Early Stopping in Gradient Descent Learning. *Constructive Approximation*, 26(2):289–315, 2007.

[96] H. Zhang, Y. Li, and Q. Lin. On the Optimality of Misspecified Spectral Algorithms. *Journal of Machine Learning Research (JMLR)*, 25(188):1–50, 2024.

Appendices

Roadmap. Appendix A presents concrete instances of (OBO), namely instrumental variable regression and value function fitting in reinforcement learning. Appendix B discusses Assumptions A to D. Appendix C establishes convergence guarantees for projected mini-batch biased SGD applied to the empirical approximation of (OBO). Within it, Appendix C.1 derives the hypergradient and the smoothness of the value function, Appendix C.2 analyzes the statistical properties of the hypergradient estimator, and Appendix C.3 proves the main convergence theorem. These guarantees in the NTK regime rely on the generalization bounds for (OBO) derived in Appendix D. Those bounds build on Appendix E, which establishes non-asymptotic generalization bounds for neural network gradient flow in the NTK regime and provides the theoretical foundation used to approximate both the inner solution $h^\star_\omega$ and the regression function $b^\star$. Differentiability of the objectives and properties of the optimal solutions are provided in Appendix F. Appendix G shows that the RKHS of the infinite-width NTK and its interpolation spaces are Sobolev spaces. Background on integral and covariance operators as well as interpolation spaces is given in Appendix H. Auxiliary technical lemmas used throughout the proofs are collected in Appendix I. Finally, further experimental details and additional numerical results are provided in Appendix J.

Notations. $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ denote the Euclidean inner product and its induced norm, $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and $\| \cdot \|_{\mathcal{H}}$ denote the inner product and norm of a Hilbert space $\mathcal{H}$, $\| \cdot \|_{\mathrm{op}}$ denotes the operator norm, $\| \cdot \|_{\mathrm{F}}$ denotes the Frobenius norm, and $\| \cdot \|_{\mathrm{HS}}$ denotes the Hilbert-Schmidt norm. $\odot$ denotes the Hadamard product. For any $v_1, \dots, v_p \in \mathbb{R}$, $\mathbf{diag}(v_1, \dots, v_p) \in \mathbb{R}^{p \times p}$ denotes the diagonal matrix of size $p \times p$, where the diagonal entries are $v_1, \dots, v_p$ and all the off-diagonal entries are 0. $\mathbb{1}_p$ denotes a vector of size p where all entries are 1. $\mathbb{1}_{p \times p}$ denotes the identity matrix of size p. For $x \in \mathbb{R}$, $\lfloor x \rfloor$ denotes the largest integer less than or equal to x, and $\lceil x \rceil$ denotes the smallest integer greater than or equal to x. For any event $\mathcal{E}$, $\mathbb{1}_{\mathcal{E}}$ denotes the indicator function of $\mathcal{E}$. Given a compact set $\mathcal{K}$, $\mathrm{diam}(\mathcal{K})$ denotes its diameter. $\mathrm{Id}_{\mathcal{H}}$ denotes the identity operator on a Hilbert space $\mathcal{H}$. For a set $\mathcal{A}$, $\mathcal{A}^{\complement}$ denotes its complement. $\mathcal{H}_1 \cong \mathcal{H}_2$ means $\mathcal{H}_1 = \mathcal{H}_2$ as sets with equivalent norms. $a \asymp b$ means that there exist constants $c, C > 0$ such that $cb \leq a \leq Cb$.

Contents

A Concrete Examples of (OBO) **18**

B Discussion of Assumptions A to D **18**

C Convergence of Projected Mini-Batch Biased SGD **19**

C.1 Hypergradient and smoothness property of the value function 19

C.2 Statistical properties of the hypergradient estimator 19

C.3 Main theorem: convergence of projected mini-batch biased SGD 21

D Generalization Bounds for (OBO) in the NTK Regime **25**

D.1 Lipschitz continuity of the inner NTK approximation error in the outer variable . . 26

D.2 Generalization bounds for the inner solution, regression and adjoint functions . . . 27

E Generalization Analysis of Neural Network Gradient Flow in the NTK Regime **31**

E.1 Problem setting 31

E.2 Neural network architecture and infinite-width training dynamics 32

E.3 Assumptions 34

E.4 Lower-bound on the smallest eigenvalue of $\boldsymbol{K}^{\mathrm{NTK}}$ 35

E.5 Eigenvalue decay rate of $\mathcal{H}^{\mathrm{NTK}}$ 40

E.6 Bounds on the training dynamics 41
E.7 Uniform NTK approximation bounds 44
E.8 Generalization guarantees 47

F Differentiability of the Objectives and Properties of the Optimal Solutions 56

G A Sobolev Space Characterization of the Infinite-Width NTK's RKHS 56
G.1 Preliminaries 57
G.2 The sphere kernel: parity gap and the hemisphere lemma 59
G.3 The RKHS $\mathcal{H}^{\mathrm{NTK}}$ is a Sobolev space 61
G.4 The interpolation spaces of $\mathcal{H}^{\mathrm{NTK}}$ are Sobolev spaces 63

H Background on Integral/Covariance Operators and Interpolation Spaces 64

I Auxiliary Technical Lemmas 65

J Further Details on Experiments and Additional Numerical Results 70
J.1 Empirical validation of the theory in the NTK regime on synthetic data 70
J.2 Insights into feature learning beyond the NTK regime 70

A Concrete Examples of (OBO)

Instrumental variable regression. It is an approach that handles endogeneity in statistical modeling, relying on instruments to identify causal relationships [73]. The aim is to learn a function $z \mapsto g_\omega(z)$, parameterized by a vector ω, satisfying $y = g_\omega(z) + \epsilon$, where $y \in \mathcal{Y}$ denotes the observed outcome, $z \in \mathcal{Z}$ the treatment, and ϵ the error term. The central difficulty lies in the endogeneity of z, namely its correlation with ϵ, which renders direct regression inconsistent. Exogeneity no longer holds, since z and ϵ are no longer independent, and the estimates of g_ω are distorted as a result. A remedy is to bring in an instrumental variable $x \in \mathcal{X}$, correlated with z but not with ϵ, which retrieves the link between y and z while remaining immune to the distortion carried by ϵ, classically through two-stage least squares regression [5]. Following [93], this formulation translates directly into a bilevel problem whose inner and outer objectives are those given in Equation (1), where h is taken in $L^2(\mathcal{X})$.

Value function fitting in reinforcement learning. A second instance arises in reinforcement learning. Let $x = (s, a)$ be a state–action pair and $z = (s, a, r, s')$ the transition it generates, where r is the reward received and s' is the successor state. Let $g_\omega(z) = r + \gamma V_\omega(s')$ be the one-step Bellman target associated with a parametric value function V_ω, where $\gamma \in (0, 1)$ is the discount factor. This target is random given x, since it depends on which reward and successor state are drawn. The inner problem averages it over that randomness, returning the Bellman backup $h^\star_\omega(s, a) = \mathbb{E}_{r,s'}[r + \gamma V_\omega(s') \mid s, a]$, which fitted value iteration estimates by regressing the target on (s, a) [71]. The outer problem then fits this backup to the discounted return y observed along a trajectory started at (s, a), whose conditional expectation given (s, a) is the true action-value function. Minimizing $\mathcal{F}$ therefore selects the value function whose averaged one-step target $h^\star_\omega(s, a)$ is closest, in $L^2(\mathcal{X})$, to the observed return y.

B Discussion of Assumptions A to D

Assumption C is mild and standard in practice, as inputs are naturally bounded in most applications, *e.g.*, images and physical measurements, or can be made so through preprocessing procedures such as normalization or clipping. Assumption B is common in supervised learning, as the output naturally takes values in a bounded range in most regression tasks, *e.g.*, probabilities in $[0, 1]$, normalized scores, or physical measurements on a known scale, and can always be enforced by output normalization

otherwise. Assumption A holds in practice for a broad class of models, including linear models with bounded features, *e.g.*, random Fourier features [80], when the parameter space is compact. It also applies to feedforward neural networks or transformers with smooth activations, *e.g.*, SiLU [81] or GeLU [48], provided that the parameter space and the input domain are compact. Assumption D is a standard regularity condition on the input data distribution. Together with Assumption C, it is satisfied, for instance, by the uniform distribution on any compact set with nonempty interior, by Gaussian distributions truncated to a ball, and by finite mixtures of such distributions. It excludes, however, distributions without a Lebesgue density, such as discrete inputs (*e.g.*, binary or one-hot features) and distributions concentrated on lower-dimensional sets (*e.g.*, the uniform distribution on the unit sphere $\mathbb{S}^{p-1}$, or compositional data constrained to the simplex), as well as densities with singularities (*e.g.*, the arcsine distribution $\mathrm{Beta}(\frac{1}{2}, \frac{1}{2})$ on $[0, 1]$, whose density diverges at both endpoints).

C Convergence of Projected Mini-Batch Biased SGD

C.1 Hypergradient and smoothness property of the value function

Proof of Proposition 2.1. The differentiability of $\mathcal{F}$ as well as the expression of its hypergradient $\nabla\mathcal{F}$ are a direct consequence of [78, Proposition 2.2] together with Lemmas F.1 and F.2. Next, we prove the smoothness property of $\mathcal{F}$. Let $\omega_1, \omega_2 \in U$. We have:

$$
\begin{aligned}
&\|\nabla\mathcal{F}(\omega_1) - \nabla\mathcal{F}(\omega_2)\| \\
&= \left\|\mathbb{E}_{\mathbb{P}}\left[\partial_\omega g_{\omega_1}(z) a^\star_{\omega_1}(x) - \partial_\omega g_{\omega_2}(z) a^\star_{\omega_2}(x)\right]\right\| \\
&\leq \mathbb{E}_{\mathbb{P}}\left[\left\|\partial_\omega g_{\omega_1}(z) a^\star_{\omega_1}(x) - \partial_\omega g_{\omega_2}(z) a^\star_{\omega_2}(x)\right\|\right] \\
&= \mathbb{E}_{\mathbb{P}}\left[\left\|\partial_\omega g_{\omega_1}(z)(a^\star_{\omega_1}(x) - a^\star_{\omega_2}(x)) + (\partial_\omega g_{\omega_1}(z) - \partial_\omega g_{\omega_2}(z)) a^\star_{\omega_2}(x)\right\|\right] \\
&\leq \mathbb{E}_{\mathbb{P}}\left[\|\partial_\omega g_{\omega_1}(z)\| \left|a^\star_{\omega_1}(x) - a^\star_{\omega_2}(x)\right| + \|\partial_\omega g_{\omega_1}(z) - \partial_\omega g_{\omega_2}(z)\| \left|a^\star_{\omega_2}(x)\right|\right] \\
&= \mathbb{E}_{\mathbb{P}}\left[\|\partial_\omega g_{\omega_1}(z)\| \left|h^\star_{\omega_1}(x) - h^\star_{\omega_2}(x)\right| + \|\partial_\omega g_{\omega_1}(z) - \partial_\omega g_{\omega_2}(z)\| \left|a^\star_{\omega_2}(x)\right|\right], \qquad (8)
\end{aligned}
$$

where, in the third line, we use Jensen's inequality, and in the last line, we use $a^\star_{\omega_1} - a^\star_{\omega_2} = -(h^\star_{\omega_1} - h^\star_{\omega_2})$, the function $b^\star$ being independent of ω. By Assumption A, $\sup_{\omega\in U, z\in\mathcal{Z}} \|\partial_\omega g_\omega(z)\| \leq G$ and $\omega \mapsto g_\omega(z)$ is L-smooth for any $z \in \mathcal{Z}$, that is, $\|\partial_\omega g_{\omega_1}(z) - \partial_\omega g_{\omega_2}(z)\| \leq L\|\omega_1 - \omega_2\|$. Furthermore, by Lemma F.3, $\|a^\star_{\omega_2}\|_\infty \leq \sup_{\omega\in U} \|a^\star_\omega\|_\infty \leq D + B$. Combining these bounds in Equation (8), we obtain:

$$
\begin{aligned}
\|\nabla\mathcal{F}(\omega_1) - \nabla\mathcal{F}(\omega_2)\| &\leq G\,\mathbb{E}_{\mathbb{P}}\left[\left|h^\star_{\omega_1}(x) - h^\star_{\omega_2}(x)\right|\right] + L(D+B)\,\|\omega_1 - \omega_2\| \\
&= G\left\|h^\star_{\omega_1} - h^\star_{\omega_2}\right\|_{L^1(\mathcal{X})} + L(D+B)\,\|\omega_1 - \omega_2\| \\
&\leq G\left\|h^\star_{\omega_1} - h^\star_{\omega_2}\right\|_{L^2(\mathcal{X})} + L(D+B)\,\|\omega_1 - \omega_2\| \\
&\leq \underbrace{\left(G^2 + L(D+B)\right)}_{\mathcal{L}}\|\omega_1 - \omega_2\|,
\end{aligned}
$$

where, in the third line, we use $\|\cdot\|_{L^1(\mathcal{X})} \leq \|\cdot\|_{L^2(\mathcal{X})}$, valid since $\mathbb{P}_X$ is a probability measure, and where $L^1(\mathcal{X})$ denotes the space of integrable functions on $\mathcal{X}$ with respect to $\mathbb{P}_X$, endowed with the norm $\|f\|_{L^1(\mathcal{X})} = \mathbb{E}_{x\sim\mathbb{P}_X}[|f(x)|]$. In the last line, we use the fact that $\omega \mapsto h^\star_\omega$ is G-Lipschitz continuous, as established in Proposition F.4. This concludes the proof. □

C.2 Statistical properties of the hypergradient estimator

Lemma C.1 (Conditional mean of the hypergradient estimator)**.** *Fix $\omega \in \Omega$. Let $\hat{h}_\omega$ and $\hat{b}$ be any estimators of the inner solution $h^\star_\omega$ and of the regression function $b^\star$, built from $\mathcal{D}_{in}$ and $\mathcal{D}_{out}$, and set $\hat{a}_\omega = -\hat{h}_\omega + \hat{b}$. Then:*

$$
\widehat{\nabla\mathcal{F}}(\omega) \coloneqq \mathbb{E}\left[\widehat{\nabla\mathcal{F}}(\omega) \;\middle|\; \hat{a}_\omega\right] = -\mathbb{E}_{\mathbb{P}}\left[\partial_\omega g_\omega(z)\hat{a}_\omega(x)\right],
$$

where $\mathbb{E}_{\mathbb{P}}$ integrates over $(x, z) \sim \mathbb{P}_{XZ}$ only, $\hat{a}_\omega$ being held fixed.

Proof. Let $\omega \in \Omega$. By Equation (4), we have:

$$\widehat{\nabla \mathcal{F}}(\omega) = -\frac{1}{n_g} \sum_{l=1}^{n_g} \partial_\omega g_\omega(\underline{z}_l) \hat{a}_\omega(\underline{x}_l),$$

where $\mathcal{B}_g = \{(\underline{x}_l, \underline{z}_l)\}_{l=1}^{n_g}$ is the mini-batch used at ω. Since $\hat{a}_\omega$ is built from $\mathcal{D}_{in}$ and $\mathcal{D}_{out}$, which are independent of $\mathcal{B}_g$, the pairs $(\underline{x}_l, \underline{z}_l)$ remain *i.i.d.* with law $\mathbb{P}_{XZ}$ conditionally on $\hat{a}_\omega$. Hence, each summand has the same conditional expectation:

$$\mathbb{E}\left[\partial_\omega g_\omega(\underline{z}_l) \hat{a}_\omega(\underline{x}_l) \;\middle|\; \hat{a}_\omega\right] = \mathbb{E}_{\mathbb{P}}\left[\partial_\omega g_\omega(z) \hat{a}_\omega(x)\right], \qquad l = 1, \dots, n_g,$$

and averaging over l yields the announced identity by linearity of the expectation. □

Lemma C.2 (Bias control)**.** *Let Assumption A hold. Let $\{\hat{h}_\omega\}_{\omega\in\Omega}$ and $\hat{b}$ be any estimators of $\{h^\star_\omega\}_{\omega\in\Omega}$ and $b^\star$ built from $\mathcal{D}_{in}$ and $\mathcal{D}_{out}$, possibly using additional randomness independent of the mini-batches, and set $\hat{a}_\omega = -\hat{h}_\omega + \hat{b}$ and $a^\star_\omega = -h^\star_\omega + b^\star$ for every $\omega \in \Omega$. Then:*

$$\mathbb{E}\left[\sup_{\omega\in\Omega} \left\| \widetilde{\nabla \mathcal{F}}(\omega) - \nabla \mathcal{F}(\omega) \right\|^2\right] \le G^2\, \mathbb{E}\left[\sup_{\omega\in\Omega} \|\hat{a}_\omega - a^\star_\omega\|^2_{L^2(\mathcal{X})}\right] \le 2G^2 \left(e_h + e_b\right),$$

where $\widetilde{\nabla \mathcal{F}}(\omega) = \mathbb{E}[\widehat{\nabla \mathcal{F}}(\omega) \mid \hat{a}_\omega]$, $e_h = \mathbb{E}[\sup_{\omega\in\Omega} \|\hat{h}_\omega - h^\star_\omega\|^2_{L^2(\mathcal{X})}]$, $e_b = \mathbb{E}[\|\hat{b} - b^\star\|^2_{L^2(\mathcal{X})}]$, and the expectation is taken over the randomness used to construct $\hat{h}_\omega$ and $\hat{b}$.

Proof. Fix $\omega \in \Omega$. By Lemma C.1 and Proposition 2.1, we have:

$$\widetilde{\nabla \mathcal{F}}(\omega) - \nabla \mathcal{F}(\omega) = -\mathbb{E}_{\mathbb{P}}\left[\partial_\omega g_\omega(z) \left(\hat{a}_\omega(x) - a^\star_\omega(x)\right)\right].$$

Since $\|\cdot\|^2$ is convex, Jensen's inequality gives:

$$\left\| \widetilde{\nabla \mathcal{F}}(\omega) - \nabla \mathcal{F}(\omega) \right\|^2 \le \mathbb{E}_{\mathbb{P}}\left[\|\partial_\omega g_\omega(z)\|^2 \left(\hat{a}_\omega(x) - a^\star_\omega(x)\right)^2\right] \le G^2 \|\hat{a}_\omega - a^\star_\omega\|^2_{L^2(\mathcal{X})},$$

where the last inequality uses $\sup_{\omega\in\Omega, z\in\mathcal{Z}} \|\partial_\omega g_\omega(z)\| \le G$ from Assumption A, together with the fact that $L^2(\mathcal{X})$ is taken with respect to $\mathbb{P}_X$, so that $\mathbb{E}_{\mathbb{P}}[(\hat{a}_\omega(x) - a^\star_\omega(x))^2] = \|\hat{a}_\omega - a^\star_\omega\|^2_{L^2(\mathcal{X})}$. This inequality holds for every $\omega \in \Omega$ and every realization of the randomness used to construct $\hat{h}_\omega$ and $\hat{b}$, so we may take the supremum over $\omega \in \Omega$ on both sides and then the expectation, which gives the first bound. For the second, the triangle inequality and $(u+v)^2 \le 2u^2 + 2v^2$ yield, for every $\omega \in \Omega$:

$$\begin{aligned}
\|\hat{a}_\omega - a^\star_\omega\|^2_{L^2(\mathcal{X})} &\le \left(\left\| \hat{h}_\omega - h^\star_\omega \right\|_{L^2(\mathcal{X})} + \left\| \hat{b} - b^\star \right\|_{L^2(\mathcal{X})} \right)^2 \\
&\le 2 \left(\left\| \hat{h}_\omega - h^\star_\omega \right\|^2_{L^2(\mathcal{X})} + \left\| \hat{b} - b^\star \right\|^2_{L^2(\mathcal{X})} \right).
\end{aligned}$$

Taking the supremum over $\omega \in \Omega$ and then the expectation concludes the proof. □

Lemma C.3 (Variance control)**.** *Let Assumptions A and B hold. Let $\{\hat{h}_\omega\}_{\omega\in\Omega}$ and $\hat{b}$ be any estimators of $\{h^\star_\omega\}_{\omega\in\Omega}$ and $b^\star$ built from $\mathcal{D}_{in}$ and $\mathcal{D}_{out}$, possibly using additional randomness independent of the mini-batches, and set $\hat{a}_\omega = -\hat{h}_\omega + \hat{b}$ and $a^\star_\omega = -h^\star_\omega + b^\star$ for every $\omega \in \Omega$. Then, for every $\omega \in \Omega$, we have:*

$$\begin{aligned}
\mathbb{E}\left[\left\| \widehat{\nabla \mathcal{F}}(\omega) - \widetilde{\nabla \mathcal{F}}(\omega) \right\|^2\right] &\le \frac{2G^2}{n_g} \left(\mathbb{E}\left[\sup_{\omega'\in\Omega} \|\hat{a}_{\omega'} - a^\star_{\omega'}\|^2_{L^2(\mathcal{X})}\right] + (D+B)^2 \right) \\
&\le \frac{2G^2}{n_g} \left[2\left(e_h + e_b\right) + (D+B)^2 \right],
\end{aligned}$$

where $\widetilde{\nabla \mathcal{F}}(\omega) = \mathbb{E}[\widehat{\nabla \mathcal{F}}(\omega) \mid \hat{a}_\omega]$, $e_h = \mathbb{E}[\sup_{\omega\in\Omega} \|\hat{h}_\omega - h^\star_\omega\|^2_{L^2(\mathcal{X})}]$, $e_b = \mathbb{E}[\|\hat{b} - b^\star\|^2_{L^2(\mathcal{X})}]$, and the expectation is taken over all sources of randomness.

Proof. Fix $\omega \in \Omega$. Let $\mathcal{B}_g = \{(\underline{x}_l, \underline{z}_l)\}_{l=1}^{n_g}$ be the mini-batch used at ω, and set:

$$U_l(\omega) = -\partial_\omega g_\omega(\underline{z}_l)\hat{a}_\omega(\underline{x}_l), \qquad l = 1, \dots, n_g,$$

so that $\widehat{\nabla\mathcal{F}}(\omega) = \frac{1}{n_g}\sum_{l=1}^{n_g} U_l(\omega)$ and, by Lemma C.1, $\widetilde{\nabla\mathcal{F}}(\omega) = \mathbb{E}[U_l(\omega) \mid \hat{a}_\omega]$ for every l. As $\mathcal{B}_g$ is independent of $\hat{a}_\omega$, the variables $U_1(\omega), \dots, U_{n_g}(\omega)$ are *i.i.d.* conditionally on $\hat{a}_\omega$, so that the conditional variance of their average scales as $1/n_g$:

$$\mathbb{E}\left[\left\|\widehat{\nabla\mathcal{F}}(\omega) - \widetilde{\nabla\mathcal{F}}(\omega)\right\|^2 \,\middle|\, \hat{a}_\omega\right] \le \frac{1}{n_g}\mathbb{E}\left[\|U_1(\omega)\|^2 \,\middle|\, \hat{a}_\omega\right].$$

By Assumption A, $\|U_1(\omega)\|^2 \le G^2|\hat{a}_\omega(\underline{x}_1)|^2$, and since $\underline{x}_1 \sim \mathbb{P}_X$ conditionally on $\hat{a}_\omega$:

$$\begin{aligned}\mathbb{E}\left[\|U_1(\omega)\|^2 \,\middle|\, \hat{a}_\omega\right] &\le G^2\,\|\hat{a}_\omega\|_{L^2(\mathcal{X})}^2 \le 2G^2\left(\|\hat{a}_\omega - a_\omega^\star\|_{L^2(\mathcal{X})}^2 + \|a_\omega^\star\|_{L^2(\mathcal{X})}^2\right)\\ &\le 2G^2\left(\sup_{\omega'\in\Omega}\|\hat{a}_{\omega'} - a_{\omega'}^\star\|_{L^2(\mathcal{X})}^2 + (D+B)^2\right),\end{aligned}$$

where the second inequality follows from the elementary inequality $\|u+v\|^2 \le 2(\|u\|^2 + \|v\|^2)$, and the third bounds $\|\hat{a}_\omega - a_\omega^\star\|_{L^2(\mathcal{X})}^2$ by its supremum over Ω and uses $\|a_\omega^\star\|_{L^2(\mathcal{X})} \le \|a_\omega^\star\|_\infty$, valid since $\mathbb{P}_X$ is a probability measure, together with the bound $\|a_\omega^\star\|_\infty \le D + B$ established in Lemma F.3. Taking the total expectation gives the first bound. For the second, the triangle inequality and $(u+v)^2 \le 2u^2 + 2v^2$ yield, for every $\omega' \in \Omega$:

$$\|\hat{a}_{\omega'} - a_{\omega'}^\star\|_{L^2(\mathcal{X})}^2 \le 2\left\|\hat{h}_{\omega'} - h_{\omega'}^\star\right\|_{L^2(\mathcal{X})}^2 + 2\left\|\hat{b} - b^\star\right\|_{L^2(\mathcal{X})}^2,$$

so that taking the supremum over $\omega' \in \Omega$ and then the expectation gives $\mathbb{E}[\sup_{\omega'\in\Omega}\|\hat{a}_{\omega'} - a_{\omega'}^\star\|_{L^2(\mathcal{X})}^2] \le 2(e_h + e_b)$, since $\hat{b}$ does not depend on ω'. □

C.3 Main theorem: convergence of projected mini-batch biased SGD

We write δ_Ω for the *indicator function* of Ω, that is $\delta_\Omega(\omega) = 0$ if $\omega \in \Omega$ and $+\infty$ otherwise — a standard device that folds the constraint $\omega \in \Omega$ into the objective itself, so that minimizing $\mathcal{F} + \delta_\Omega$ over all of $\mathbb{R}^d$ is the same problem as minimizing $\mathcal{F}$ over Ω alone. We set $f := \mathcal{F} + \delta_\Omega$. For ω with $f(\omega) < \infty$, $\partial f(\omega)$ denotes the (limiting) subdifferential of f at ω: a *set* of generalized slopes, reducing to $\{\nabla f(\omega)\}$ where f is differentiable, with $0 \in \partial f(\omega^\star)$ the first-order stationarity condition for $\min_{\omega\in\Omega}\mathcal{F}(\omega)$. For a point p and a set S, $\operatorname{dist}(p, S) := \inf_{s\in S}\|p - s\|$ is the ordinary distance from p to S; we use it only as $\operatorname{dist}(0, \partial f(\omega))$, a continuous relaxation of exact stationarity (zero exactly when $0 \in \partial f(\omega)$, small when *some* element of $\partial f(\omega)$ is nearly zero). A function ϕ is *ρ-weakly convex* if $\phi + \frac{\rho}{2}\|\cdot\|^2$ is convex. We note that every $\mathcal{L}$-smooth function is $\mathcal{L}$-weakly convex.

The Moreau envelope. Since $\mathcal{F}$ is $\mathcal{L}$-smooth on Ω (Proposition 2.1), $\mathcal{F} + \frac{\mathcal{L}}{2}\|\cdot\|^2$ is convex, so $\mathcal{F}$ is $\mathcal{L}$-weakly convex. Moreover, since Ω is convex, δ_Ω is convex, so $f = \mathcal{F} + \delta_\Omega$ is $\mathcal{L}$-weakly convex as well, being a sum of the two. Fix $\lambda \in (0, \mathcal{L}^{-1})$ and define, for $\omega \in \Omega$, the Moreau envelope and the proximal map of $\mathcal{F}$ as:

$$\mathcal{F}_\lambda(\omega) := \min_{y\in\Omega}\left\{\mathcal{F}(y) + \frac{1}{2\lambda}\|y - \omega\|^2\right\}, \qquad \operatorname{prox}_{\lambda\mathcal{F}}(\omega) := \operatorname*{argmin}_{y\in\Omega}\left\{\mathcal{F}(y) + \frac{1}{2\lambda}\|y - \omega\|^2\right\}. \tag{9}$$

Since $\lambda < \mathcal{L}^{-1}$, the minimized objective is $(\lambda^{-1} - \mathcal{L})$-strongly convex on Ω, so the minimum is attained at a unique point and $\operatorname{prox}_{\lambda\mathcal{F}}(\omega)$ is well defined. By the standard Moreau-envelope calculus for weakly convex functions [70, 83], $\mathcal{F}_\lambda$ is C^1 with:

$$\nabla\mathcal{F}_\lambda(\omega) = \frac{1}{\lambda}\left(\omega - \operatorname{prox}_{\lambda\mathcal{F}}(\omega)\right), \qquad \|\operatorname{prox}_{\lambda\mathcal{F}}(\omega) - \omega\| = \lambda\,\|\nabla\mathcal{F}_\lambda(\omega)\|. \tag{10}$$

Writing $u_k := \operatorname{prox}_{\lambda\mathcal{F}}(\omega_k)$, we have $f(u_k) \le f(\omega_k)$, since ω_k itself is a feasible, zero-penalty competitor in the minimization defining $\mathcal{F}_\lambda(\omega_k)$, and $\operatorname{dist}(0, \partial f(u_k)) \le \|\nabla\mathcal{F}_\lambda(\omega_k)\|$, since the first-order optimality condition for u_k reads exactly $\nabla\mathcal{F}_\lambda(\omega_k) \in \partial f(u_k)$ by the subdifferential sum rule, valid here because the added quadratic term is smooth. A small $\|\nabla\mathcal{F}_\lambda(\omega_k)\|$ therefore certifies that ω_k is close to a point u_k that is itself near-stationary for (OBO), with no loss of objective value. This is the C^1 surrogate for $\operatorname{dist}(0, \partial f(\omega_k))$ itself, which is generally discontinuous in ω_k and hence unsuitable for a descent-type stochastic analysis.

Lemma C.4 (One-step envelope descent)**.** *Let Assumptions A and B hold and let $\lambda \in (0, \mathcal{L}^{-1})$, where $\mathcal{L}$ is the smoothness parameter of $\mathcal{F}$ provided in Proposition 2.1. Let $\{\omega_k\}_{k\geq 0}$ be generated by $\omega_{k+1} = \Pi_\Omega\left(\omega_k - \eta\widehat{\nabla\mathcal{F}}(\omega_k)\right)$ with $0 < \eta \leq \frac{1}{2\mathcal{L}}$. Then, for every $k \geq 0$:*

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{F}_\lambda(\omega_{k+1}) \mid \hat{a}_{\omega_k}\right] \\
&\leq \mathcal{F}_\lambda(\omega_k) - \frac{3\eta}{4}\left(1 - \lambda\mathcal{L}\right)\|\nabla\mathcal{F}_\lambda(\omega_k)\|^2 + \eta\left(\frac{1}{1-\lambda\mathcal{L}} + \frac{2\eta}{\lambda}\right)\left\|\widetilde{\nabla\mathcal{F}}(\omega_k) - \nabla\mathcal{F}(\omega_k)\right\|^2 \\
&\quad + \frac{2\eta^2}{\lambda}\,\mathbb{E}\left[\left\|\widehat{\nabla\mathcal{F}}(\omega_k) - \widetilde{\nabla\mathcal{F}}(\omega_k)\right\|^2 \,\middle|\, \hat{a}_{\omega_k}\right] + \frac{\eta}{\lambda}\Big(\mathcal{F}(\omega_k) - \mathbb{E}\left[\mathcal{F}(\omega_{k+1}) \mid \hat{a}_{\omega_k}\right]\Big).
\end{aligned}$$

Proof. Let $k \geq 0$. Write $u_k := \mathrm{prox}_{\lambda\mathcal{F}}(\omega_k) \in \Omega$ for the proximal point of $\mathcal{F}$ at ω_k, that is, the unique minimizer in Equation (9), so that $\|u_k - \omega_k\| = \lambda\|\nabla\mathcal{F}_\lambda(\omega_k)\|$ (Equation (10)). Write also $\mathbb{E}_k[\cdot] := \mathbb{E}[\cdot \mid \hat{a}_{\omega_k}]$[3], with $g_k = \widehat{\nabla\mathcal{F}}(\omega_k)$, $v_k := \mathbb{E}_k[\widehat{\nabla\mathcal{F}}(\omega_k)] = \widetilde{\nabla\mathcal{F}}(\omega_k)$, $b_k := v_k - \nabla\mathcal{F}(\omega_k)$, $\xi_k := g_k - v_k$, and $\Delta_k := \omega_{k+1} - \omega_k$.

Step 1: A one-step bound on $\mathbb{E}_k[\mathcal{F}_\lambda]$. By construction, it holds that $\omega_{k+1} = \Pi_\Omega(\omega_k - \eta g_k) = \mathrm{argmin}_{y\in\Omega}\, h(y)$ with $h(y) := \langle g_k, y - \omega_k\rangle + \frac{1}{2\eta}\|y - \omega_k\|^2$, which is $\frac{1}{\eta}$-strongly convex. Since h is quadratic, for any $y \in \Omega$, one has the exact expansion $h(y) = h(\omega_{k+1}) + \langle\nabla h(\omega_{k+1}), y - \omega_{k+1}\rangle + \frac{1}{2\eta}\|y - \omega_{k+1}\|^2$. In addition, the first-order optimality condition for the constrained minimizer ω_{k+1} gives $\langle\nabla h(\omega_{k+1}), y - \omega_{k+1}\rangle \geq 0$ for all $y \in \Omega$ [22, Proposition 1.3]. Hence, taking $y = u_k \in \Omega$, we obtain:

$$h(u_k) \;\geq\; h(\omega_{k+1}) + \frac{1}{2\eta}\|u_k - \omega_{k+1}\|^2.$$

Expanding both sides via the definition of h and multiplying through by $2\eta > 0$ yields:

$$\|u_k - \omega_{k+1}\|^2 \leq \|u_k - \omega_k\|^2 - \|\Delta_k\|^2 + 2\eta\langle g_k, u_k - \omega_{k+1}\rangle. \tag{11}$$

This transfers to $\mathcal{F}_\lambda$ as follows. Since $u_k \in \Omega$ is feasible for the minimization defining $\mathcal{F}_\lambda(\omega_{k+1})$, but not necessarily its minimizer, we get:

$$\mathcal{F}_\lambda(\omega_{k+1}) \leq \mathcal{F}(u_k) + \frac{1}{2\lambda}\|u_k - \omega_{k+1}\|^2,$$

and by definition $\mathcal{F}(u_k) + \frac{1}{2\lambda}\|u_k - \omega_k\|^2 = \mathcal{F}_\lambda(\omega_k)$ exactly. Multiplying Equation (11) by $\frac{1}{2\lambda}$ and substituting:

$$\mathcal{F}_\lambda(\omega_{k+1}) \leq \mathcal{F}_\lambda(\omega_k) - \frac{1}{2\lambda}\|\Delta_k\|^2 + \frac{\eta}{\lambda}\langle g_k, u_k - \omega_{k+1}\rangle. \tag{12}$$

Since $g_k = \nabla\mathcal{F}(\omega_k) + b_k + \xi_k$ and $u_k - \omega_{k+1} = (u_k - \omega_k) - \Delta_k$, we have:

$$\langle g_k, u_k - \omega_{k+1}\rangle = \langle g_k, u_k - \omega_k\rangle - \langle\nabla\mathcal{F}(\omega_k), \Delta_k\rangle - \langle b_k, \Delta_k\rangle - \langle\xi_k, \Delta_k\rangle.$$

By $\mathcal{L}$-smoothness of $\mathcal{F}$ (Proposition 2.1), it holds that $-\langle\nabla\mathcal{F}(\omega_k), \Delta_k\rangle \leq \mathcal{F}(\omega_k) - \mathcal{F}(\omega_{k+1}) + \frac{\mathcal{L}}{2}\|\Delta_k\|^2$ [72, Lemma 1.2.3]. By Young's inequality with parameter $\frac{1}{4\eta}$, $-\langle b_k, \Delta_k\rangle \leq 2\eta\|b_k\|^2 + \frac{1}{8\eta}\|\Delta_k\|^2$ and likewise $-\langle\xi_k, \Delta_k\rangle \leq 2\eta\|\xi_k\|^2 + \frac{1}{8\eta}\|\Delta_k\|^2$. Multiplying these three bounds by $\frac{\eta}{\lambda}$ and inserting into Equation (12), the coefficient of $\|\Delta_k\|^2$ becomes $\frac{1}{\lambda}\big(-\frac{1}{2} + \frac{\eta\mathcal{L}}{2} + \frac{1}{8} + \frac{1}{8}\big) = \frac{1}{\lambda}\big(\frac{\eta\mathcal{L}}{2} - \frac{1}{4}\big)$, which is ≤ 0 for $\eta \leq \frac{1}{2\mathcal{L}}$, independently of λ. Dropping this nonpositive term, we get:

$$\mathcal{F}_\lambda(\omega_{k+1}) \leq \mathcal{F}_\lambda(\omega_k) + \frac{\eta}{\lambda}\langle g_k, u_k - \omega_k\rangle + \frac{\eta}{\lambda}\big(\mathcal{F}(\omega_k) - \mathcal{F}(\omega_{k+1})\big) + \frac{2\eta^2}{\lambda}\big(\|b_k\|^2 + \|\xi_k\|^2\big). \tag{13}$$

The quantities $\mathcal{F}_\lambda(\omega_k)$, $\mathcal{F}(\omega_k)$ and b_k are $\hat{a}_{\omega_k}$-measurable, and $\mathbb{E}_k[\xi_k] = 0$ since $\mathcal{B}_g^k$ is fresh, so that $\mathbb{E}_k[\langle g_k, u_k - \omega_k\rangle] = \langle v_k, u_k - \omega_k\rangle$. Taking $\mathbb{E}_k$ in Equation (13) gives:

$$\mathbb{E}_k[\mathcal{F}_\lambda(\omega_{k+1})] \leq \mathcal{F}_\lambda(\omega_k) + \frac{\eta}{\lambda}\langle v_k, u_k - \omega_k\rangle + \frac{\eta}{\lambda}\big(\mathcal{F}(\omega_k) - \mathbb{E}_k[\mathcal{F}(\omega_{k+1})]\big) + \frac{2\eta^2}{\lambda}\big(\|b_k\|^2 + \mathbb{E}_k[\|\xi_k\|^2]\big). \tag{14}$$

[3]Precisely, $\hat{a}_{\omega_k}$ is shorthand for the pair $(\omega_k, \hat{a}_{\omega_k})$: Step 1 treats $u_k - \omega_k$, $\mathcal{F}_\lambda(\omega_k)$, $\mathcal{F}(\omega_k)$ and b_k as known given $\mathbb{E}_k$, which requires ω_k itself to be fixed, not just the function $\hat{a}_{\omega_k}$. This causes no loss of generality: the fresh mini-batch $\mathcal{B}_g^k$ is independent of the *entire* history up to step k (not merely of $\hat{a}_{\omega_k}$), so every conditional expectation of a quantity built from $(\omega_k, \hat{a}_{\omega_k}, \mathcal{B}_g^k)$ takes the same value whether one conditions on $(\omega_k, \hat{a}_{\omega_k})$ or on the full history.

Step 2: Control on the cross term $\langle v_k, u_k - \omega_k\rangle$. Write $v_k = \nabla\mathcal{F}(\omega_k) + b_k$, so:

$$\langle v_k, u_k - \omega_k\rangle = \langle \nabla\mathcal{F}(\omega_k), u_k - \omega_k\rangle + \langle b_k, u_k - \omega_k\rangle. \tag{15}$$

Since $\mathcal{F}$ is $\mathcal{L}$-smooth (Proposition 2.1), [72, Lemma 1.2.3] gives:

$$\mathcal{F}(u_k) \geq \mathcal{F}(\omega_k) + \langle \nabla\mathcal{F}(\omega_k), u_k - \omega_k\rangle - \frac{\mathcal{L}}{2}\|u_k - \omega_k\|^2,$$

so

$$\langle \nabla\mathcal{F}(\omega_k), u_k - \omega_k\rangle \leq \mathcal{F}(u_k) - \mathcal{F}(\omega_k) + \frac{\mathcal{L}}{2}\|u_k - \omega_k\|^2. \tag{16}$$

We have by Young's inequality with parameter $\frac{1}{2}\left(\lambda^{-1} - \mathcal{L}\right) > 0$:

$$\langle b_k, u_k - \omega_k\rangle \leq \frac{1}{\lambda^{-1} - \mathcal{L}}\|b_k\|^2 + \frac{\lambda^{-1} - \mathcal{L}}{4}\|u_k - \omega_k\|^2. \tag{17}$$

The map $\psi_k(y) := \mathcal{F}(y) + \frac{1}{2\lambda}\|y - \omega_k\|^2$ is $(\lambda^{-1} - \mathcal{L})$-strongly convex on Ω (sum of the convex function $\mathcal{F} + \frac{\mathcal{L}}{2}\|\cdot\|^2$, up to an affine term, and a quadratic of curvature $\lambda^{-1} - \mathcal{L}$, which is positive since $\lambda < \mathcal{L}^{-1}$), with unique minimizer u_k. Comparing to the feasible point $y = \omega_k$, strong convexity gives:

$$\psi_k(\omega_k) \geq \psi_k(u_k) + \frac{\lambda^{-1} - \mathcal{L}}{2}\|\omega_k - u_k\|^2.$$

Since $\psi_k(\omega_k) = \mathcal{F}(\omega_k)$ and $\psi_k(u_k) = \mathcal{F}(u_k) + \frac{1}{2\lambda}\|u_k - \omega_k\|^2$, this rearranges to:

$$\mathcal{F}(u_k) - \mathcal{F}(\omega_k) \leq -\left(\lambda^{-1} - \frac{\mathcal{L}}{2}\right)\|u_k - \omega_k\|^2. \tag{18}$$

Adding Equation (16) and Equation (17) reconstructs the right-hand side of Equation (15). Substituting Equation (18) for $\mathcal{F}(u_k) - \mathcal{F}(\omega_k)$ and combining the three $\|u_k - \omega_k\|^2$ coefficients, $-\left(\lambda^{-1} - \frac{\mathcal{L}}{2}\right) + \frac{\mathcal{L}}{2} + \frac{\lambda^{-1} - \mathcal{L}}{4} = -\frac{3}{4}\left(\lambda^{-1} - \mathcal{L}\right)$:

$$\langle v_k, u_k - \omega_k\rangle \leq -\frac{3}{4}\left(\lambda^{-1} - \mathcal{L}\right)\|u_k - \omega_k\|^2 + \frac{1}{\lambda^{-1} - \mathcal{L}}\|b_k\|^2. \tag{19}$$

Step 3: Assembling the descent inequality. Plugging Equation (19) into Equation (14) and using $\|u_k - \omega_k\|^2 = \lambda^2\|\nabla\mathcal{F}_\lambda(\omega_k)\|^2$, the descent term becomes:

$$-\frac{\eta}{\lambda}\cdot\frac{3}{4}\left(\lambda^{-1} - \mathcal{L}\right)\cdot\lambda^2\|\nabla\mathcal{F}_\lambda(\omega_k)\|^2 = -\frac{3\eta}{4}\left(1 - \lambda\mathcal{L}\right)\|\nabla\mathcal{F}_\lambda(\omega_k)\|^2,$$

while the coefficient of $\|b_k\|^2$ becomes $\frac{\eta}{\lambda}\cdot\frac{1}{\lambda^{-1} - \mathcal{L}} + \frac{2\eta^2}{\lambda} = \eta\left(\frac{1}{1 - \lambda\mathcal{L}} + \frac{2\eta}{\lambda}\right)$. Hence:

$$\mathbb{E}_k[\mathcal{F}_\lambda(\omega_{k+1})] \leq \mathcal{F}_\lambda(\omega_k) - \frac{3\eta}{4}\left(1 - \lambda\mathcal{L}\right)\|\nabla\mathcal{F}_\lambda(\omega_k)\|^2 + \eta\left(\frac{1}{1 - \lambda\mathcal{L}} + \frac{2\eta}{\lambda}\right)\|b_k\|^2$$
$$+ \frac{2\eta^2}{\lambda}\,\mathbb{E}_k[\|\xi_k\|^2] + \frac{\eta}{\lambda}\big(\mathcal{F}(\omega_k) - \mathbb{E}_k[\mathcal{F}(\omega_{k+1})]\big).$$

Recalling $b_k = \widetilde{\nabla\mathcal{F}}(\omega_k) - \nabla\mathcal{F}(\omega_k)$ and $\xi_k = \widehat{\nabla\mathcal{F}}(\omega_k) - \widetilde{\nabla\mathcal{F}}(\omega_k)$ gives exactly the claimed inequality. □

Theorem C.5 (Convergence guarantee of projected mini-batch biased SGD)**.** *Fix $\omega_0 \in \Omega$. Let $\mathcal{F}^\star := \min_{\omega\in\Omega}\mathcal{F}(\omega)$ and $\Delta_\mathcal{F} := \mathcal{F}(\omega_0) - \mathcal{F}^\star$. Let $\{\omega_k\}_{k\geq 0}$ be generated by $\omega_{k+1} = \Pi_\Omega(\omega_k - \eta\widehat{\nabla\mathcal{F}}(\omega_k))$ with stepsize $0 < \eta \leq \frac{1}{2\mathcal{L}}$, where $\mathcal{L}$ is the smoothness parameter of $\mathcal{F}$ provided in Proposition 2.1. Under Assumptions A and B, for any $K \geq 1$ and any $\lambda \in (0, \mathcal{L}^{-1})$, the iterates $\{\omega_k\}_{k\geq 0}$ satisfy:*

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla\mathcal{F}_\lambda(\omega_k)\|^2\right] \leq \frac{4}{3\left(1 - \lambda\mathcal{L}\right)}\Bigg[\left(\frac{1}{\eta} + \frac{1}{\lambda}\right)\frac{\Delta_\mathcal{F}}{K} + 2G^2\left(\frac{1}{1 - \lambda\mathcal{L}} + \frac{2\eta}{\lambda}\right)(e_h + e_b)$$
$$+ \frac{4G^2\eta}{\lambda n_g}\left((D + B)^2 + 2\left(e_h + e_b\right)\right)\Bigg],$$

where $e_h = \mathbb{E}[\sup_{\omega\in\Omega}\|\hat{h}_\omega - h^\star_\omega\|^2_{L^2(\mathcal{X})}]$, $e_b = \mathbb{E}[\|\hat{b} - b^\star\|^2_{L^2(\mathcal{X})}]$, and B, D, G are the constants of Assumptions A and B. The expectation is over all sources of randomness, namely the samples $\mathcal{D}_{in}$ and $\mathcal{D}_{out}$ used to construct $\hat{h}_\omega$ and $\hat{b}$, and the mini-batches $\{\mathcal{B}_g^k\}_{k\geq 0}$; the expectations defining e_h and e_b are over $\mathcal{D}_{in}$ and $\mathcal{D}_{out}$ only.

Proof. Let $k \geq 0$, and recall that $b_k := \widetilde{\nabla \mathcal{F}}(\omega_k) - \nabla \mathcal{F}(\omega_k)$ is the bias of the plug-in hypergradient estimator $\widehat{\nabla \mathcal{F}}(\omega_k)$ conditional on $\hat{a}_{\omega_k}$, and $\xi_k := \widehat{\nabla \mathcal{F}}(\omega_k) - \widetilde{\nabla \mathcal{F}}(\omega_k)$ is its conditionally zero-mean noise.

Taking full expectation in Lemma C.4 and summing over $k = 0, \dots, K-1$, both the $\mathcal{F}_\lambda$- and the $\mathcal{F}$-sums telescope, and ω_0 is deterministic, so:

$$\begin{aligned}
&\frac{3\eta}{4}\left(1-\lambda\mathcal{L}\right)\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\mathcal{F}_\lambda(\omega_k)\right\|^2\right] \\
&\leq \mathcal{F}_\lambda(\omega_0) - \mathbb{E}\left[\mathcal{F}_\lambda(\omega_K)\right] + \frac{\eta}{\lambda}\left(\mathcal{F}(\omega_0) - \mathbb{E}\left[\mathcal{F}(\omega_K)\right]\right) + \eta\left(\frac{1}{1-\lambda\mathcal{L}} + \frac{2\eta}{\lambda}\right)\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|b_k\right\|^2\right] \\
&\quad + \frac{2\eta^2}{\lambda}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\xi_k\right\|^2\right].
\end{aligned}$$

Since $\mathcal{F}_\lambda(\omega) \leq \mathcal{F}(\omega)$ and $\mathcal{F}_\lambda(\omega), \mathcal{F}(\omega) \geq \mathcal{F}^\star$ for every $\omega \in \Omega$, the two differences on the right-hand side are bounded by $\left(1 + \frac{\eta}{\lambda}\right)\Delta_{\mathcal{F}}$. Dividing by $\frac{3\eta}{4}(1-\lambda\mathcal{L})K$, which is positive because $\lambda < \mathcal{L}^{-1}$, we get:

$$\begin{aligned}
&\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\mathcal{F}_\lambda(\omega_k)\right\|^2\right] \\
&\leq \frac{4}{3\left(1-\lambda\mathcal{L}\right)}\left(\frac{1}{\eta}+\frac{1}{\lambda}\right)\frac{\Delta_{\mathcal{F}}}{K} + \frac{4}{3\left(1-\lambda\mathcal{L}\right)}\left(\frac{1}{1-\lambda\mathcal{L}}+\frac{2\eta}{\lambda}\right)\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|b_k\right\|^2\right] \\
&\quad + \frac{8\eta}{3\lambda\left(1-\lambda\mathcal{L}\right)}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\xi_k\right\|^2\right]. \qquad (20)
\end{aligned}$$

The iterate ω_k depends on $\mathcal{D}_{in}$ and $\mathcal{D}_{out}$, so we bound both error terms uniformly over Ω. By Lemmas C.2 and C.3, for every k, we have:

$$\mathbb{E}\left[\left\|b_k\right\|^2\right] \leq 2G^2\left(e_h + e_b\right) \quad \text{and} \quad \mathbb{E}\left[\left\|\xi_k\right\|^2\right] \leq \frac{2G^2}{n_g}\left[\left(D+B\right)^2 + 2\left(e_h+e_b\right)\right].$$

Substituting these bounds into Equation (20) gives exactly the claimed bound. □

Proof of Theorem 2.4. We specialize Theorem C.5 to $\lambda = \frac{1}{2\mathcal{L}}$, which is admissible since $\frac{1}{2\mathcal{L}} < \mathcal{L}^{-1}$. With this choice, $\lambda\mathcal{L} = \frac{1}{2}$, so the prefactor equals $\frac{4}{3(1-\lambda\mathcal{L})} = \frac{8}{3}$ and $\frac{1}{\lambda} = 2\mathcal{L}$. Using $\eta \leq \frac{1}{2\mathcal{L}}$, that is $2\mathcal{L} \leq \frac{1}{\eta}$ and $\eta\mathcal{L} \leq \frac{1}{2}$, the three brackets are bounded by:

$$\frac{1}{\eta}+\frac{1}{\lambda} = \frac{1}{\eta}+2\mathcal{L} \leq \frac{2}{\eta}, \quad 2G^2\left(\frac{1}{1-\lambda\mathcal{L}}+\frac{2\eta}{\lambda}\right) = 2G^2\left(2+4\eta\mathcal{L}\right) \leq 8G^2, \quad \frac{4G^2\eta}{\lambda n_g} = \frac{8\mathcal{L}G^2\eta}{n_g},$$

so that Theorem C.5 gives:

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\mathcal{F}_{1/2\mathcal{L}}(\omega_k)\right\|^2\right] \leq \frac{16\Delta_{\mathcal{F}}}{3\eta K} + \frac{64G^2\left(e_h+e_b\right)}{3} + \frac{64\mathcal{L}G^2\eta}{3n_g}\left(\left(D+B\right)^2 + 2\left(e_h+e_b\right)\right). \qquad (21)$$

It remains to pass from the Moreau envelope to the gradient mapping. Applying [30, Section 2.2], we obtain, for every $\omega \in \Omega$:

$$\left\|\mathcal{G}_{1/\mathcal{L}}(\omega)\right\| \leq \frac{3}{2}\left(1+\frac{1}{\sqrt{2}}\right)\left\|\nabla\mathcal{F}_{1/2\mathcal{L}}(\omega)\right\|, \qquad \text{hence} \qquad \left\|\mathcal{G}_{1/\mathcal{L}}(\omega)\right\|^2 \leq c\left\|\nabla\mathcal{F}_{1/2\mathcal{L}}(\omega)\right\|^2,$$

with $c = \frac{9}{4}\left(\frac{3}{2}+\sqrt{2}\right)$. Taking expectations, averaging over $k = 0, \dots, K-1$ and combining with Equation (21) gives the claimed bound. □

Corollary C.6 (Balanced stepsize)**.** *Let the assumptions of Theorem 2.4 hold, assume $\Delta_{\mathcal{F}} > 0$, and set $\sigma^2 := (D+B)^2 + 2(e_h + e_b)$. For any $K \geq 1$, the choice*

$$\eta = \min\left\{\frac{1}{2\mathcal{L}},\ \eta_\star\right\}, \qquad \eta_\star := \frac{1}{2G\sigma}\sqrt{\frac{\Delta_{\mathcal{F}}\, n_g}{\mathcal{L}K}}, \tag{22}$$

is admissible and yields:

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\mathcal{G}_{1/\mathcal{L}}(\omega_k)\|^2\right] \leq 2c\left[\frac{\mathcal{L}\Delta_{\mathcal{F}}}{K} + 2G\sigma\sqrt{\frac{\mathcal{L}\Delta_{\mathcal{F}}}{Kn_g}} + 2G^2(e_h+e_b)\right], \tag{23}$$

where $c = 18 + 12\sqrt{2}$ is the constant of Theorem 2.4.

Proof. Write the bound of Theorem 2.4 as $c\,[c_{\mathrm{opt}}/\eta + c_{\mathrm{var}}\eta + c_{\mathrm{bias}}]$, valid for every $0 < \eta \leq \frac{1}{2\mathcal{L}}$, where:

$$c_{\mathrm{opt}} = \frac{\Delta_{\mathcal{F}}}{K}, \qquad c_{\mathrm{var}} = \frac{4\mathcal{L}G^2\sigma^2}{n_g}, \qquad c_{\mathrm{bias}} = 4G^2(e_h+e_b), \qquad c = 18+12\sqrt{2}.$$

Since the constant c does not depend on η, minimizing this bound amounts to minimizing $\eta \mapsto c_{\mathrm{opt}}/\eta + c_{\mathrm{var}}\eta$, which is convex on $(0,+\infty)$ and attains its minimum at $\sqrt{c_{\mathrm{opt}}/c_{\mathrm{var}}}$, and

$$\sqrt{\frac{c_{\mathrm{opt}}}{c_{\mathrm{var}}}} = \sqrt{\frac{\Delta_{\mathcal{F}}/K}{4\mathcal{L}G^2\sigma^2/n_g}} = \frac{1}{2G\sigma}\sqrt{\frac{\Delta_{\mathcal{F}}n_g}{\mathcal{L}K}} = \eta_\star,$$

so that Equation (22) reads $\eta = \min\{\frac{1}{2\mathcal{L}}, \eta_\star\}$ and is admissible in Theorem 2.4. Since $\frac{1}{\eta} = \max\{2\mathcal{L}, \eta_\star^{-1}\} \leq 2\mathcal{L} + \eta_\star^{-1}$ and $\eta \leq \eta_\star$, we obtain:

$$\frac{c_{\mathrm{opt}}}{\eta} + c_{\mathrm{var}}\eta \ \leq\ 2\mathcal{L}c_{\mathrm{opt}} + \frac{c_{\mathrm{opt}}}{\eta_\star} + c_{\mathrm{var}}\eta_\star \ =\ 2\mathcal{L}c_{\mathrm{opt}} + 2\sqrt{c_{\mathrm{opt}}c_{\mathrm{var}}},$$

where we used $c_{\mathrm{opt}}/\eta_\star = c_{\mathrm{opt}}\sqrt{c_{\mathrm{var}}/c_{\mathrm{opt}}} = \sqrt{c_{\mathrm{opt}}c_{\mathrm{var}}}$ and $c_{\mathrm{var}}\eta_\star = c_{\mathrm{var}}\sqrt{c_{\mathrm{opt}}/c_{\mathrm{var}}} = \sqrt{c_{\mathrm{opt}}c_{\mathrm{var}}}$. Finally:

$$2\mathcal{L}c_{\mathrm{opt}} = \frac{2\mathcal{L}\Delta_{\mathcal{F}}}{K}, \qquad 2\sqrt{c_{\mathrm{opt}}c_{\mathrm{var}}} = 2\sqrt{\frac{\Delta_{\mathcal{F}}}{K}\cdot\frac{4\mathcal{L}G^2\sigma^2}{n_g}} = 4\,G\sigma\sqrt{\frac{\mathcal{L}\Delta_{\mathcal{F}}}{Kn_g}},$$

which, together with c_{bias} and after multiplying by c, gives Equation (23). □

Remark C.7 (Dominant regime). The second branch of Equation (22) is active as soon as $K \geq \mathcal{L}\Delta_{\mathcal{F}}n_g/(G^2\sigma^2)$. In that regime, the ratio of the first term of Equation (23) to the second equals $\frac{1}{2G\sigma}\sqrt{\mathcal{L}\Delta_{\mathcal{F}}n_g/K}$, which is at most $\frac{1}{2}$, so that the algorithmic error is of order $G\sigma\sqrt{\mathcal{L}\Delta_{\mathcal{F}}/(Kn_g)}$. Only the total number Kn_g of hypergradient samples then matters: a given accuracy is reached either with many small batches or with fewer larger ones, as long as K remains above this threshold.

D Generalization Bounds for (OBO) in the NTK Regime

Appendix D proves the bounds on e_h and e_b stated in Proposition 3.1. The bound on e_b follows directly from the single-regression result of Proposition E.25. The bound on e_h is more delicate because of the supremum over ω, and proceeds in four steps: (i) we split the error of $\hat{h}_\omega$ into the gap between the finite-width network and its infinite-width NTK counterpart, which is controlled uniformly in ω since the truncation event does not depend on ω, and the error of the NTK predictor itself; (ii) we show that the NTK predictor is linear in the target vector, so that its error is Lipschitz continuous in ω; (iii) we control the NTK error on a finite net of Ω by a union bound and extend it to all of Ω by Lipschitz continuity; (iv) we choose the confidence level δ to balance the resulting terms.

Throughout Appendix D, we fix a nonempty, compact and convex parameter set $\Omega \subset \mathbb{R}^d$ with diameter $\mathrm{diam}(\Omega)$. We write $\mathbf{K} = [K^{\mathrm{NTK}}(x_i,x_j)]_{i,j=1}^{n_{in}} \in \mathbb{R}^{n_{in}\times n_{in}}$ for the inner infinite-width NTK Gram matrix on the samples of $\mathcal{D}_{\mathrm{in}}$, and λ_0 for its smallest eigenvalue. For $\omega \in \Omega$, let $Y_\omega = [g_\omega(z_1),\dots,g_\omega(z_{n_{in}})]^\top \in \mathbb{R}^{n_{in}}$ denote the inner target vector. The event $\mathcal{E}^h_{\lambda_0} = \{\lambda_0 \geq \lambda_\star\}$ is the

NTK non-degeneracy event of Section 3.1 and Appendix E.8.2, whose threshold $\lambda_\star = \lambda_\star(n_{in}, \delta) > 0$ depends only on n_{in}, δ, p, A and $M_{\mathbb{P}}$. We denote by $h^{\mathrm{NTK}}_{t,\omega}$ the inner NTK gradient flow predictor of $h^\star_\omega$ whose dynamics are described as follows for any $x \in \mathcal{X}$:

$$\frac{\mathrm{d}}{\mathrm{d}t} h^{\mathrm{NTK}}_{t,\omega}(x) = -\frac{1}{n_{in}} \sum_{i=1}^{n_{in}} K^{\mathrm{NTK}}(x, x_i)(h^{\mathrm{NTK}}_{t,\omega}(x_i) - g_\omega(z_i)), \qquad h^{\mathrm{NTK}}_{0,\omega} = 0. \tag{24}$$

For brevity, we state the generalization bounds without writing the exact lower bounds on the inner and outer sample sizes n_{in} and n_{out} required for the results to hold. We refer the reader to Proposition E.25 for a detailed statement including these conditions.

D.1 Lipschitz continuity of the inner NTK approximation error in the outer variable

Lemma D.1 (Linear representation of the inner NTK predictor)**.** *On the event $\mathcal{E}^h_{\lambda_0}$, the inner NTK gradient flow predictor started from the mirrored initialization admits, for every $\omega \in \Omega$ and every $t \geq 0$, the following representation:*

$$h^{\mathrm{NTK}}_{t,\omega} = \Phi_t Y_\omega, \qquad \Phi_t := \frac{1}{n_{in}} K^{\mathrm{NTK}}(\cdot, \boldsymbol{X})^\top \left(\mathbb{1}_{n_{in} \times n_{in}} - \exp\left(-\frac{t}{n_{in}} \mathbf{K} \right) \right) \left(\frac{1}{n_{in}} \mathbf{K} \right)^{-1},$$

where $K^{\mathrm{NTK}}(\cdot, \boldsymbol{X}) = [K^{\mathrm{NTK}}(\cdot, x_1), \dots, K^{\mathrm{NTK}}(\cdot, x_{n_{in}})]^\top$. The operator Φ_t depends only on $\mathbf{K}$, the inputs $x_1, \dots, x_{n_{in}}$, and t; all dependence on ω enters through Y_ω.

Proof. We write $\mathbf{A} := \frac{1}{n_{in}} \mathbf{K}$ and let $H(t) := [h^{\mathrm{NTK}}_{t,\omega}(x_1), \dots, h^{\mathrm{NTK}}_{t,\omega}(x_{n_{in}})]^\top$. Evaluating the kernel gradient flow in Equation (24) at the training points gives the linear inhomogeneous ODE $\dot{H}(t) = -\mathbf{A}(H(t) - Y_\omega)$ with $H(0) = 0$ due to the mirrored initialization. Its unique solution is $H(t) = (\mathbb{1}_{n_{in} \times n_{in}} - \exp(-t\mathbf{A}))Y_\omega$, so the training residual is $H(t) - Y_\omega = -\exp(-t\mathbf{A})Y_\omega$. Substituting into the flow at an arbitrary $x \in \mathcal{X}$, we get:

$$\frac{\mathrm{d}}{\mathrm{d}t} h^{\mathrm{NTK}}_{t,\omega}(x) = -\frac{1}{n_{in}} K^{\mathrm{NTK}}(x, \boldsymbol{X})^\top (H(t) - Y_\omega) = \frac{1}{n_{in}} K^{\mathrm{NTK}}(x, \boldsymbol{X})^\top \exp(-t\mathbf{A}) Y_\omega.$$

The right-hand side does not involve $h^{\mathrm{NTK}}_{t,\omega}(x)$. Integrating from 0 to t with $h^{\mathrm{NTK}}_{0,\omega} = 0$ and using $\int_0^t \exp(-s\mathbf{A})\mathrm{d}s = \mathbf{A}^{-1}(\mathbb{1}_{n_{in} \times n_{in}} - \exp(-t\mathbf{A}))$ (valid since $\mathbf{A}$ is invertible on $\mathcal{E}^h_{\lambda_0}$) yields $h^{\mathrm{NTK}}_{t,\omega}(x) = \frac{1}{n_{in}} K^{\mathrm{NTK}}(x, \boldsymbol{X})^\top \mathbf{A}^{-1}(\mathbb{1}_{n_{in} \times n_{in}} - \exp(-t\mathbf{A}))Y_\omega$, which concludes the proof. □

Lemma D.2 (Operator-norm bound on Φ_t)**.** *Let Assumption C hold. On the event $\mathcal{E}^h_{\lambda_0}$, for every $t \geq 0$, we have:*

$$\|\Phi_t\|_{\ell^2 \to L^2(\mathcal{X})} \leq \frac{\kappa^{\mathrm{NTK}} \sqrt{n_{in}}}{\lambda_0}, \text{ with } \kappa^{\mathrm{NTK}} := (p+2)(A^2+1) + 1,$$

where $\|\Phi_t\|_{\ell^2 \to L^2(\mathcal{X})} := \sup_{Y \neq 0} \|\Phi_t Y\|_{L^2(\mathcal{X})} / \|Y\|$ denotes the operator norm of Φ_t as a map from $(\mathbb{R}^{n_{in}}, \|\cdot\|)$ to $L^2(\mathcal{X})$.

Proof. We factor $\Phi_t = \mathcal{E} M_t$, where $M_t := (\mathbb{1}_{n_{in} \times n_{in}} - \exp(-\frac{t}{n_{in}}\mathbf{K}))(\frac{1}{n_{in}}\mathbf{K})^{-1} \in \mathbb{R}^{n_{in} \times n_{in}}$ and $\mathcal{E} : \mathbb{R}^{n_{in}} \to L^2(\mathcal{X})$ is the kernel-evaluation operator defined by:

$$(\mathcal{E}c)(x) = \frac{1}{n_{in}} \sum_{i=1}^{n_{in}} K^{\mathrm{NTK}}(x, x_i) c_i, \quad \text{for } c \in \mathbb{R}^{n_{in}} \text{ and } x \in \mathcal{X}.$$

Control of $\|M_t\|_{\mathrm{op}}$. $M_t = g_t(\mathbf{K})$ is a function of $\mathbf{K}$ with scalar filter $g_t(\lambda) = \frac{n_{in}}{\lambda}(1 - \exp(-\frac{t}{n_{in}}\lambda))$. Since $\mathbf{K}$ is symmetric positive definite on $\mathcal{E}^h_{\lambda_0}$, $\|M_t\|_{\mathrm{op}} = \max_{1 \leq i \leq n_{in}} g_t(\lambda_i)$ where $\{\lambda_i\}_{i=1}^{n_{in}}$ are the eigenvalues of $\mathbf{K}$. Using $0 \leq 1 - \exp(-a) \leq 1$ provided that $a \geq 0$, we obtain:

$$g_t(\lambda) = \frac{n_{in}}{\lambda}\left(1 - \exp\left(-\frac{t}{n_{in}}\lambda\right)\right) \leq \frac{n_{in}}{\lambda}, \quad \text{hence } \|M_t\|_{\mathrm{op}} \leq \frac{n_{in}}{\min_{1 \leq i \leq n_{in}} \lambda_i} = \frac{n_{in}}{\lambda_0},$$

a bound independent of t.

Control of $\|\mathcal{E}\|_{\ell^2 \to L^2(\mathcal{X})}$. Fix $c \in \mathbb{R}^{n_{in}}$ and $x \in \mathcal{X}$. By the Cauchy-Schwarz inequality and $\left|K^{\mathrm{NTK}}(x, x_i)\right| \leq \kappa^{\mathrm{NTK}}$ (Lemma E.2), we get:

$$\left| \frac{1}{n_{in}} \sum_{i=1}^{n_{in}} K^{\mathrm{NTK}}(x, x_i) c_i \right| \leq \frac{1}{n_{in}} \left\| K^{\mathrm{NTK}}(x, \boldsymbol{X}) \right\| \|c\| \leq \frac{\kappa^{\mathrm{NTK}}}{\sqrt{n_{in}}} \|c\|,$$

where $K^{\mathrm{NTK}}(x, \boldsymbol{X}) = [K^{\mathrm{NTK}}(x, x_1), \ldots, K^{\mathrm{NTK}}(x, x_{n_{in}})]^\top$. Since the marginal on $\mathcal{X}$ is a probability measure, we have:

$$\|\mathcal{E}c\|^2_{L^2(\mathcal{X})} \leq \left(\frac{\kappa^{\mathrm{NTK}}}{\sqrt{n_{in}}} \right)^2 \|c\|^2, \text{ which implies that } \|\mathcal{E}\|_{\ell^2 \to L^2(\mathcal{X})} \leq \frac{\kappa^{\mathrm{NTK}}}{\sqrt{n_{in}}}.$$

By submultiplicativity of the operator norm, we get:

$$\|\Phi_t\|_{\ell^2 \to L^2(\mathcal{X})} \leq \|\mathcal{E}\|_{\ell^2 \to L^2(\mathcal{X})} \|M_t\|_{\mathrm{op}} \leq \frac{\kappa^{\mathrm{NTK}}}{\sqrt{n_{in}}} \cdot \frac{n_{in}}{\lambda_0} = \frac{\kappa^{\mathrm{NTK}} \sqrt{n_{in}}}{\lambda_0}. \qquad \square$$

Lemma D.3 (Lipschitz continuity in ω of the inner NTK error)**.** *Let Assumptions A to C hold. On the event $\mathcal{E}^h_{\lambda_0}$, for any $t \geq 0$, the map $\Psi : \omega \mapsto h^{\mathrm{NTK}}_{t,\omega} - h^\star_\omega$ is L_Ψ-Lipschitz continuous from Ω to $L^2(\mathcal{X})$ with:*

$$L_\Psi := G \left(1 + \frac{\kappa^{\mathrm{NTK}} n_{in}}{\lambda_0} \right), \text{ with } \kappa^{\mathrm{NTK}} := (p+2)(A^2+1) + 1.$$

Proof. Let $t \geq 0$ and $\omega, \omega' \in \Omega$. By the triangle inequality, we have:

$$\|\Psi(\omega) - \Psi(\omega')\|_{L^2(\mathcal{X})} \leq \left\| h^{\mathrm{NTK}}_{t,\omega} - h^{\mathrm{NTK}}_{t,\omega'} \right\|_{L^2(\mathcal{X})} + \|h^\star_\omega - h^\star_{\omega'}\|_{L^2(\mathcal{X})}.$$

For the target term, Proposition F.4 gives $\|h^\star_\omega - h^\star_{\omega'}\|_{L^2(\mathcal{X})} \leq G \|\omega - \omega'\|$. For the estimator term, Lemma D.1 and the linearity of Φ_t give $h^{\mathrm{NTK}}_{t,\omega} - h^{\mathrm{NTK}}_{t,\omega'} = \Phi_t(Y_\omega - Y_{\omega'})$, hence by Lemma D.2:

$$\left\| h^{\mathrm{NTK}}_{t,\omega} - h^{\mathrm{NTK}}_{t,\omega'} \right\|_{L^2(\mathcal{X})} \leq \|\Phi_t\|_{\ell^2 \to L^2(\mathcal{X})} \|Y_\omega - Y_{\omega'}\| \leq \frac{\kappa^{\mathrm{NTK}} \sqrt{n_{in}}}{\lambda_0} \|Y_\omega - Y_{\omega'}\|.$$

Since the open set $U \supset \Omega$ of Assumption A is convex, the segment $\omega_s := \omega' + s(\omega - \omega')$, $s \in [0,1]$, lies in U. Hence, applying the Cauchy-Schwarz inequality and Assumption A, we get for each $1 \leq i \leq n_{in}$:

$$|g_\omega(z_i) - g_{\omega'}(z_i)| = \left| \int_0^1 \langle \partial_\omega g_{\omega_s}(z_i), \omega - \omega' \rangle \, \mathrm{d}s \right| \leq G \|\omega - \omega'\|,$$

and therefore $\|Y_\omega - Y_{\omega'}\| \leq \sqrt{n_{in}} G \|\omega - \omega'\|$. Combining both, we get:

$$\left\| h^{\mathrm{NTK}}_{t,\omega} - h^{\mathrm{NTK}}_{t,\omega'} \right\|_{L^2(\mathcal{X})} \leq \frac{\kappa^{\mathrm{NTK}} n_{in}}{\lambda_0} G \|\omega - \omega'\|.$$

Adding the two contributions concludes the proof. $\square$

D.2 Generalization bounds for the inner solution, regression and adjoint functions

Proposition D.4 (Uniform-in-ω expected error in the inner solution)**.** *Consider n_{in} i.i.d. samples $\{(x_i, z_i)\}_{i=1}^{n_{in}} \sim \mathbb{P}_{XZ}$ and assume $p \geq 2$. Fix a compact $\Omega \subset \mathbb{R}^d$ and let Assumptions A and C to E hold. If we choose a stopping time $t_{\mathrm{op}_h} = n_{in}^{\frac{p+1}{s_h(p+1)+p}}$ with $\frac{1}{p+1} < s_h \leq 2$, then, considering the early-stopped truncated neural network predictor $\hat{h}_\omega$ for the inner solution $h^\star_\omega$, there exist a positive constant $C_1 > 0$ and a polynomial poly_h such that when n_{in} is sufficiently large and the width satisfies $m_h \geq \mathrm{poly}_h(n_{in}, \ln n_{in})$, the following holds:*

$$\mathbb{E}\left[\sup_{\omega \in \Omega} \left\| \hat{h}_\omega - h^\star_\omega \right\|^2_{L^2(\mathcal{X})} \right] \leq C_1 d^4 (\ln n_{in})^4 n_{in}^{-\frac{s_h(p+1)}{s_h(p+1)+p}},$$

where the constant $C_1 > 0$ is independent of n_{in} and t_{op_h}, but depends on s_h, p, A, D, G, $M_{\mathbb{P}}$, $\mathrm{diam}(\Omega)$, and R_h, and the coefficients of poly_h depend on s_h, A, D, G, $M_{\mathbb{P}}$, $\mathrm{diam}(\Omega)$, p, and R_h. The expectation here is with respect to random samples and random initialization.

Proof. Throughout, we write $r_h \coloneqq \frac{1}{2} s_h \gamma_h$ with $\gamma_h = \frac{p+1}{s_h(p+1)+p}$. By the triangle inequality and the identity $(a+b)^2 \le 2(a^2+b^2)$, we have:

$$\mathbb{E}\left[\sup_{\omega\in\Omega}\left\|\hat{h}_\omega - h^\star_\omega\right\|^2_{L^2(\mathcal{X})}\right] \le 2\Bigg(\mathbb{E}\left[\sup_{\omega\in\Omega}\left\|\hat{h}_\omega - h^{\mathrm{NTK}}_{t_{\mathrm{op}_h},\omega}\right\|^2_{L^2(\mathcal{X})}\right] + \mathbb{E}\left[\sup_{\omega\in\Omega}\left\|h^{\mathrm{NTK}}_{t_{\mathrm{op}_h},\omega} - h^\star_\omega\right\|^2_{L^2(\mathcal{X})}\right]\Bigg). \tag{25}$$

Upper-bound on $\mathbb{E}\left[\sup_{\omega\in\Omega}\left\|\hat{h}_\omega - h^{\mathrm{NTK}}_{t_{\mathrm{op}_h},\omega}\right\|^2_{L^2(\mathcal{X})}\right]$. The bound on this term is obtained by following the first part of the proof of Proposition E.25, the only difference being the presence of the supremum over Ω. The supremum is harmless here for two reasons. First, the favorable event $\mathcal{E}^h$ is independent of ω: the truncated predictor $\hat{h}_\omega$ therefore vanishes on $(\mathcal{E}^h)^{\complement}$ for all $\omega \in \Omega$ at once, so the decomposition over $\mathcal{E}^h$ and $(\mathcal{E}^h)^{\complement}$ and the ensuing control of $\Pr((\mathcal{E}^h)^{\complement})$ are unchanged. Second, the outer variable enters the argument only through the uniform target bound $|g_\omega(z_i)| \le D$ (Assumption A), so all constants remain ω-independent and each estimate holds uniformly over Ω. We thus obtain, exactly as in Proposition E.25:

$$\mathbb{E}\left[\sup_{\omega\in\Omega}\left\|\hat{h}_\omega - h^{\mathrm{NTK}}_{t_{\mathrm{op}_h},\omega}\right\|^2_{L^2(\mathcal{X})}\right] \le c_1^2 m_h^{-\frac{1}{12}} \ln m_h + 4(\kappa^{\mathrm{NTK}})^2 D^2 n_{in}^{2\gamma_h}\delta, \tag{26}$$

where $\delta \in (0,1)$, $\kappa^{\mathrm{NTK}} = (p+2)(A^2+1)+1$, and $c_1 > 0$ is a constant that is independent of δ, n_{in}, and t_{op_h}, but depends on A. We note that Equation (26) is valid whenever $m_h \ge \mathrm{poly}(n_{in}, 1/\delta, \ln n_{in}, \ln(1/\delta))$, where the coefficients of poly depend on A, D, $M_{\mathbb{P}}$, and p.

Upper-bound on $\mathbb{E}\left[\sup_{\omega\in\Omega}\left\|h^{\mathrm{NTK}}_{t_{\mathrm{op}_h},\omega} - h^\star_\omega\right\|^2_{L^2(\mathcal{X})}\right]$. Let $\lambda_\star \coloneqq \lambda_\star(n_{in},\delta) > 0$ denote the threshold in the definition of $\mathcal{E}^h_{\lambda_0}$, so that $\mathcal{E}^h_{\lambda_0} = \{\lambda_0 \ge \lambda_\star\}$. It depends only on n_{in}, δ, p, A and $M_{\mathbb{P}}$, and not on $\mathcal{D}_{in}$. Define the deterministic constants:

$$L^\star_\Psi \coloneqq G\left(1 + \frac{\kappa^{\mathrm{NTK}} n_{in}}{\lambda_\star}\right) \qquad \text{and} \qquad \varepsilon \coloneqq (L^\star_\Psi)^{-1} n_{in}^{-r_h}.$$

On $\mathcal{E}^h_{\lambda_0}$, the Lipschitz constant L_Ψ of Lemma D.3 satisfies $L_\Psi \le L^\star_\Psi$, since $\lambda_0 \ge \lambda_\star$ there. Let $\mathcal{N} = \{\omega_1, \dots, \omega_N\} \subset \Omega$ be a maximal ε-separated subset of Ω, *i.e.*, $\|\omega_j - \omega_k\| \ge \varepsilon$ for all $j \ne k$ and no point of Ω can be added to $\mathcal{N}$ without violating this property. By maximality, every $\omega \in \Omega$ satisfies $\|\omega - \omega_k\| < \varepsilon$ for some $1 \le k \le N$, so $\mathcal{N}$ is an ε-net of Ω whose points lie in Ω. In particular, Lemma D.3 applies to every pair (ω, ω_k). Note that since ε is deterministic, $\mathcal{N}$ and N are fixed before the data are drawn, and are therefore independent of $\mathcal{D}_{in}$. Let also $\mathcal{E}_{\mathrm{net}} \coloneqq \bigcap_{k=1}^N \mathcal{E}^{\mathrm{NTK}}_{\omega_k}$, where, for each $1 \le k \le N$, $\mathcal{E}^{\mathrm{NTK}}_{\omega_k}$ is the event of Lemma E.21 at ω_k with deviation parameter δ/N, namely:

$$\mathcal{E}^{\mathrm{NTK}}_{\omega_k} \coloneqq \left\{\left\|h^{\mathrm{NTK}}_{t_{\mathrm{op}_h},\omega_k} - h^\star_{\omega_k}\right\|_{L^2(\mathcal{X})} \le c_2\left(\ln\frac{6N}{\delta}\right)^2 n_{in}^{-r_h}\right\},$$

where $c_2 > 0$ is a constant that is independent of δ, $n_{in}, t_{\mathrm{op}_h}$, and ω_k, but depends on s_h, p, A, D, $M_{\mathbb{P}}$, and R_h. Since ω_k is deterministic, Lemma E.21 applies to the regression problem with target $h^\star_{\omega_k}$ and gives $\Pr\big((\mathcal{E}^{\mathrm{NTK}}_{\omega_k})^{\complement}\big) \le \delta/N$. Therefore, by a union bound, $\Pr(\mathcal{E}^{\complement}_{\mathrm{net}}) \le \delta$. On $\mathcal{E}_{\mathrm{net}} \cap \mathcal{E}^h_{\lambda_0}$, for every $\omega \in \Omega$ and its nearest net point ω_k, *i.e.*, $\|\omega - \omega_k\| \le \varepsilon$, applying Lemma D.3 together with $L_\Psi \le L^\star_\Psi$ gives:

$$\left\|h^{\mathrm{NTK}}_{t_{\mathrm{op}_h},\omega} - h^\star_\omega\right\|_{L^2(\mathcal{X})} \le \left\|h^{\mathrm{NTK}}_{t_{\mathrm{op}_h},\omega_k} - h^\star_{\omega_k}\right\|_{L^2(\mathcal{X})} + L^\star_\Psi\varepsilon \le c_2\left(\ln\frac{6N}{\delta}\right)^2 n_{in}^{-r_h} + n_{in}^{-r_h},$$

by the choice $L_\Psi^\star \varepsilon = n_{in}^{-r_h}$. Setting $\mathscr{F} \coloneqq \mathcal{E}_{\text{net}} \cap \mathcal{E}_{\lambda_0}^h$, so that $\Pr(\mathscr{F}^\complement) \le 2\delta$, decomposing over $\mathscr{F}$ and $\mathscr{F}^\complement$, and using the identity $(a+b)^2 \le 2(a^2+b^2)$, we obtain:

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\omega\in\Omega}\left\|h^{\text{NTK}}_{t_{\text{op}_h},\omega} - h^\star_\omega\right\|^2_{L^2(\mathcal{X})}\right] \\
=&\mathbb{E}\left[\sup_{\omega\in\Omega}\left\|h^{\text{NTK}}_{t_{\text{op}_h},\omega} - h^\star_\omega\right\|^2_{L^2(\mathcal{X})} \mathbb{1}_{\mathscr{F}}\right] + \mathbb{E}\left[\sup_{\omega\in\Omega}\left\|h^{\text{NTK}}_{t_{\text{op}_h},\omega} - h^\star_\omega\right\|^2_{L^2(\mathcal{X})} \mathbb{1}_{\mathscr{F}^\complement}\right] \\
\le& 2c_2^2\left(\ln\frac{6N}{\delta}\right)^4 n_{in}^{-2r_h} + 2n_{in}^{-2r_h} + 4\left((\kappa^{\text{NTK}})^2 D^2 n_{in}^{2\gamma_h} + D^2\right)\delta, \qquad (27)
\end{aligned}$$

where we apply Lemmas E.15 and F.3 on the event $\mathscr{F}^\complement$ to bound $\left\|h^{\text{NTK}}_{t_{\text{op}_h},\omega}\right\|^2_{L^2(\mathcal{X})}$ by $(\kappa^{\text{NTK}})^2 D^2 n_{in}^{2\gamma_h}$ and $\|h^\star_\omega\|^2_{L^2(\mathcal{X})}$ by D^2 for any $\omega \in \Omega$.

Expected generalization bound. Plugging Equations (26) and (27) into Equation (25) and recalling that $n_{in}^{-2r_h} = n_{in}^{-s_h\gamma_h}$, we obtain:

$$\mathbb{E}\left[\sup_{\omega\in\Omega}\left\|\hat{h}_\omega - h^\star_\omega\right\|^2_{L^2(\mathcal{X})}\right] \le 2c_1^2 m_h^{-\frac{1}{12}} \ln m_h + 4c_2^2\left(\ln\frac{6N}{\delta}\right)^4 n_{in}^{-s_h\gamma_h} + 4n_{in}^{-s_h\gamma_h} + a_1 n_{in}^{2\gamma_h}\delta, \quad (28)$$

where $a_1 \coloneqq 16(\kappa^{\text{NTK}})^2 D^2 + 8D^2$ collects the three δ-terms, using $n_{in}^{2\gamma_h} \ge 1$. For each fixed $\delta \in (0,1)$, the net above is deterministic, so Equation (28) holds for every $\delta \in (0,1)$. We now control the δ-dependent terms. The balls $B(\omega_k, \varepsilon/2)$, $1 \le k \le N$, are pairwise disjoint and are all contained in $B(\omega_1, \operatorname{diam}(\Omega) + \varepsilon/2)$, since $\|\omega_k - \omega_1\| \le \operatorname{diam}(\Omega)$. Comparing volumes, and using $\varepsilon \le \operatorname{diam}(\Omega)$ for n_{in} large, we obtain by [89, Proposition 4.2.10]:

$$N \le \left(1 + \frac{2\operatorname{diam}(\Omega)}{\varepsilon}\right)^d \le \left(\frac{3\operatorname{diam}(\Omega)}{\varepsilon}\right)^d, \text{ hence } \ln N \le d\Big[\ln\big(3\operatorname{diam}(\Omega)\big) + \ln L_\Psi^\star + r_h \ln n_{in}\Big].$$

By Corollary E.10, the deterministic threshold $\lambda_\star$ satisfies:

$$\lambda_\star^{-1} \le c_3 n_{in}^{\frac{8}{p}} \delta^{-\frac{4}{p}} \left(\ln n_{in} + \ln\frac{1}{\delta}\right)^3,$$

where $c_3 > 0$ depends only on p, A and $M_{\mathbb{P}}$. Since $L_\Psi^\star = G(1 + \kappa^{\text{NTK}} n_{in}/\lambda_\star) \le 2G\kappa^{\text{NTK}} n_{in}\, \lambda_\star^{-1}$ for n_{in} large, taking logarithms, and using the bound on $\lambda_\star^{-1}$ above and the inequality $\ln y \le y$ for $y > 0$ to bound the last term, we get:

$$\begin{aligned}
\ln L_\Psi^\star &\le \ln\left(2G\kappa^{\text{NTK}}\right) + \ln n_{in} + \ln \lambda_\star^{-1} \\
&\le \ln\left(2G\kappa^{\text{NTK}} c_3\right) + \left(1 + \frac{8}{p}\right)\ln n_{in} + \frac{4}{p}\ln\frac{1}{\delta} + 3\ln\left(\ln n_{in} + \ln\frac{1}{\delta}\right) \\
&\le C_\Psi\left(\ln n_{in} + \ln\frac{1}{\delta}\right),
\end{aligned}$$

with $C_\Psi \coloneqq \left|\ln(2G\kappa^{\text{NTK}} c_3)\right| + 4 + \frac{8}{p} > 0$, where the last inequality uses $\ln n_{in} \ge 1$ for $n_{in} \ge 3$. Using also $r_h \le 1$ and $\ln n_{in} \ge 1$, we obtain:

$$\ln N \le cd\left(\ln n_{in} + \ln\frac{1}{\delta}\right), \qquad \text{with} \qquad c \coloneqq \left|\ln(3\operatorname{diam}(\Omega))\right| + C_\Psi + 1.$$

This implies that $\ln\frac{6N}{\delta} = \ln 6 + \ln N + \ln\frac{1}{\delta} \le c' d\big(\ln n_{in} + \ln\frac{1}{\delta}\big)$ with $c' \coloneqq c + 2$, using $\ln 6 \le 2\ln n_{in}$ for $n_{in} \ge 3$ and $d \ge 1$, where $c' > 0$ is a constant depending on $\operatorname{diam}(\Omega)$, p, A, $M_{\mathbb{P}}$, and G. Hence, using the identity $(u+v)^4 \le 8(u^4+v^4)$ and $\ln\frac{1}{\delta} \le \ln\frac{6}{\delta}$, we obtain:

$$\left(\ln\frac{6N}{\delta}\right)^4 \le 8c'^4 d^4\left[(\ln n_{in})^4 + \left(\ln\frac{1}{\delta}\right)^4\right] \le 8c'^4 d^4\left[(\ln n_{in})^4 + \left(\ln\frac{6}{\delta}\right)^4\right].$$

Substituting into Equation (28), the δ-dependent part of the bound is at most $\varphi(\delta)$, where:

$$\varphi(\delta) \coloneqq a_2 d^4\left(\ln\frac{6}{\delta}\right)^4 n_{in}^{-s_h\gamma_h} + a_1 n_{in}^{2\gamma_h}\delta, \qquad a_2 \coloneqq 32c_2^2 c'^4,$$

and the remaining, δ-free part is $2c_1^2 m_h^{-\frac{1}{12}} \ln m_h + \big(a_2 d^4 (\ln n_{in})^4 + 4\big) n_{in}^{-s_h\gamma_h}$. Minimizing φ over $\delta \in (0,1)$ exactly as in the proof of Proposition E.25 gives:

$$\delta^\star = \frac{108 a_2 d^4}{a_1} n_{in}^{-(s_h+2)\gamma_h} W_0^3(z_h), \qquad z_h := \left(\frac{a_1}{18 a_2 d^4}\right)^{\frac{1}{3}} n_{in}^{\frac{1}{3}(s_h+2)\gamma_h}, \tag{29}$$

and $\varphi(\delta^\star) = 27 a_2 d^4 W_0^3(z_h)\big(3W_0(z_h)+4\big) n_{in}^{-s_h\gamma_h}$, where W_0 is the principal branch of the Lambert W function. Since $d \geq 1$, we have $z_h \leq (a_1/(18a_2))^{1/3} n_{in}^{(s_h+2)\gamma_h/3}$, so Lemma I.3 bounds $W_0^3(z_h)(3W_0(z_h)+4)$ by a constant multiple of $(\ln n_{in})^4$ that does not involve d. This yields the rate $d^4 (\ln n_{in})^4 n_{in}^{-s_h\gamma_h}$. For the width, note that taking n_{in} large enough so that $n_{in}^{(s_h+2)\gamma_h} \geq 18 e^3 a_2 d^4 / a_1$ gives $z_h \geq e$, hence $W_0(z_h) \geq 1$ and

$$\delta^\star \geq \frac{108\, a_2 d^4}{a_1} n_{in}^{-(s_h+2)\gamma_h} \geq \frac{108\, a_2}{a_1} n_{in}^{-(s_h+2)\gamma_h},$$

so that, using $(s_h+2)\gamma_h \leq 4$ and $\ln n_{in} \geq 1$, we obtain:

$$\frac{1}{\delta^\star} \leq \frac{a_1}{108 a_2} n_{in}^4, \qquad \ln \frac{1}{\delta^\star} \leq \left(\left|\ln \frac{a_1}{108 a_2}\right| + 4\right) \ln n_{in}.$$

Both bounds are free of d, so the condition $m_h \geq \mathrm{poly}_h(n_{in}, 1/\delta^\star, \ln n_{in}, \ln(1/\delta^\star))$ is implied by $m_h \geq \mathrm{poly}_h(n_{in}, \ln n_{in})$. As a consequence, we obtain:

$$\mathbb{E}\left[\sup_{\omega\in\Omega} \left\|\hat{h}_\omega - h_\omega^\star\right\|_{L^2(\mathcal{X})}^2\right] \leq C_1 d^4 (\ln n_{in})^4 n_{in}^{-s_h\gamma_h},$$

provided that $m_h \geq \mathrm{poly}_h(n_{in}, \ln n_{in})$, where $C_1 > 0$ is a constant independent of n_{in} and t_{op_h}, but depends on s_h, p, A, D, G, $M_{\mathbb{P}}$, $\mathrm{diam}(\Omega)$, and R_h, and the coefficients of poly_h depend on s_h, A, D, G, $M_{\mathbb{P}}$, $\mathrm{diam}(\Omega)$, p, and R_h. This concludes the proof. □

Proposition D.5 (Expected error in the regression function)**.** *Consider n_{out} i.i.d. samples $\{(\tilde{x}_j, \tilde{y}_j)\}_{j=1}^{n_{out}} \sim \mathbb{P}_{XY}$ and assume $p \geq 2$. Let Assumptions B to D and F hold. If we choose a stopping time $t_{\mathrm{op}_b} = n_{out}^{\frac{p+1}{s_b(p+1)+p}}$ with $\frac{1}{p+1} < s_b \leq 2$, then, considering the early-stopped truncated neural network predictor $\hat{b}$ for the regression function $b^\star$, there exist a positive constant $C_2 > 0$ and a polynomial poly_b such that when n_{out} is sufficiently large and the width satisfies $m_b \geq \mathrm{poly}_b(n_{out}, \ln n_{out})$, the following holds:*

$$\mathbb{E}\left[\left\|\hat{b} - b^\star\right\|_{L^2(\mathcal{X})}^2\right] \leq C_2 \left(\ln n_{out}\right)^4 n_{out}^{-\frac{s_b(p+1)}{s_b(p+1)+p}},$$

where $C_2 > 0$ is a constant that is independent of n_{out} and t_{op_b}, but depends on s_b, p, A, B, $M_{\mathbb{P}}$, and R_b, and the coefficients of poly_b depend on s_b, A, B, $M_{\mathbb{P}}$, R_b, and p. The expectation here is with respect to random samples and random initialization.

Proof. This is a direct consequence of Proposition E.25, where the neural network approximating $b^\star$ of the form in Equation (5) is trained using the empirical loss $\widehat{L}_{out}$, defined for any $h \in L^2(\mathcal{X})$ by:

$$\widehat{L}_{out}(h) = \frac{1}{2n_{out}} \sum_{j=1}^{n_{out}} \left(h(\tilde{x}_j) - \tilde{y}_j\right)^2.$$ □

Proposition D.6 (Uniform-in-ω expected error in the adjoint function)**.** *Consider n_{in} i.i.d. samples $\{(x_i, z_i)\}_{i=1}^{n_{in}} \sim \mathbb{P}_{XZ}$, n_{out} i.i.d. samples $\{(\tilde{x}_j, \tilde{y}_j)\}_{j=1}^{n_{out}} \sim \mathbb{P}_{XY}$, and assume $p \geq 2$. Let Assumptions A to F hold. If we choose stopping times $t_{\mathrm{op}_h} = n_{in}^{\frac{p+1}{s_h(p+1)+p}}$ and $t_{\mathrm{op}_b} = n_{out}^{\frac{p+1}{s_b(p+1)+p}}$ with $\frac{1}{p+1} < s_h \leq 2$ and $\frac{1}{p+1} < s_b \leq 2$, then, for every $\omega \in \Omega$, considering the early-stopped truncated neural network predictors $\hat{h}_\omega$ and $\hat{b}$ for the inner solution $h_\omega^\star$ and the regression function $b^\star$, respectively, there exist two positive constants $C_1, C_2 > 0$ and two polynomials poly_h and poly_b*

such that when n_{in} and n_{out} are sufficiently large and the widths satisfy $m_h \geq \mathrm{poly}_h(n_{in}, \ln n_{in})$ and $m_b \geq \mathrm{poly}_b(n_{out}, \ln n_{out})$, the following holds:

$$\mathbb{E}\left[\sup_{\omega\in\Omega} \|\hat{a}_\omega - a^\star_\omega\|^2_{L^2(\mathcal{X})}\right] \leq 2C_1 d^4 \left(\ln n_{in}\right)^4 n_{in}^{-\frac{s_h(p+1)}{s_h(p+1)+p}} + 2C_2 \left(\ln n_{out}\right)^4 n_{out}^{-\frac{s_b(p+1)}{s_b(p+1)+p}},$$

where $\hat{a}_\omega = -\hat{h}_\omega + \hat{b}$ and $a^\star_\omega = -h^\star_\omega + b^\star$. $C_1, C_2 > 0$ and $\mathrm{poly}_h, \mathrm{poly}_b$ are the same as in Propositions D.4 and D.5, respectively. The expectation here is with respect to random samples and random initialization.

Proof. Let $\omega \in \Omega$. Using the triangle inequality and the identity $(a+b)^2 \leq 2(a^2+b^2)$, we have:

$$\begin{aligned}\|\hat{a}_\omega - a^\star_\omega\|^2_{L^2(\mathcal{X})} &= \left\|\left(-\hat{h}_\omega + \hat{b}\right) - \left(-h^\star_\omega + b^\star\right)\right\|^2_{L^2(\mathcal{X})} \\ &\leq 2\left(\left\|\hat{h}_\omega - h^\star_\omega\right\|^2_{L^2(\mathcal{X})} + \left\|\hat{b} - b^\star\right\|^2_{L^2(\mathcal{X})}\right).\end{aligned}$$

Taking the supremum over $\omega \in \Omega$ followed by the expectation over the random samples and random initialization on both sides, we obtain:

$$\mathbb{E}\left[\sup_{\omega\in\Omega} \|\hat{a}_\omega - a^\star_\omega\|^2_{L^2(\mathcal{X})}\right] \leq 2\left(\mathbb{E}\left[\sup_{\omega\in\Omega}\left\|\hat{h}_\omega - h^\star_\omega\right\|^2_{L^2(\mathcal{X})}\right] + \mathbb{E}\left[\left\|\hat{b} - b^\star\right\|^2_{L^2(\mathcal{X})}\right]\right). \tag{30}$$

By Proposition D.4, we know that there exist a positive constant $C_1 > 0$ and a polynomial poly_h such that when n_{in} is sufficiently large and $m_h \geq \mathrm{poly}_h(n_{in}, \ln n_{in})$, the following holds:

$$\mathbb{E}\left[\sup_{\omega\in\Omega}\left\|\hat{h}_\omega - h^\star_\omega\right\|^2_{L^2(\mathcal{X})}\right] \leq C_1 d^4 \left(\ln n_{in}\right)^4 n_{in}^{-\frac{s_h(p+1)}{s_h(p+1)+p}},$$

where the exact dependence of C_1 and the coefficients of poly_h are specified in Proposition D.4. Additionally, by Proposition D.5, we know that there exist a positive constant $C_2 > 0$ and a polynomial poly_b such that when n_{out} is sufficiently large and $m_b \geq \mathrm{poly}_b(n_{out}, \ln n_{out})$, the following holds:

$$\mathbb{E}\left[\left\|\hat{b} - b^\star\right\|^2_{L^2(\mathcal{X})}\right] \leq C_2 \left(\ln n_{out}\right)^4 n_{out}^{-\frac{s_b(p+1)}{s_b(p+1)+p}},$$

where the exact dependence of $C_2 > 0$ and the coefficients of poly_b are provided in Proposition D.5. After combining both upper-bounds in Equation (30), we get:

$$\mathbb{E}\left[\sup_{\omega\in\Omega} \|\hat{a}_\omega - a^\star_\omega\|^2_{L^2(\mathcal{X})}\right] \leq 2C_1 d^4 \left(\ln n_{in}\right)^4 n_{in}^{-\frac{s_h(p+1)}{s_h(p+1)+p}} + 2C_2 \left(\ln n_{out}\right)^4 n_{out}^{-\frac{s_b(p+1)}{s_b(p+1)+p}}. \qquad \square$$

E Generalization Analysis of Neural Network Gradient Flow in the NTK Regime

E.1 Problem setting

Let p be a fixed positive integer, $\mathcal{X} \subset \mathbb{R}^p$, $\mathcal{Y} \subset \mathbb{R}$, and ρ be a *fixed* but *unknown* probability distribution supported on $\mathcal{X} \times \mathcal{Y}$. We denote by μ the marginal distribution of ρ on $\mathcal{X}$, which we assume is *absolutely continuous* with respect to the Lebesgue measure on $\mathbb{R}^p$, *i.e.*, it admits a *density function* f_μ. Let $L^2(\mathcal{X}, \mu)$ and $L^\infty(\mathcal{X})$ denote respectively the spaces of *square-integrable* with respect to μ and *essentially bounded* functions defined on the input space $\mathcal{X}$ and taking values in $\mathcal{Y}$, endowed with the following norms:

$$\|f\|_{L^2(\mathcal{X},\mu)} = \left(\mathbb{E}_{x\sim\mu}\left[|f(x)|^2\right]\right)^{\frac{1}{2}}, \quad \|f\|_\infty = \operatorname*{ess\,sup}_{x\in\mathcal{X}} |f(x)|.$$

For conciseness, we write $L^2(\mathcal{X})$ for $L^2(\mathcal{X}, \mu)$. Suppose we are given a training set of n *i.i.d.* samples $\{(x_i, y_i)\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$ drawn from the distribution ρ. Using those n samples, our goal is to *learn a neural network* that approximates the regression function $f^\star : \mathcal{X} \to \mathcal{Y}$ in $L^2(\mathcal{X})$, defined for any $x \in \mathcal{X}$ by $f^\star(x) = \mathbb{E}_{y\sim\rho}[y \mid x]$, which minimizes the following *population loss* over $L^2(\mathcal{X})$:

$$L(f) = \frac{1}{2}\mathbb{E}_{(x,y)\sim\rho}\left[|f(x) - y|^2\right].$$

Throughout Appendix E, we denote by $\boldsymbol{X} = [x_1, \ldots, x_n]^\top \in \mathbb{R}^{n\times p}$ the design matrix, by $\boldsymbol{Y} = [y_1, \ldots, y_n]^\top \in \mathbb{R}^n$ the target vector, and by $\mathbb{S}^p = \{x \in \mathbb{R}^{p+1} \mid \|x\| = 1\}$ the unit sphere in $\mathbb{R}^{p+1}$.

E.2 Neural network architecture and infinite-width training dynamics

Neural network architecture. We consider a *two-layer* ReLU neural network predictor f_θ^{NN} for $f^\star$, defined for any $x \in \mathcal{X}$ by:

$$f_\theta^{\mathrm{NN}}(x) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} v_i \sigma(\tilde{\boldsymbol{u}}_i^\top \tilde{x}) + a, \quad \text{with parameters } \theta = (\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{b}, a) \in \Theta, \tag{31}$$

where the augmented input is $\tilde{x} = [x^\top, 1]^\top \in \mathbb{R}^{p+1}$ and the augmented inner weights $\tilde{\boldsymbol{u}}_i \in \mathbb{R}^{p+1}$ incorporate a neuron-wise bias b_i via:

$$\tilde{\boldsymbol{u}}_i = \begin{bmatrix} \boldsymbol{u}_i \\ b_i \end{bmatrix} \in \mathbb{R}^{p+1},$$

where $\boldsymbol{u} = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_m]^\top \in \mathbb{R}^{m \times p}$ are the inner layer weights, $\boldsymbol{b} = [b_1, \dots, b_m]^\top \in \mathbb{R}^m$ are the inner layer biases, $\boldsymbol{v} = [v_1, \dots, v_m]^\top \in \mathbb{R}^m$ are the outer layer weights, $a \in \mathbb{R}$ is the outer bias, $\Theta = \mathbb{R}^m \times \mathbb{R}^{m \times p} \times \mathbb{R}^m \times \mathbb{R}$ is the parameter space, and σ is the ReLU activation function, defined as $\sigma(x) = \max(0, x)$. The derivative of σ is $\sigma'(x) = 1$ for $x > 0$ and $\sigma'(x) = 0$ for $x < 0$. Since σ is not differentiable at 0, we adopt the convention $\sigma'(0) = 0$; hence we write $\sigma'(x) = \mathbb{1}_{x>0}$.

Initialization. We use a *mirrored initialization* to ensure that the network output is exactly zero at initialization. Specifically, the m neurons are arranged in $m/2$ pairs (assuming m is even). For each pair $(i, i + m/2)$ with $1 \le i \le m/2$, the inner weights and biases are initialized symmetrically:

$$\boldsymbol{u}_i = \boldsymbol{u}_{i+m/2} \sim \mathcal{N}(0, \mathbb{1}_{p \times p}), \qquad b_i = b_{i+m/2} \sim \mathcal{N}(0, 1),$$

and the outer weights are initialized with opposite signs:

$$v_i = -v_{i+m/2} \sim \mathcal{N}(0, 1).$$

The outer bias is initialized as $a = 0$. Moreover, $v_i, \boldsymbol{u}_i$, and b_i are drawn *independently* from one another. Under this initialization, for each pair $(i, i + m/2)$ and any input $x \in \mathcal{X}$:

$$v_i \, \sigma(\tilde{\boldsymbol{u}}_i^\top \tilde{x}) + v_{i+m/2} \, \sigma(\tilde{\boldsymbol{u}}_{i+m/2}^\top \tilde{x}) = v_i \, \sigma(\tilde{\boldsymbol{u}}_i^\top \tilde{x}) - v_i \, \sigma(\tilde{\boldsymbol{u}}_i^\top \tilde{x}) = 0.$$

Summing over all pairs, the network output at initialization satisfies $f_\theta^{\mathrm{NN}}(x) = 0$ for all $x \in \mathcal{X}$.

Training dynamics. Given a training set of n *i.i.d.* samples $\{(x_i, y_i)\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$ drawn from the distribution ρ, we consider the following *empirical loss*:

$$\widehat{L}(f_\theta^{\mathrm{NN}}) \coloneqq \widehat{\mathscr{L}}(\theta) = \frac{1}{2n} \sum_{i=1}^{n} \left(f_\theta^{\mathrm{NN}}(x_i) - y_i \right)^2,$$

and we train the network by *gradient flow*, where the parameters are updated according to the following ordinary differential equation:

$$\frac{\mathrm{d}}{\mathrm{d}t} \theta_t = -\partial_\theta \widehat{\mathscr{L}}(\theta_t) = -\frac{1}{n} \sum_{i=1}^{n} \partial_\theta f_{\theta_t}^{\mathrm{NN}}(x_i) \left(f_{\theta_t}^{\mathrm{NN}}(x_i) - y_i \right),$$

where $\theta_t = (\boldsymbol{v}_t, \boldsymbol{u}_t, \boldsymbol{b}_t, a_t)$ and $\partial_\theta f_{\theta_t}^{\mathrm{NN}}(x_i) \in \mathbb{R}^{\dim(\Theta)}$ denotes the Jacobian. Using the chain rule, for any $x \in \mathcal{X}$, the network's output evolves as follows:

$$\frac{\mathrm{d}}{\mathrm{d}t} f_{\theta_t}^{\mathrm{NN}}(x) = \left(\partial_\theta f_{\theta_t}^{\mathrm{NN}}(x) \right)^\top \frac{\mathrm{d}}{\mathrm{d}t} \theta_t = -\frac{1}{n} \sum_{i=1}^{n} \left(\partial_\theta f_{\theta_t}^{\mathrm{NN}}(x) \right)^\top \partial_\theta f_{\theta_t}^{\mathrm{NN}}(x_i) \left(f_{\theta_t}^{\mathrm{NN}}(x_i) - y_i \right).$$

We define the *finite-width Neural Tangent Kernel* (NTK) K_θ^m for any $x, x' \in \mathcal{X}$ by:

$$K_\theta^m(x, x') = \left(\partial_\theta f_\theta^{\mathrm{NN}}(x) \right)^\top \partial_\theta f_\theta^{\mathrm{NN}}(x') \in \mathbb{R}. \tag{32}$$

Using this, the output dynamics becomes:

$$\frac{\mathrm{d}}{\mathrm{d}t} f_{\theta_t}^{\mathrm{NN}}(x) = -\frac{1}{n} \sum_{i=1}^{n} K_{\theta_t}^m(x, x_i) \left(f_{\theta_t}^{\mathrm{NN}}(x_i) - y_i \right). \tag{33}$$

For the neural network defined above in Equation (31), K_θ^m admits an explicit expression. Indeed, computing the gradients with respect to the parameters for any $1 \leq i \leq m$, we get:

$$\partial_{v_i} f_\theta^{\mathrm{NN}}(x) = \frac{1}{\sqrt{m}}\sigma(\tilde{\boldsymbol{u}}_i^\top \tilde{x}), \quad \partial_{\tilde{\boldsymbol{u}}_i} f_\theta^{\mathrm{NN}}(x) = \frac{1}{\sqrt{m}} v_i\, \sigma'(\tilde{\boldsymbol{u}}_i^\top \tilde{x})\, \tilde{x}, \quad \partial_a f_\theta^{\mathrm{NN}}(x) = 1.$$

Using the definition of K_θ^m in Equation (32), we obtain:

$$\begin{aligned} K_\theta^m(x,x') =& \sum_{i=1}^m \left\langle \partial_{v_i} f_\theta^{\mathrm{NN}}(x), \partial_{v_i} f_\theta^{\mathrm{NN}}(x') \right\rangle + \sum_{i=1}^m \left\langle \partial_{\tilde{\boldsymbol{u}}_i} f_\theta^{\mathrm{NN}}(x), \partial_{\tilde{\boldsymbol{u}}_i} f_\theta^{\mathrm{NN}}(x') \right\rangle + 1 \\ =& \frac{1}{m}\sum_{i=1}^m \sigma(\tilde{\boldsymbol{u}}_i^\top \tilde{x})\sigma(\tilde{\boldsymbol{u}}_i^\top \tilde{x}') + \frac{1}{m}\sum_{i=1}^m v_i^2\, \sigma'(\tilde{\boldsymbol{u}}_i^\top \tilde{x})\sigma'(\tilde{\boldsymbol{u}}_i^\top \tilde{x}')\, \tilde{x}^\top \tilde{x}' + 1. \end{aligned} \tag{34}$$

Infinite-width limit. When the network's width m goes to $+\infty$, $K_{\theta_t}^m$ converges to a *time-invariant deterministic* kernel K^{NTK}, referred to as the *infinite-width* NTK [4, 53, 59]. Let $f_{\theta_t}^{\mathrm{NTK}}$ denote the function that describes the limiting training dynamics of $f_{\theta_t}^{\mathrm{NN}}$ when $m \to +\infty$. Then, in this regime, the dynamics of $f_{\theta_t}^{\mathrm{NTK}}$ are described by:

$$\frac{\mathrm{d}}{\mathrm{d}t} f_{\theta_t}^{\mathrm{NTK}}(x) = -\frac{1}{n}\sum_{i=1}^n K^{\mathrm{NTK}}(x, x_i)\left(f_{\theta_t}^{\mathrm{NTK}}(x_i) - y_i\right). \tag{35}$$

Similarly to the neural network initialization, we also have that $f_{\theta_0}^{\mathrm{NTK}} = 0$ due to the mirrored initialization. For the neural network defined above in Equation (31), K^{NTK} also admits an explicit expression. Indeed, as $m \to +\infty$, by the law of large numbers and the Gaussian initialization of the parameters, the kernel K_θ^m, whose closed-form expression is provided in Equation (34), converges to a deterministic kernel K^{NTK} given for any $x, x' \in \mathcal{X}$ by:

$$\begin{aligned} K^{\mathrm{NTK}}(x,x') =& \mathbb{E}_{\tilde{\boldsymbol{u}}\sim\mathcal{N}(0,\mathbb{1}_{(p+1)\times(p+1)})}\left[\sigma(\tilde{\boldsymbol{u}}^\top \tilde{x})\sigma(\tilde{\boldsymbol{u}}^\top \tilde{x}')\right] \\ &+ \mathbb{E}_{v\sim\mathcal{N}(0,1),\tilde{\boldsymbol{u}}\sim\mathcal{N}(0,\mathbb{1}_{(p+1)\times(p+1)})}\left[v^2\sigma'(\tilde{\boldsymbol{u}}^\top \tilde{x})\sigma'(\tilde{\boldsymbol{u}}^\top \tilde{x}')\right]\tilde{x}^\top \tilde{x}' + 1 \\ =& \mathbb{E}_{\tilde{\boldsymbol{u}}\sim\mathcal{N}(0,\mathbb{1}_{(p+1)\times(p+1)})}\left[\sigma(\tilde{\boldsymbol{u}}^\top \tilde{x})\sigma(\tilde{\boldsymbol{u}}^\top \tilde{x}')\right] \\ &+ \mathbb{E}_{\tilde{\boldsymbol{u}}\sim\mathcal{N}(0,\mathbb{1}_{(p+1)\times(p+1)})}\left[\sigma'(\tilde{\boldsymbol{u}}^\top \tilde{x})\sigma'(\tilde{\boldsymbol{u}}^\top \tilde{x}')\right]\tilde{x}^\top \tilde{x}' + 1, \end{aligned} \tag{36}$$

where in the second equality, we use the independence between v and $\tilde{\boldsymbol{u}}$, as well as $\mathbb{E}_{v\sim\mathcal{N}(0,1)}[v^2] = 1$. The NTK matrix $\boldsymbol{K}^{\mathrm{NTK}} = [K^{\mathrm{NTK}}(x_i, x_j)]_{i,j=1}^n$ can thus be expressed as:

$$\begin{aligned} \boldsymbol{K}^{\mathrm{NTK}} =& \mathbb{E}_{\tilde{\boldsymbol{u}}\sim\mathcal{N}(0,\mathbb{1}_{(p+1)\times(p+1)})}\left[\sigma(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{u}})\sigma(\tilde{\boldsymbol{u}}^\top \tilde{\boldsymbol{X}}^\top)\right] \\ &+ \mathbb{E}_{\tilde{\boldsymbol{u}}\sim\mathcal{N}(0,\mathbb{1}_{(p+1)\times(p+1)})}\left[\sigma'(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{u}})\sigma'(\tilde{\boldsymbol{u}}^\top \tilde{\boldsymbol{X}}^\top)\right] \odot \tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top + \mathbb{1}_n\mathbb{1}_n^\top, \end{aligned} \tag{37}$$

where $\tilde{\boldsymbol{X}} = [\tilde{x}_1, \ldots, \tilde{x}_n]^\top \in \mathbb{R}^{n\times(p+1)}$ is the design matrix of the dataset $\{\tilde{x}_i\}_{i=1}^n$ and $\odot$ is the Hadamard product. In Equation (37), σ and σ' are applied entrywise to vectors or matrices. Moreover, we have the following explicit formula for the infinite-width NTK [27]:

$$K^{\mathrm{NTK}}(x,x') = \frac{\|\tilde{x}\|\|\tilde{x}'\|}{2\pi}(\sin\varphi + (\pi-\varphi)\cos\varphi) + \frac{\pi-\varphi}{2\pi}\cdot \tilde{x}^\top \tilde{x}' + 1, \tag{38}$$

where $\varphi = \arccos\left(\frac{\tilde{x}^\top \tilde{x}'}{\|\tilde{x}\|\|\tilde{x}'\|}\right)$ is the angle between the augmented inputs $\tilde{x} = [x^\top, 1]^\top$ and $\tilde{x}' = [x'^\top, 1]^\top$. In the following, for notational simplicity, we denote $f_{\theta_t}^{\mathrm{NN}}$, $f_{\theta_t}^{\mathrm{NTK}}$, and $K_{\theta_t}^m$ by f_t^{NN}, f_t^{NTK}, and K_t^m, respectively. In the remainder of Appendix E, we denote by $\mathcal{H}^{\mathrm{NTK}}$ the RKHS associated with the infinite-width kernel K^{NTK}, by $[\mathcal{H}^{\mathrm{NTK}}]^s$ its interpolation space, and by $(\lambda_i)_{i\geq 1} \subset (0, +\infty)$ the non-increasing sequence of non-zero eigenvalues of the integral operator $\mathcal{T}$ (equivalently, of the covariance operator Σ), with associated orthonormal family $(e_i)_{i\geq 1}$ in $L^2(\mathcal{X})$. We refer the reader to Appendix H for details on interpolation spaces and the spectral decomposition. We further denote by λ_0 the smallest eigenvalue of the infinite-width NTK matrix $\boldsymbol{K}^{\mathrm{NTK}} = [K^{\mathrm{NTK}}(x_i, x_j)]_{i,j=1}^n \in \mathbb{R}^{n\times n}$, which controls the convergence of the NTK gradient flow.

E.3 Assumptions

We make the following assumptions.

Assumption a. (Compactness of $\mathcal{X}$) The input space $\mathcal{X} \subset \mathbb{R}^p$ is compact with a constant $A > 0$ such that $\|x\| \leq A$ for any $x \in \mathcal{X}$.
Assumption b. (Boundedness of $\mathcal{Y}$) The target space $\mathcal{Y} \subset \mathbb{R}$ is bounded, *i.e.*, there exists a constant $B > 0$ such that $|y| \leq B$ for any $y \in \mathcal{Y}$.
Assumption c. (Source condition for $f^\star$) There exist $s, R > 0$ such that $\|f^\star\|_{[\mathcal{H}^{\mathrm{NTK}}]^s} \leq R$.
Assumption d. (Bounded density) There exists a positive constant $M_\mu > 0$ such that $f_\mu \leq M_\mu$. Moreover, there exist an open ball $\Omega_\mu \subset \mathcal{X}$ and a constant $c_\mu > 0$ such that $f_\mu \geq c_\mu$ on Ω_μ.

We now derive some consequences of these assumptions that will be useful later.

Lemma E.1 (Boundedness of $f^\star$)**.** *Under Assumption b, $f^\star \in L^\infty(\mathcal{X})$ and $\|f^\star\|_\infty \leq B$.*

Proof. For any $x \in \mathcal{X}$, we have $|f^\star(x)| = |\mathbb{E}_{y\sim\rho}[y \mid x]| \leq B$ by Jensen's inequality and Assumption b. Taking the supremum over $x \in \mathcal{X}$ allows us to conclude. □

Lemma E.2 (Boundedness of K^{NTK})**.** *Under Assumption a, K^{NTK} is bounded by $\kappa^{\mathrm{NTK}} := (p+2)(A^2+1)+1$ on $\mathcal{X} \times \mathcal{X}$.*

Proof. We start by recalling the expression of K^{NTK} given in Equation (36) for any $x, x' \in \mathcal{X}$ by:

$$\begin{aligned} K^{\mathrm{NTK}}(x,x') =& \mathbb{E}_{\tilde{\boldsymbol{u}}\sim\mathcal{N}(0,\mathbb{1}_{(p+1)\times(p+1)})}\left[\sigma(\tilde{\boldsymbol{u}}^\top\tilde{x})\sigma(\tilde{\boldsymbol{u}}^\top\tilde{x}')\right] \\ &+ \mathbb{E}_{\tilde{\boldsymbol{u}}\sim\mathcal{N}(0,\mathbb{1}_{(p+1)\times(p+1)})}\left[\sigma'(\tilde{\boldsymbol{u}}^\top\tilde{x})\sigma'(\tilde{\boldsymbol{u}}^\top\tilde{x}')\right]\tilde{x}^\top\tilde{x}' + 1, \end{aligned}$$

By the triangle inequality, we have:

$$\begin{aligned} \left|K^{\mathrm{NTK}}(x,x')\right| \leq& \left|\mathbb{E}_{\tilde{\boldsymbol{u}}\sim\mathcal{N}(0,\mathbb{1}_{(p+1)\times(p+1)})}\left[\sigma(\tilde{\boldsymbol{u}}^\top\tilde{x})\sigma(\tilde{\boldsymbol{u}}^\top\tilde{x}')\right]\right| \\ &+ \left|\mathbb{E}_{\tilde{\boldsymbol{u}}\sim\mathcal{N}(0,\mathbb{1}_{(p+1)\times(p+1)})}\left[\sigma'(\tilde{\boldsymbol{u}}^\top\tilde{x})\sigma'(\tilde{\boldsymbol{u}}^\top\tilde{x}')\right]\tilde{x}^\top\tilde{x}'\right| + 1, \end{aligned} \tag{39}$$

Using the triangle inequality, the identities $|\sigma(z)| \leq |z|$ and $\sigma'(z) \leq 1$, Cauchy-Schwarz inequality, and Assumption a, we get:

$$\begin{aligned} \left|\mathbb{E}_{\tilde{\boldsymbol{u}}\sim\mathcal{N}(0,\mathbb{1}_{(p+1)\times(p+1)})}\left[\sigma(\tilde{\boldsymbol{u}}^\top\tilde{x})\sigma(\tilde{\boldsymbol{u}}^\top\tilde{x}')\right]\right| \leq& \mathbb{E}_{\tilde{\boldsymbol{u}}\sim\mathcal{N}(0,\mathbb{1}_{(p+1)\times(p+1)})}\left[\left|\tilde{\boldsymbol{u}}^\top\tilde{x}\right|\left|\tilde{\boldsymbol{u}}^\top\tilde{x}'\right|\right] \\ \leq& \underbrace{\mathbb{E}_{\tilde{\boldsymbol{u}}\sim\mathcal{N}(0,\mathbb{1}_{(p+1)\times(p+1)})}\left[\|\tilde{\boldsymbol{u}}\|^2\right]}_{=p+1}\|\tilde{x}\|\|\tilde{x}'\| \\ \leq& (p+1)(A^2+1), \end{aligned}$$

$$\text{and} \quad \left|\mathbb{E}_{\tilde{\boldsymbol{u}}\sim\mathcal{N}(0,\mathbb{1}_{(p+1)\times(p+1)})}\left[\sigma'(\tilde{\boldsymbol{u}}^\top\tilde{x})\sigma'(\tilde{\boldsymbol{u}}^\top\tilde{x}')\right]\tilde{x}^\top\tilde{x}'\right| \leq \|\tilde{x}\|\|\tilde{x}'\| \leq A^2+1.$$

Substituting both inequalities into Equation (39) concludes the proof. □

Remark E.3. As pointed out in Appendix H, Lemma E.2 implies that the RKHS $\mathcal{H}^{\mathrm{NTK}}$ is included in $L^2(\mathcal{X})$. This allows us to consider the closure of $\mathcal{H}^{\mathrm{NTK}}$ in $L^2(\mathcal{X})$, as we do in Lemma E.4.

Lemma E.4. *Under Assumptions a and c, $f^\star$ lies in the closure of $\mathcal{H}^{\mathrm{NTK}}$ in $L^2(\mathcal{X})$.*

Proof. It suffices to show that there exists a sequence in $\mathcal{H}^{\mathrm{NTK}}$ that converges to $f^\star$ in $L^2(\mathcal{X})$. Under Assumption c, we know that $f^\star \in [\mathcal{H}^{\mathrm{NTK}}]^s$ for some $s > 0$. Therefore, $f^\star$ can be written as (see Appendix H):

$$f^\star = \sum_{i\geq 1} a_i \lambda_i^{\frac{s}{2}} e_i, \qquad \text{with } \sum_{i\geq 1} a_i^2 < +\infty.$$

Let $(h_n)_{n\geq 1}$ be a sequence in $\mathcal{H}^{\mathrm{NTK}}$ defined for any $n \geq 1$ by:

$$h_n = \sum_{i=1}^n a_i \lambda_i^{\frac{s}{2}} e_i = \sum_{i=1}^n \left(a_i \lambda_i^{\frac{s-1}{2}}\right) \lambda_i^{\frac{1}{2}} e_i.$$

Indeed, $h_n \in \mathcal{H}^{\mathrm{NTK}}$ for any $n \geq 1$ since h_n is a finite linear combination of the basis functions $\lambda_i^{\frac{1}{2}} e_i$. Moreover, we have:

$$\|h_n - f^\star\|^2_{L^2(\mathcal{X})} = \left\| \sum_{i>n} a_i \lambda_i^{\frac{s}{2}} e_i \right\|^2_{L^2(\mathcal{X})} = \sum_{i>n} a_i^2 \lambda_i^s \leq \lambda_1^s \sum_{i>n} a_i^2,$$

where the second equality follows from Parseval's identity, and the inequality uses the fact that the eigenvalues $(\lambda_i)_{i\geq 1}$ are sorted in non-increasing order. Given that the series $(a_i^2)_{i\geq 1}$ is convergent, we know that its tail $\sum_{i>n} a_i^2 \longrightarrow 0$ as $n \to \infty$. Hence, we obtain:

$$\|h_n - f^\star\|^2_{L^2(\mathcal{X})} \xrightarrow[n\to\infty]{} 0,$$

which shows that the sequence $(h_n)_{n\geq 1}$ converges to $f^\star$ in $L^2(\mathcal{X})$. This concludes the proof. □

E.4 Lower-bound on the smallest eigenvalue of $\boldsymbol{K}^{\mathrm{NTK}}$

Recall that we are given a dataset $x_1, \ldots, x_n \in \mathcal{X} \subset \mathbb{R}^p$ sampled from a distribution μ and an augmented dataset $\tilde{x}_1, \ldots, \tilde{x}_n \in \mathbb{R}^{p+1}$ where $\tilde{x}_i = [x_i^\top, 1]^\top$ for $1 \leq i \leq n$, with design matrices $\boldsymbol{X} = [x_1, \ldots, x_n]^\top \in \mathbb{R}^{n\times p}$ and $\tilde{\boldsymbol{X}} = [\tilde{x}_1, \ldots, \tilde{x}_n]^\top \in \mathbb{R}^{n\times(p+1)}$, respectively. Recall also that σ is the ReLU activation function, and σ' is its derivative. For $\psi \in \{\sqrt{p+1}\sigma, \sigma'\}$, let $\boldsymbol{K}^\infty_\psi \in \mathbb{R}^{n\times n}$ be the matrix defined as:

$$\boldsymbol{K}^\infty_\psi = \mathbb{E}_{w\sim\mathcal{U}(\mathbb{S}^p)}\left[\psi(\hat{\boldsymbol{X}}w)\psi(w^\top \hat{\boldsymbol{X}}^\top)\right],$$

where $\mathcal{U}(\mathbb{S}^p)$ denotes the uniform distribution on the unit sphere $\mathbb{S}^p$, ψ is applied entrywise to the vector $\hat{\boldsymbol{X}}w \in \mathbb{R}^n$ and its transpose, with $\hat{\boldsymbol{X}} = [\hat{x}_1, \ldots, \hat{x}_n]^\top \in \mathbb{R}^{n\times(p+1)}$ being the design matrix of the normalized counterparts of $\tilde{x}_1, \ldots, \tilde{x}_n$, defined by:

$$\hat{x}_i = \frac{\tilde{x}_i}{\|\tilde{x}_i\|} = \frac{1}{\sqrt{\|x_i\|^2+1}} \begin{bmatrix} x_i \\ 1 \end{bmatrix} \in \mathbb{S}^p \quad \text{for } 1 \leq i \leq n. \tag{40}$$

Definition E.5 (ϵ-separated dataset [55])**.** A dataset $\{x_i\}_{i=1}^n$ is ϵ-separated for $\epsilon > 0$ if:

$$\min_{i\neq k} \min\left(\|x_i - x_k\|, \|x_i + x_k\|\right) \geq \epsilon.$$

Lemma E.6 (Lemma 7 in [55])**.** *Let $\hat{\boldsymbol{X}} \in \mathbb{R}^{n\times(p+1)}$ be the design matrix of the dataset $\hat{x}_1, \ldots, \hat{x}_n \in \mathbb{S}^p$ with $p \geq 2$, and $\psi \in \{\sqrt{p+1}\sigma, \sigma'\}$. Assume $0 < \epsilon \leq 1$. Suppose that this dataset is ϵ-separated. Then, for all $w \in \mathbb{R}^n$ with $\|w\| = 1$, we have:*

$$\left\langle \boldsymbol{K}^\infty_\psi w, w \right\rangle \geq \begin{cases} C(1+g(\epsilon))^{-3}\epsilon^2 & \text{if } \psi = \sigma' \\ C(1+g(\epsilon))^{-3}\epsilon^4 & \text{if } \psi = \sqrt{p+1}\sigma, \end{cases}$$

where $C > 0$ is an absolute positive constant and $g(\epsilon) = \frac{(p+1)\ln(1/\epsilon)}{\ln(p+1)}$.

Lemma E.7 (Deterministic lower-bound on the smallest eigenvalue of $\boldsymbol{K}^{\mathrm{NTK}}$)**.** *Let $\tilde{\boldsymbol{X}} \in \mathbb{R}^{n\times(p+1)}$ be the design matrix of the augmented dataset $\tilde{x}_i = [x_i^\top, 1]^\top$ and $\hat{\boldsymbol{X}} \in \mathbb{R}^{n\times(p+1)}$ be the design matrix of the normalized dataset $\hat{x}_i = \tilde{x}_i/\|\tilde{x}_i\|$, with $p \geq 2$. Assume $0 < \epsilon \leq 1$. Suppose that the normalized dataset $\hat{x}_1, \ldots, \hat{x}_n \in \mathbb{S}^p$ is ϵ-separated. Then, we have the following lower-bound on the smallest eigenvalue λ_0 of $\boldsymbol{K}^{\mathrm{NTK}}$:*

$$\lambda_0 \geq C(1+g(\epsilon))^{-3}\epsilon^4,$$

where $C > 0$ is an absolute positive constant and $g(\epsilon) = \frac{(p+1)\ln(1/\epsilon)}{\ln(p+1)}$.

Proof. By Equation (37), the infinite-width NTK can be written as:

$$\begin{aligned} \boldsymbol{K}^{\mathrm{NTK}} =& \mathbb{E}_{\tilde{\boldsymbol{u}}\sim\mathcal{N}(0,\mathbb{1}_{(p+1)\times(p+1)})}\left[\sigma(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{u}})\sigma(\tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{X}}^\top)\right] \\ &+ \mathbb{E}_{\tilde{\boldsymbol{u}}\sim\mathcal{N}(0,\mathbb{1}_{(p+1)\times(p+1)})}\left[\sigma'(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{u}})\sigma'(\tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{X}}^\top)\right] \odot \tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top + \mathbb{1}_n\mathbb{1}_n^\top. \end{aligned}$$

Since $\tilde{\boldsymbol{u}} \sim \mathcal{N}(0, \mathbb{1}_{(p+1)\times(p+1)})$, using the Gaussian polar decomposition, it holds that $\tilde{\boldsymbol{u}} = r\boldsymbol{w}$, where $r = \|\tilde{\boldsymbol{u}}\|$, $\boldsymbol{w} \sim \mathcal{U}(\mathbb{S}^p)$, and r and $\boldsymbol{w}$ are independent. We note also that $r^2 = \|\tilde{\boldsymbol{u}}\|^2 = \tilde{u}_1^2 + \cdots + \tilde{u}_{p+1}^2$, where $\tilde{u}_1, \ldots, \tilde{u}_{p+1}$ are independent and sampled from the standard Gaussian distribution $\mathcal{N}(0,1)$. This implies that $r^2 \sim \chi^2_{p+1}$ (the chi-squared distribution with $p+1$ degrees of freedom) with $\mathbb{E}[r^2] = p+1$. Writing $\tilde{x}_i = \|\tilde{x}_i\|\hat{x}_i$ and letting $\boldsymbol{D} = \mathbf{diag}(\|\tilde{x}_1\|, \ldots, \|\tilde{x}_n\|)$, so that $\tilde{\boldsymbol{X}}\boldsymbol{w} = \boldsymbol{D}\hat{\boldsymbol{X}}\boldsymbol{w}$, and using the facts that σ is homogeneous of degree 1 and σ' is scale-invariant together with the independence of r and $\boldsymbol{w}$, we obtain:

$$
\begin{aligned}
&\boldsymbol{K}^{\mathrm{NTK}} \\
=&\mathbb{E}_{r,\boldsymbol{w}\sim\mathcal{U}(\mathbb{S}^p)}\left[\sigma(\tilde{\boldsymbol{X}}r\boldsymbol{w})\sigma(r\boldsymbol{w}^\top\tilde{\boldsymbol{X}}^\top)\right] + \mathbb{E}_{r,\boldsymbol{w}\sim\mathcal{U}(\mathbb{S}^p)}\left[\sigma'(\tilde{\boldsymbol{X}}r\boldsymbol{w})\sigma'(r\boldsymbol{w}^\top\tilde{\boldsymbol{X}}^\top)\right] \odot \tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top + \mathbb{1}_n\mathbb{1}_n^\top \\
=&\underbrace{\mathbb{E}_{r^2\sim\chi^2_{p+1}}[r^2]}_{=p+1}\mathbb{E}_{\boldsymbol{w}\sim\mathcal{U}(\mathbb{S}^p)}\left[\sigma(\tilde{\boldsymbol{X}}\boldsymbol{w})\sigma(\boldsymbol{w}^\top\tilde{\boldsymbol{X}}^\top)\right] + \mathbb{E}_{\boldsymbol{w}\sim\mathcal{U}(\mathbb{S}^p)}\left[\sigma'(\tilde{\boldsymbol{X}}\boldsymbol{w})\sigma'(\boldsymbol{w}^\top\tilde{\boldsymbol{X}}^\top)\right] \odot \tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top \\
&+ \mathbb{1}_n\mathbb{1}_n^\top \\
=&\boldsymbol{D}\boldsymbol{K}^\infty_{\sqrt{p+1}\sigma}\boldsymbol{D} + \boldsymbol{K}^\infty_{\sigma'} \odot \tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top + \mathbb{1}_n\mathbb{1}_n^\top,
\end{aligned}
$$

where the last line uses $\sigma(\tilde{\boldsymbol{X}}\boldsymbol{w}) = \boldsymbol{D}\,\sigma(\hat{\boldsymbol{X}}\boldsymbol{w})$ and $\sigma'(\tilde{\boldsymbol{X}}\boldsymbol{w}) = \sigma'(\hat{\boldsymbol{X}}\boldsymbol{w})$. The matrices $\boldsymbol{D}\boldsymbol{K}^\infty_{\sqrt{p+1}\sigma}\boldsymbol{D}$, $\boldsymbol{K}^\infty_{\sigma'} \odot \tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top$, and $\mathbb{1}_n\mathbb{1}_n^\top$ are all positive semi-definite (the second by the Schur product theorem). Hence, by Weyl's inequality, we obtain:

$$
\lambda_0 \geq \lambda_{\min}(\boldsymbol{D}\boldsymbol{K}^\infty_{\sqrt{p+1}\sigma}\boldsymbol{D}) + \lambda_{\min}(\boldsymbol{K}^\infty_{\sigma'} \odot \tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top) + \lambda_{\min}(\mathbb{1}_n\mathbb{1}_n^\top) \geq \lambda_{\min}(\boldsymbol{D}\boldsymbol{K}^\infty_{\sqrt{p+1}\sigma}\boldsymbol{D}), \quad (41)
$$

where we use the fact that $\lambda_{\min}(\boldsymbol{K}^\infty_{\sigma'} \odot \tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top) \geq 0$ and $\lambda_{\min}(\mathbb{1}_n\mathbb{1}_n^\top) \geq 0$ given that $\boldsymbol{K}^\infty_{\sigma'} \odot \tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top$ and $\mathbb{1}_n\mathbb{1}_n^\top$ are both positive semi-definite. By definition, for any unit vector $w \in \mathbb{R}^n$, we have:

$$
\begin{aligned}
\lambda_{\min}(\boldsymbol{D}\boldsymbol{K}^\infty_{\sqrt{p+1}\sigma}\boldsymbol{D}) &= \inf_{\|w\|=1}\langle \boldsymbol{D}\boldsymbol{K}^\infty_{\sqrt{p+1}\sigma}\boldsymbol{D}w, w\rangle \\
&= \inf_{\|w\|=1}\|\boldsymbol{D}w\|^2\left\langle \boldsymbol{K}^\infty_{\sqrt{p+1}\sigma}\frac{\boldsymbol{D}w}{\|\boldsymbol{D}w\|}, \frac{\boldsymbol{D}w}{\|\boldsymbol{D}w\|}\right\rangle \\
&\geq \lambda_{\min}(\boldsymbol{K}^\infty_{\sqrt{p+1}\sigma})\cdot\lambda_{\min}(\boldsymbol{D}^2),
\end{aligned}
$$

since $\boldsymbol{D}w/\|\boldsymbol{D}w\|$ is a unit vector and $\|\boldsymbol{D}w\|^2 \geq \lambda_{\min}(\boldsymbol{D}^2)\|w\|^2 = \lambda_{\min}(\boldsymbol{D}^2)$. Since each $\tilde{x}_i$ has an appended 1, we have $\lambda_{\min}(\boldsymbol{D}^2) = \min_i\|\tilde{x}_i\|^2 = \min_i(\|x_i\|^2+1) \geq 1$. Moreover, by definition, we have:

$$
\lambda_{\min}(\boldsymbol{K}^\infty_{\sqrt{p+1}\sigma}) = \inf_{\|w\|=1}\left\langle \boldsymbol{K}^\infty_{\sqrt{p+1}\sigma}w, w\right\rangle.
$$

It then follows from Lemma E.6 that:

$$
\lambda_{\min}(\boldsymbol{D}\boldsymbol{K}^\infty_{\sqrt{p+1}\sigma}\boldsymbol{D}) \geq C(1+g(\epsilon))^{-3}\epsilon^4,
$$

where $C > 0$ is an absolute positive constant and $g(\epsilon) = \frac{(p+1)\ln(1/\epsilon)}{\ln(p+1)}$. Substituting this lower-bound into Equation (41) concludes the proof. □

Lemma E.8 (High-probability separation of homogeneous normalized points on $\mathbb{S}^p$)**.** *Let Assumptions a and d hold. Define the homogeneous normalized points $\{\hat{x}_i\}_{i=1}^n$ as in Equation (40). Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$ with respect to random samples, the normalized dataset $\{\hat{x}_i\}_{i=1}^n$ is ϵ-separated in the sense of Definition E.5 with*

$$
\epsilon \leq \min\left(\frac{2}{\sqrt{A^2+1}}, \frac{2}{\pi}\left(\frac{2\delta}{M_\mu(A^2+1)^{(p+1)/2}v_p n^2}\right)^{\frac{1}{p}}\right),
$$

where $v_p = \mathrm{Vol}(B(0,1))$ is the volume of the unit ball in $\mathbb{R}^p$.

Proof. Let $\phi : \mathcal{X} \subset \mathbb{R}^p \to \mathbb{S}^p \subset \mathbb{R}^{p+1}$ be the normalization map defined for any $x \in \mathcal{X}$ by:

$$
\phi(x) = \frac{1}{\sqrt{\|x\|^2+1}}\begin{bmatrix} x \\ 1\end{bmatrix},
$$

so that $\hat{x}_i = \phi(x_i)$. Define the event that the dataset is not ϵ-separated:

$$\mathcal{E}_\epsilon = \left\{ \min_{i \neq k} \min \left(\|\hat{x}_i - \hat{x}_k\|, \|\hat{x}_i + \hat{x}_k\| \right) < \epsilon \right\},$$

i.e., there exists at least one pair $i \neq k$ for which the points are either too close together, or too close to antipodal. Hence, we obtain the following inclusion:

$$\mathcal{E}_\epsilon \subset \bigcup_{i<k} \left(\{ \|\hat{x}_i - \hat{x}_k\| < \epsilon \} \cup \{ \|\hat{x}_i + \hat{x}_k\| < \epsilon \} \right).$$

Note that the antipodal term $\|\hat{x}_i + \hat{x}_k\| < \epsilon$ need not be considered provided that ϵ is chosen small enough such that $\epsilon \leq 2/\sqrt{A^2+1}$. Indeed, for any $x \in \mathcal{X}$ with $\|x\| \leq A$ (Assumption a), the last coordinate of $\phi(x)$ is $1/\sqrt{\|x\|^2+1} \geq 1/\sqrt{A^2+1} > 0$, so all points $\hat{x}_i$ lie in the open upper hemisphere $\mathbb{S}^p_+ := \{z \in \mathbb{S}^p \mid z_{p+1} > 0\}$, bounded away from the equator by the constant $1/\sqrt{A^2+1}$. Therefore $(\hat{x}_i + \hat{x}_k)_{p+1} \geq 2/\sqrt{A^2+1}$ for all $i \neq k$. Since the last coordinate of a vector is dominated by its Euclidean norm, we have $\|\hat{x}_i + \hat{x}_k\| \geq (\hat{x}_i + \hat{x}_k)_{p+1} \geq 2/\sqrt{A^2+1}$. Hence, provided that $\epsilon \leq 2/\sqrt{A^2+1}$, the event $\{\|\hat{x}_i + \hat{x}_k\| < \epsilon\}$ is empty for every pair $i \neq k$. This implies that the antipodal event cannot occur. Thus:

$$\mathcal{E}_\epsilon \subset \bigcup_{i<k} \{ \|\hat{x}_i - \hat{x}_k\| < \epsilon \}.$$

Therefore, by the union bound, we get:

$$\Pr(\mathcal{E}_\epsilon) \leq \sum_{i<k} \Pr\left(\|\hat{x}_i - \hat{x}_k\| < \epsilon \right). \tag{42}$$

Let $\nu = \phi_{\#}\mu$ be the pushforward distribution of μ under the map ϕ, so that $\hat{x}_i \sim \nu$ on $\mathbb{S}^p$. Since $\hat{x}_i$ and $\hat{x}_k$ are independent and identically distributed samples from ν on $\mathbb{S}^p$, conditioning on $\hat{x}_k = z$ and applying the tower property yield:

$$\Pr\left(\|\hat{x}_i - \hat{x}_k\| < \epsilon \right) = \mathbb{E}\left[\Pr\left(\|\hat{x}_i - z\| < \epsilon \mid \hat{x}_k = z \right) \right] = \mathbb{E}\left[\nu\left(B_{\mathbb{S}^p}(z, \epsilon) \right) \right], \tag{43}$$

where $B_{\mathbb{S}^p}(z, \epsilon) = \{\hat{x} \in \mathbb{S}^p \mid \|\hat{x} - z\| < \epsilon\}$ is the Euclidean ball of radius ϵ centered at z intersected with $\mathbb{S}^p$. We now derive the density of ν with respect to the surface measure σ on $\mathbb{S}^p$. For any measurable set $\mathcal{M} \subset \mathbb{S}^p$, we have:

$$\nu(\mathcal{M}) = \mu(\phi^{-1}(\mathcal{M})) = \int_{\phi^{-1}(\mathcal{M})} f_\mu(x)\, \mathrm{d}x = \int_{\mathbb{R}^p} g(x) J_\phi(x)\, \mathrm{d}x, \tag{44}$$

where $g(x) = f_\mu(x)/J_\phi(x) \cdot \mathbb{1}_{\phi(x) \in \mathcal{M}}$, $J_\phi(x) = \sqrt{\det\left(D\phi(x)^\top D\phi(x) \right)}$ with $D\phi(x) \in \mathbb{R}^{(p+1)\times p}$ being the Jacobian matrix. Since ϕ is C^∞ on $\mathbb{R}^p$, it is Lipschitz on the compact domain $\mathcal{X}$. The area formula [36, Theorem 3.9] therefore gives:

$$\int_{\mathbb{R}^p} g(x) J_\phi(x)\, \mathrm{d}x = \int_{\mathbb{S}^p} \left[\sum_{x \in \phi^{-1}(z)} g(x) \right] \mathrm{d}\sigma(z).$$

Given that ϕ is injective on $\mathbb{R}^p$, the preimage $\phi^{-1}(z)$ contains exactly one point for each $z \in \mathbb{S}^p_+$ and is empty for $z \notin \mathbb{S}^p_+$. Using injectivity so that the sum has exactly one term for $z \in \mathbb{S}^p_+$, we obtain:

$$\int_{\mathbb{R}^p} g(x) J_\phi(x)\, \mathrm{d}x = \int_{\mathbb{S}^p} g(\phi^{-1}(z)) \mathrm{d}\sigma(z) = \int_{\mathcal{M}} \frac{f_\mu(\phi^{-1}(z))}{J_\phi(\phi^{-1}(z))} \mathrm{d}\sigma(z).$$

Next, we find the expression of $J_\phi(x)$. Write $s = \|x\|^2 + 1$ for brevity. $D\phi(x)$ has entries, for $1 \leq i \leq p$ and $1 \leq j \leq p$, obtained by the product rule applied to $\phi_i(x) = x_i s^{-1/2}$:

$$\frac{\partial \phi_i}{\partial x_j}(x) = \delta_{ij} s^{-1/2} + x_i \cdot \left(-x_j s^{-3/2} \right) = \frac{\delta_{ij} s - x_i x_j}{s^{3/2}},$$

where δ_{ij} is the Kronecker delta, and for the last row, using $\phi_{p+1}(x) = s^{-1/2}$ and the chain rule:

$$\frac{\partial \phi_{p+1}}{\partial x_j}(x) = -\frac{x_j}{s^{3/2}}.$$

Therefore, the full Jacobian matrix is:

$$D\phi(x) = \frac{1}{s^{3/2}} \begin{bmatrix} s\mathbb{1}_{p\times p} - xx^\top \\ -x^\top \end{bmatrix} \in \mathbb{R}^{(p+1)\times p}.$$

We compute the Gram matrix $D\phi(x)^\top D\phi(x) \in \mathbb{R}^{p\times p}$. Expanding the block matrix product and using $(xx^\top)(xx^\top) = \|x\|^2 xx^\top = (s-1)xx^\top$, we obtain:

$$\begin{aligned} D\phi(x)^\top D\phi(x) =& \frac{1}{s^3}\left[(s\mathbb{1}_{p\times p} - xx^\top)^2 + xx^\top\right] = \frac{1}{s^3}\left[s^2\mathbb{1}_{p\times p} - (s+1)xx^\top + xx^\top\right] \\ =& \frac{1}{s}\left(\mathbb{1}_{p\times p} - \frac{xx^\top}{s}\right). \end{aligned}$$

Applying the matrix determinant lemma $\det(\mathbb{1}_{p\times p} + uv^\top) = 1 + v^\top u$ with $u = -x/\sqrt{s}$ and $v = x/\sqrt{s}$:

$$\det\left(\mathbb{1}_{p\times p} - \frac{xx^\top}{s}\right) = 1 - \frac{\|x\|^2}{s} = 1 - \frac{s-1}{s} = \frac{1}{s}.$$

Therefore:

$$\det\left(D\phi(x)^\top D\phi(x)\right) = \frac{1}{s^p}\cdot\frac{1}{s} = \frac{1}{(\|x\|^2+1)^{p+1}},$$

and the Jacobian factor is:

$$J_\phi(x) = \sqrt{\det\left(D\phi(x)^\top D\phi(x)\right)} = \frac{1}{(\|x\|^2+1)^{(p+1)/2}}.$$

Consequently, we obtain:

$$\int_{\mathbb{R}^p} g(x)J_\phi(x)\,\mathrm{d}x = \int_{\mathcal{M}} f_\mu\big(\phi^{-1}(z)\big)\cdot\big(\|\phi^{-1}(z)\|^2+1\big)^{(p+1)/2}\,\mathrm{d}\sigma(z).$$

Substituting this into Equation (44), we get:

$$\nu(\mathcal{M}) = \int_{\mathcal{M}} f_\mu\big(\phi^{-1}(z)\big)\cdot\big(\|\phi^{-1}(z)\|^2+1\big)^{(p+1)/2}\,\mathrm{d}\sigma(z).$$

Since this holds for any measurable set $\mathcal{M}\subset\mathbb{S}^p$, by uniqueness of the Radon-Nikodym derivative, ν has the following density with respect to σ:

$$f_\nu(z) = f_\mu\big(\phi^{-1}(z)\big)\cdot\big(\|\phi^{-1}(z)\|^2+1\big)^{(p+1)/2}.$$

The inverse map is $\phi^{-1}(z) = z_{1:p}/z_{p+1}$, where $z_{1:p} = [z_1,\dots,z_p]^\top$ denotes the first p coordinates of z. Since $z\in\mathbb{S}^p$ we have $\|z_{1:p}\|^2 + z_{p+1}^2 = 1$, therefore:

$$\|\phi^{-1}(z)\|^2 + 1 = \frac{\|z_{1:p}\|^2}{z_{p+1}^2} + 1 = \frac{\|z_{1:p}\|^2 + z_{p+1}^2}{z_{p+1}^2} = \frac{1}{z_{p+1}^2} = \|x\|^2 + 1.$$

Given that $\|x\|\le A$ for all $x\in\mathcal{X}$ by Assumption a, we have $\|\phi^{-1}(z)\|^2+1 \le A^2+1$. Moreover, we have by Assumption d that $\|f_\mu\|_\infty \le M_\mu$. This yields:

$$\|f_\nu\|_\infty \le \|f_\mu\|_\infty \cdot \sup_{x\in\mathcal{X}}\big(\|x\|^2+1\big)^{(p+1)/2} \le M_\mu(A^2+1)^{(p+1)/2}.$$

From this bound and Lemma I.4, we bound the ν-measure of $B_{\mathbb{S}^p}(z,\epsilon)$ as follows:

$$\nu\left(B_{\mathbb{S}^p}(z,\epsilon)\right) = \int_{B_{\mathbb{S}^p}(z,\epsilon)} f_\nu(\hat{x})\,\mathrm{d}\sigma(\hat{x}) \le \|f_\nu\|_\infty\cdot\sigma\left(B_{\mathbb{S}^p}(z,\epsilon)\right) \le M_\mu(A^2+1)^{(p+1)/2} v_p\left(\frac{\pi}{2}\epsilon\right)^p,$$

where v_p is the volume of the unit ball in $\mathbb{R}^p$. Since this bound is uniform in z, after substituting it into Equation (43), we obtain:

$$\Pr\left(\|\hat{x}_i - \hat{x}_k\| < \epsilon\right) \le M_\mu(A^2+1)^{(p+1)/2} v_p\left(\frac{\pi}{2}\epsilon\right)^p.$$

Putting this bound into the union bound in Equation (42) and using $\binom{n}{2} \leq n^2/2$, we get:

$$\begin{aligned}
\Pr(\mathcal{E}_\epsilon) &\leq \sum_{i<k} M_\mu (A^2+1)^{(p+1)/2} v_p \left(\frac{\pi}{2}\epsilon\right)^p \\
&= M_\mu (A^2+1)^{(p+1)/2} v_p \left(\frac{\pi}{2}\epsilon\right)^p \underbrace{\sum_{i<k} 1}_{\text{number of pairs with } i<k} \\
&= M_\mu (A^2+1)^{(p+1)/2} v_p \left(\frac{\pi}{2}\epsilon\right)^p \binom{n}{2} \\
&\leq M_\mu (A^2+1)^{(p+1)/2} v_p \left(\frac{\pi}{2}\epsilon\right)^p \frac{n^2}{2}.
\end{aligned}$$

Setting this quantity less than or equal to δ and solving for ϵ:

$$M_\mu (A^2+1)^{(p+1)/2} v_p \left(\frac{\pi}{2}\epsilon\right)^p \frac{n^2}{2} \leq \delta \implies \epsilon \leq \frac{2}{\pi}\left(\frac{2\delta}{M_\mu (A^2+1)^{(p+1)/2} v_p n^2}\right)^{\frac{1}{p}}.$$

Therefore, for any $\epsilon \leq \frac{2}{\pi}\left(\frac{2\delta}{M_\mu (A^2+1)^{(p+1)/2} v_p n^2}\right)^{1/p}$, we have $\Pr(\mathcal{E}_\epsilon) \leq \delta$, *i.e.*, the normalized dataset $\{\hat{x}_i\}_{i=1}^n$ is ϵ-separated with probability at least $1-\delta$. Combining both upper-bounds on ϵ, we obtain:

$$\epsilon \leq \min\left(\frac{2}{\sqrt{A^2+1}}, \frac{2}{\pi}\left(\frac{2\delta}{M_\mu (A^2+1)^{(p+1)/2} v_p n^2}\right)^{\frac{1}{p}}\right),$$

which concludes the proof. □

Lemma E.9 (High-probability lower-bound on the smallest eigenvalue of $\boldsymbol{K}^{\mathrm{NTK}}$)**.** *Fix $p \geq 2$ and let Assumptions a and d hold. Define the homogeneous normalized points $\{\hat{x}_i\}_{i=1}^n$ as in Equation (40). Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$ with respect to random samples, the smallest eigenvalue λ_0 of $\boldsymbol{K}^{\mathrm{NTK}}$ satisfies:*

$$\lambda_0 \geq C(1+g(\epsilon))^{-3}\epsilon^4 \quad \textit{for every} \quad 0 < \epsilon \leq \min\left(1, \frac{2}{\sqrt{A^2+1}}, \frac{2}{\pi}\left(\frac{2\delta}{M_\mu (A^2+1)^{(p+1)/2} v_p n^2}\right)^{\frac{1}{p}}\right),$$

where $C > 0$ is an absolute positive constant, $g(\epsilon) = \frac{(p+1)\ln(1/\epsilon)}{\ln(p+1)}$, and $v_p = \mathrm{Vol}(B(0,1))$ is the volume of the unit ball in $\mathbb{R}^p$.

Proof. By Lemma E.8, with probability at least $1-\delta$ with respect to random samples, the dataset $\{\hat{x}_i\}_{i=1}^n$ is $\bar{\epsilon}$-separated with:

$$\bar{\epsilon} \leq \min\left(\frac{2}{\sqrt{A^2+1}}, \frac{2}{\pi}\left(\frac{2\delta}{M_\mu (A^2+1)^{(p+1)/2} v_p n^2}\right)^{\frac{1}{p}}\right).$$

On this event, it is also ϵ-separated for every $0 < \epsilon \leq \bar{\epsilon}$, since the condition in Definition E.5 becomes weaker as ϵ decreases. Consequently, by Lemma E.7, we obtain on this event, for every $0 < \epsilon \leq \min(1, \bar{\epsilon})$, the following lower-bound on λ_0:

$$\lambda_0 \geq C(1+g(\epsilon))^{-3}\epsilon^4,$$

where $C > 0$ is an absolute positive constant and $g(\epsilon) = \frac{(p+1)\ln(1/\epsilon)}{\ln(p+1)}$. Note that we require $\epsilon \leq 1$ to ensure that $g(\epsilon) \geq 0$, and hence $1+g(\epsilon) \geq 1$, so that the lower-bound is well defined. This concludes the proof. □

Corollary E.10 (High-probability upper-bound on the inverse of the smallest eigenvalue of $\boldsymbol{K}^{\mathrm{NTK}}$)**.** *Fix $p \geq 2$ and let Assumptions a and d hold. Define the homogeneous normalized points $\{\hat{x}_i\}_{i=1}^n$ as in Equation (40), and, for $n \geq 1$ and $\delta \in (0,1)$, the deterministic threshold $\lambda_\star(n,\delta) := h(\min(\epsilon_A, \epsilon_0(n,\delta))) > 0$, where $h(\epsilon) := C(1+g(\epsilon))^{-3}\epsilon^4$, $\epsilon_A := \min(1, 2/\sqrt{A^2+1})$*

and $\epsilon_0(n,\delta) \coloneqq \frac{2}{\pi}\left(\frac{2\delta}{M_\mu(A^2+1)^{(p+1)/2}v_p n^2}\right)^{\frac{1}{p}}$*, with* C*,* g *and* v_p *as in Lemma E.9. Then, there exists a constant* $C' > 0$*, depending only on* A*,* M_μ*, and* p*, such that, for any* $n \geq 3$ *and* $\delta \in (0,1)$*, with probability* $1-\delta$ *with respect to random samples, the smallest eigenvalue* λ_0 *of* $\boldsymbol{K}^{\mathrm{NTK}}$ *satisfies* $\lambda_0 \geq \lambda_\star(n,\delta)$*, and hence:*

$$\lambda_0^{-1} \leq \lambda_\star(n,\delta)^{-1} \leq C' n^{8/p}\delta^{-4/p}\left(\ln n + \ln(1/\delta)\right)^3.$$

Proof. Since $\epsilon_A \leq 1$, we have $g(\epsilon_A) \geq 0$, hence $1+g(\epsilon_A) \geq 1$ and $\kappa \coloneqq h(\epsilon_A) > 0$, a constant depending on A and p. Similarly, $\lambda_\star(n,\delta) > 0$, since $\min(\epsilon_A, \epsilon_0(n,\delta)) \in (0,1]$, which is admissible in Lemma E.9. Therefore, with probability at least $1-\delta$ with respect to random samples, $\lambda_0 \geq \lambda_\star(n,\delta)$. It remains to bound $\lambda_\star(n,\delta)^{-1}$. To this end, we define:

$$L \coloneqq \ln n + \ln(1/\delta), \qquad F \coloneqq n^{8/p}\,\delta^{-4/p}L^3, \qquad M \coloneqq M_\mu(A^2+1)^{(p+1)/2}v_p.$$

Since $n \geq 3$, we have $\ln n \geq \ln 3 > 1$, and since $\delta \in (0,1)$, we have $\ln(1/\delta) > 0$; hence $L \geq 1$. Moreover, $n^{8/p} \geq 1$ and $\delta^{-4/p} \geq 1$, so that $F \geq 1$. We distinguish two cases.

Case 1: $\epsilon_0(n,\delta) > \epsilon_A$**.** Then $\lambda_\star(n,\delta) = h(\epsilon_A) = \kappa$, so that, using $F \geq 1$, we get $\lambda_\star(n,\delta)^{-1} = \kappa^{-1} \leq \kappa^{-1}F$.

Case 2: $\epsilon_0(n,\delta) \leq \epsilon_A$**.** Then $\lambda_\star(n,\delta) = h(\epsilon_0(n,\delta))$. Since $\epsilon_0(n,\delta) \leq 1$, the quantity

$$\Lambda \coloneqq \ln\frac{1}{\epsilon_0(n,\delta)} = \ln\left(\frac{\pi}{2}\right) + \frac{1}{p}\ln\left(\frac{Mn^2}{2\delta}\right)$$

is non-negative, and $1+g(\epsilon_0(n,\delta)) = 1 + \frac{(p+1)\Lambda}{\ln(p+1)} \geq 1$. Substituting $\epsilon_0(n,\delta)$ gives:

$$\lambda_\star(n,\delta)^{-1} = h(\epsilon_0(n,\delta))^{-1} = C^{-1}\left(\frac{\pi}{2}\right)^4\left(\frac{M}{2}\right)^{\frac{4}{p}}\left(1+\frac{(p+1)\Lambda}{\ln(p+1)}\right)^3 n^{8/p}\,\delta^{-4/p}.$$

Expanding the logarithm, we have $\Lambda = \ln(\frac{\pi}{2}) + \frac{1}{p}\ln\frac{M}{2} + \frac{2}{p}\ln n + \frac{1}{p}\ln\frac{1}{\delta}$. Since $p \geq 2$ and $\ln n, \ln(1/\delta) \geq 0$, this gives $\Lambda \leq a_0 + L$, where $a_0 \coloneqq \left|\ln(\frac{\pi}{2}) + \frac{1}{p}\ln\frac{M}{2}\right|$. Using $L \geq 1$, we have $a_0 + L \leq (a_0+1)L$, hence:

$$1 + \frac{(p+1)\Lambda}{\ln(p+1)} \leq L + \frac{(p+1)(a_0+1)}{\ln(p+1)}L = bL,$$

where $b \coloneqq 1 + \frac{(p+1)(a_0+1)}{\ln(p+1)}$. Both sides being positive, cubing preserves the inequality, and we obtain:

$$\lambda_\star(n,\delta)^{-1} \leq cF, \qquad \text{where} \qquad c \coloneqq C^{-1}\left(\frac{\pi}{2}\right)^4\left(\frac{M}{2}\right)^{\frac{4}{p}}b^3.$$

Combining both cases yields $\lambda_\star(n,\delta)^{-1} \leq \max(\kappa^{-1}, c)F \leq (\kappa^{-1}+c)F$. Setting $C' \coloneqq \kappa^{-1}+c$ concludes the proof, since C is an absolute constant, κ depends on A and p, and M, a_0 and b depend on A, M_μ and p. □

E.5 Eigenvalue decay rate of $\mathcal{H}^{\mathrm{NTK}}$

We now derive the eigenvalue decay rate of $\mathcal{H}^{\mathrm{NTK}}$ in our setting in the sense of Definition E.11. The eigenvalues of $\mathcal{H}^{\mathrm{NTK}}$ are understood as those of its associated integral/covariance operator (see Appendix H for more details). Throughout, $\lambda_i(k; E, \nu)$ denotes the i-th largest eigenvalue of the integral operator of a kernel k on the space $L^2(E,\nu)$.

Definition E.11 (Eigenvalue decay rate)**.** Let $K : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ be a continuous kernel associated to an RKHS $\mathcal{H}$, and $(\lambda_i)_{i\geq 1}$ be the non-zero eigenvalues of its associated integral operator in non-increasing order. We say that $\mathcal{H}$ has eigenvalue decay rate $\beta > 1$ if there exist two constants $k_1, k_2 > 0$ such that $k_1 i^{-\beta} \leq \lambda_i \leq k_2 i^{-\beta}$, for any $i \geq 1$.

Lemma E.12 (Eigenvalue decay rate of $\mathcal{H}^{\mathrm{NTK}}$)**.** *Under Assumptions a and d, the eigenvalue decay rate of the integral/covariance operator associated with* K^{NTK} *in Equation (36) is* $\frac{p+1}{p}$.

Proof. Let $\phi(x) \coloneqq \tilde{x}/\|\tilde{x}\| \in \mathbb{S}^p$, $\rho(x) \coloneqq \|\tilde{x}\|/\sqrt{2}$ and $K_0(y, y') \coloneqq \kappa(\langle y, y' \rangle)$ with $y, y' \in \mathbb{S}^p$, where, for any $u \in [-1, 1]$, $\kappa(u) = u\kappa_0(u) + \kappa_1(u)$ with $\kappa_0(u) \coloneqq \frac{1}{\pi}(\pi - \arccos(u))$ and $\kappa_1(u) \coloneqq \frac{1}{\pi}(u(\pi - \arccos(u)) + \sqrt{1-u^2})$. Thus, we can write $K^{\mathrm{NTK}} = K_1 + 1$, where $K_1(x, x') \coloneqq \rho(x)K_0(\phi(x), \phi(x'))\rho(x')$. Let $\tilde{\nu} \coloneqq \phi_{\#}(\rho^2\mu)$ be the push-forward measure.

Upper bound. Since ϕ is injective, [62, Propositions 1 and 3] give:

$$\lambda_i(K_1; \mathcal{X}, \mu) = \lambda_i(K_0 \circ (\phi \times \phi); \mathcal{X}, \rho^2\mu) = \lambda_i(K_0; \mathbb{S}^p, \tilde{\nu}).$$

By the change of variables computed in the proof of Lemma E.8, $\tilde{\nu}$ has the density $q(y) = \frac{1}{2}(\|x\|^2 + 1)^{(p+3)/2} f_\mu(x)$, with $x = \phi^{-1}(y)$, with respect to the surface measure σ on $\mathbb{S}^p$, and $q = 0$ outside $\phi(\mathcal{X})$. By Assumptions a and d, $q \le C_q \coloneqq \frac{1}{2}(A^2+1)^{(p+3)/2} M_\mu$. Hence [62, Lemma 2] gives $\lambda_i(K_0; \mathbb{S}^p, \tilde{\nu}) \le C_q \lambda_i(K_0; \mathbb{S}^p, \sigma)$. By [45, Lemma B.1], $\lambda_i(K_0; \mathbb{S}^p, \sigma) \le C i^{-(p+1)/p}$ for all $i \ge 1$, where $C > 0$. Finally, the constant kernel 1 has rank one, so Weyl's inequality gives $\lambda_{i+1}(K^{\mathrm{NTK}}; \mathcal{X}, \mu) \le \lambda_i(K_1; \mathcal{X}, \mu)$ for all $i \ge 1$. Together with $\lambda_1(K^{\mathrm{NTK}}; \mathcal{X}, \mu) \le \int_{\mathcal{X}} K^{\mathrm{NTK}}(x, x) \mathrm{d}\mu(x) \le \kappa^{\mathrm{NTK}}$ (Lemma E.2), this yields $\lambda_i(K^{\mathrm{NTK}}; \mathcal{X}, \mu) \le k_2 i^{-(p+1)/p}$ for all $i \ge 1$, with $k_2 \coloneqq \max\left(\kappa^{\mathrm{NTK}}, 2^{(p+1)/p} C_q C\right)$.

Lower bound. By Assumption d, $f_\mu \ge c_\mu$ on the open ball $\Omega_\mu \subset \mathcal{X}$. The integral operator of K^{NTK} on $L^2(\Omega_\mu, \mu)$ is the compression of the one on $L^2(\mathcal{X}, \mu)$ to functions vanishing outside Ω_μ, so the minimax principle [62, Lemma 49] gives $\lambda_i(K^{\mathrm{NTK}}; \mathcal{X}, \mu) \ge \lambda_i(K^{\mathrm{NTK}}; \Omega_\mu, \mu)$. Since $\mathrm{d}\mu/\mathrm{d}x \ge c_\mu$ on Ω_μ, [62, Lemma 2] gives $\lambda_i(K^{\mathrm{NTK}}; \Omega_\mu, \mu) \ge c_\mu \lambda_i(K^{\mathrm{NTK}}; \Omega_\mu, \mathrm{d}x)$. The ball Ω_μ is a bounded Lipschitz domain, so Theorem G.12 shows that the RKHS of K^{NTK} restricted to Ω_μ equals the Sobolev space $H^{\frac{p+1}{2}}(\Omega_\mu)$, with equivalent norms. The eigenvalues $\lambda_i(K^{\mathrm{NTK}}; \Omega_\mu, \mathrm{d}x)$ are the squares of the singular values of the embedding of this RKHS into $L^2(\Omega_\mu)$ [87], and equivalent norms change these singular values by at most constant factors. The singular values of the embedding $H^{\frac{p+1}{2}}(\Omega_\mu) \hookrightarrow L^2(\Omega_\mu)$ are of exact order $i^{-\frac{p+1}{2p}}$ [32, § 3.3.4]. Hence $\lambda_i(K^{\mathrm{NTK}}; \Omega_\mu, \mathrm{d}x) \ge c' i^{-(p+1)/p}$, and therefore $\lambda_i(K^{\mathrm{NTK}}; \mathcal{X}, \mu) \ge k_1 i^{-(p+1)/p}$ with $k_1 \coloneqq c_\mu c'$, where c' depends on p, A and the radius of Ω_μ. □

E.6 Bounds on the training dynamics

For $t \ge 0$, we recall that f_t^{NN} and f_t^{NTK} are the solutions of the gradient flows provided in Equations (33) and (35), respectively.

Lemma E.13 (Training residual bounds)**.** *Consider n i.i.d. samples $\{(x_i, y_i)\}_{i=1}^n \sim \rho$. For any $t \ge 0$, let:*

$$r^{\mathrm{NTK}}(t) = f_t^{\mathrm{NTK}}(\boldsymbol{X}) - \boldsymbol{Y} = [f_t^{\mathrm{NTK}}(x_1) - y_1, \dots, f_t^{\mathrm{NTK}}(x_n) - y_n]^\top \in \mathbb{R}^n,$$
$$r^{\mathrm{NN}}(t) = f_t^{\mathrm{NN}}(\boldsymbol{X}) - \boldsymbol{Y} = [f_t^{\mathrm{NN}}(x_1) - y_1, \dots, f_t^{\mathrm{NN}}(x_n) - y_n]^\top \in \mathbb{R}^n.$$

Under Assumption b, it holds that:

$$\left\|r^{\mathrm{NTK}}(t)\right\| \le \exp\left(-\frac{\lambda_0}{n} t\right) \sqrt{n} B,$$
$$\left\|r^{\mathrm{NN}}(t) - r^{\mathrm{NTK}}(t)\right\| \le \Delta t \exp\left(\left(\Delta - \frac{\lambda_0}{n}\right) t\right) \sqrt{n} B,$$

where λ_0 is the smallest eigenvalue of the infinite-width NTK matrix $\boldsymbol{K}^{\mathrm{NTK}}$ and

$$\Delta = \sup_{t \ge 0} \sup_{\|x\| \le A, \|x'\| \le A} \left|K_t^m(x, x') - K^{\mathrm{NTK}}(x, x')\right|.$$

Remark E.14. The quantity Δ is random, as it depends on both the random initialization of the network and the random training samples. Lemma E.18 shows that, for sufficiently large width m, it is finite with high probability, and all subsequent results involving Δ are understood to hold on this high-probability event.

Proof of Lemma E.13. Let $t \geq 0$. We recall that the gradient flows of f_t^{NN} and f_t^{NTK} evaluated at the samples $\{(x_i, y_i)\}_{i=1}^n$ are:

$$\frac{\mathrm{d}}{\mathrm{d}t} r^{\mathrm{NN}}(t) = -\frac{1}{n} \boldsymbol{K}_t^m r^{\mathrm{NN}}(t), \tag{45}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} r^{\mathrm{NTK}}(t) = -\frac{1}{n} \boldsymbol{K}^{\mathrm{NTK}} r^{\mathrm{NTK}}(t), \tag{46}$$

with $r^{\mathrm{NN}}(0) = r^{\mathrm{NTK}}(0) = -\boldsymbol{Y}$.

Upper-bound on $\left\|r^{\mathrm{NTK}}(t)\right\|$. A closed-form expression of the solution to the ordinary differential equation in Equation (46) is:

$$r^{\mathrm{NTK}}(t) = \exp\left(-\frac{1}{n} t \boldsymbol{K}^{\mathrm{NTK}}\right) r^{\mathrm{NTK}}(0) = -\exp\left(-\frac{1}{n} t \boldsymbol{K}^{\mathrm{NTK}}\right) \boldsymbol{Y},$$

where the second equality holds because $f_0^{\mathrm{NTK}} = 0$. Thus, under Assumption b, we get:

$$\left\|r^{\mathrm{NTK}}(t)\right\| \leq \left\|\exp\left(-\frac{1}{n} t \boldsymbol{K}^{\mathrm{NTK}}\right)\right\|_{\mathrm{op}} \|\boldsymbol{Y}\| = \exp\left(-\frac{\lambda_0}{n} t\right) \|\boldsymbol{Y}\| \leq \exp\left(-\frac{\lambda_0}{n} t\right) \sqrt{n} B, \tag{47}$$

where we use Lemma I.5 to find the operator norm of the matrix.

Upper-bound on $\left\|r^{\mathrm{NN}}(t) - r^{\mathrm{NTK}}(t)\right\|$. Define the error $e(t) = r^{\mathrm{NN}}(t) - r^{\mathrm{NTK}}(t)$. Using the gradient flow expressions, we obtain:

$$\frac{\mathrm{d}}{\mathrm{d}t} e(t) = \frac{\mathrm{d}}{\mathrm{d}t} r^{\mathrm{NN}}(t) - \frac{\mathrm{d}}{\mathrm{d}t} r^{\mathrm{NTK}}(t) = -\frac{1}{n} \boldsymbol{K}_t^m r^{\mathrm{NN}}(t) + \frac{1}{n} \boldsymbol{K}^{\mathrm{NTK}} r^{\mathrm{NTK}}(t).$$

Adding and subtracting $\frac{1}{n} \boldsymbol{K}^{\mathrm{NTK}} r^{\mathrm{NN}}(t)$ leads to:

$$\frac{\mathrm{d}}{\mathrm{d}t} e(t) = -\frac{1}{n} \left(\boldsymbol{K}_t^m - \boldsymbol{K}^{\mathrm{NTK}}\right) r^{\mathrm{NN}}(t) - \frac{1}{n} \boldsymbol{K}^{\mathrm{NTK}} \underbrace{\left(r^{\mathrm{NN}}(t) - r^{\mathrm{NTK}}(t)\right)}_{e(t)}.$$

This is a linear ordinary differential equation with initialization $e(0) = 0$. Applying Duhamel's principle, *i.e.*, the variation of constants method, we get:

$$e(t) = \int_0^t \exp\left(-\frac{1}{n}(t-s) \boldsymbol{K}^{\mathrm{NTK}}\right) \left(-\frac{1}{n} \left(\boldsymbol{K}_s^m - \boldsymbol{K}^{\mathrm{NTK}}\right) r^{\mathrm{NN}}(s)\right) \mathrm{d}s.$$

Taking norms and applying the triangle inequality together with Lemma I.5 gives:

$$\begin{aligned}
\|e(t)\| \leq & \frac{1}{n} \int_0^t \left\|\exp\left(-\frac{1}{n}(t-s) \boldsymbol{K}^{\mathrm{NTK}}\right)\right\|_{\mathrm{op}} \left\|\boldsymbol{K}_s^m - \boldsymbol{K}^{\mathrm{NTK}}\right\|_{\mathrm{op}} \left\|r^{\mathrm{NN}}(s)\right\| \mathrm{d}s \\
= & \frac{1}{n} \int_0^t \exp\left(-\frac{1}{n} \lambda_0 (t-s)\right) \left\|\boldsymbol{K}_s^m - \boldsymbol{K}^{\mathrm{NTK}}\right\|_{\mathrm{op}} \left\|r^{\mathrm{NN}}(s)\right\| \mathrm{d}s.
\end{aligned}$$

Noticing by the triangle inequality that $\left\|r^{\mathrm{NN}}(s)\right\| \leq \|e(s)\| + \left\|r^{\mathrm{NTK}}(s)\right\|$ yields:

$$\begin{aligned}
\|e(t)\| \leq & \frac{1}{n} \int_0^t \exp\left(-\frac{1}{n} \lambda_0 (t-s)\right) \left\|\boldsymbol{K}_s^m - \boldsymbol{K}^{\mathrm{NTK}}\right\|_{\mathrm{op}} \left\|r^{\mathrm{NTK}}(s)\right\| \mathrm{d}s \\
& + \frac{1}{n} \int_0^t \exp\left(-\frac{1}{n} \lambda_0 (t-s)\right) \left\|\boldsymbol{K}_s^m - \boldsymbol{K}^{\mathrm{NTK}}\right\|_{\mathrm{op}} \|e(s)\| \mathrm{d}s
\end{aligned}$$

Multiplying both sides by $\exp\left(\frac{1}{n} \lambda_0 t\right)$ gives $f(t) \leq g(t) + \int_0^t h(s) f(s) \mathrm{d}s$, where:

$$\begin{aligned}
f(t) = & \exp\left(\frac{1}{n} \lambda_0 t\right) \|e(t)\|, \\
g(t) = & \frac{1}{n} \int_0^t \exp\left(\frac{1}{n} \lambda_0 s\right) \left\|\boldsymbol{K}_s^m - \boldsymbol{K}^{\mathrm{NTK}}\right\|_{\mathrm{op}} \left\|r^{\mathrm{NTK}}(s)\right\| \mathrm{d}s, \\
h(t) = & \frac{1}{n} \left\|\boldsymbol{K}_t^m - \boldsymbol{K}^{\mathrm{NTK}}\right\|_{\mathrm{op}}.
\end{aligned}$$

We notice that $t \mapsto h(t)$ is non-negative and $t \mapsto g(t)$ is non-decreasing. Hence, it follows from Grönwall's inequality that:

$$f(t) \leq g(t) \exp\left(\int_0^t h(s)\mathrm{d}s\right).$$

Plugging in the definitions of $f(t)$, $g(t)$, and $h(t)$, then dividing by $\exp\left(\frac{1}{n}\lambda_0 t\right)$, we get:

$$\begin{aligned}\|e(t)\| \leq & \frac{1}{n}\int_0^t \exp\left(-\frac{1}{n}\lambda_0(t-s)\right)\left\|\boldsymbol{K}_s^m - \boldsymbol{K}^{\mathrm{NTK}}\right\|_{\mathrm{op}}\left\|r^{\mathrm{NTK}}(s)\right\|\mathrm{d}s \\ & \times \exp\left(\int_0^t \frac{1}{n}\left\|\boldsymbol{K}_s^m - \boldsymbol{K}^{\mathrm{NTK}}\right\|_{\mathrm{op}}\mathrm{d}s\right).\end{aligned}$$

Let

$$\Delta = \sup_{t\geq 0}\sup_{\|x\|\leq A, \|x'\|\leq A}\left|K_t^m(x,x') - K^{\mathrm{NTK}}(x,x')\right|.$$

Using Equation (47) to bound $\left\|r^{\mathrm{NTK}}(s)\right\|$ and the fact that $\left\|\boldsymbol{K}_s^m - \boldsymbol{K}^{\mathrm{NTK}}\right\|_{\mathrm{op}} \leq n\Delta$, we obtain:

$$\|e(t)\| \leq \Delta t \exp\left(\left(\Delta - \frac{\lambda_0}{n}\right)t\right)\sqrt{n}B. \qquad \square$$

Lemma E.15 (Bound on $\|f_t^{\mathrm{NTK}}\|_{L^2(\mathcal{X})}$)**.** *Let $t \geq 0$. Under Assumptions a and b, it holds that:*

$$\left\|f_t^{\mathrm{NTK}}\right\|_{L^2(\mathcal{X})} \leq \kappa^{\mathrm{NTK}} B t,$$

where $\kappa^{\mathrm{NTK}} := (p+2)(A^2+1)+1$.

Proof. Let $x \in \mathcal{X}$ and $t \geq 0$. We recall that the gradient flow of f_t^{NTK} is the following:

$$\frac{\mathrm{d}}{\mathrm{d}t}f_t^{\mathrm{NTK}}(x) = -\frac{1}{n}\sum_{i=1}^n K^{\mathrm{NTK}}(x,x_i)r_i^{\mathrm{NTK}}(t), \tag{48}$$

where $r_i^{\mathrm{NTK}}(t) = f_t^{\mathrm{NTK}}(x_i) - y_i$. Using the fact that $f_0^{\mathrm{NTK}} = 0$ and the triangle inequality, we obtain:

$$\left|f_t^{\mathrm{NTK}}(x)\right| = \left|f_t^{\mathrm{NTK}}(x) - f_0^{\mathrm{NTK}}(x)\right| = \left|\int_0^t \frac{\mathrm{d}}{\mathrm{d}s}f_s^{\mathrm{NTK}}(x)\mathrm{d}s\right| \leq \int_0^t \left|\frac{\mathrm{d}}{\mathrm{d}s}f_s^{\mathrm{NTK}}(x)\right|\mathrm{d}s. \tag{49}$$

Next, we find an upper-bound on $\left|\frac{\mathrm{d}}{\mathrm{d}s}f_s^{\mathrm{NTK}}(x)\right|$. We have:

$$\left|\frac{\mathrm{d}}{\mathrm{d}s}f_s^{\mathrm{NTK}}(x)\right| = \frac{1}{n}\left|\sum_{i=1}^n K^{\mathrm{NTK}}(x,x_i)r_i^{\mathrm{NTK}}(s)\right| \leq \frac{\kappa^{\mathrm{NTK}}}{n}\sum_{i=1}^n\left|r_i^{\mathrm{NTK}}(s)\right|,$$

where we use Lemma E.2 to bound $K^{\mathrm{NTK}}(x,x_i)$ by $\kappa^{\mathrm{NTK}} := (p+2)(A^2+1)+1$ in the inequality. Using the notation $r^{\mathrm{NTK}}(s) = [r_1^{\mathrm{NTK}}(s), \ldots, r_n^{\mathrm{NTK}}(s)]^\top \in \mathbb{R}^n$, we obtain by the Cauchy-Schwarz inequality:

$$\left|\frac{\mathrm{d}}{\mathrm{d}s}f_s^{\mathrm{NTK}}(x)\right| \leq \frac{\kappa^{\mathrm{NTK}}}{\sqrt{n}}\left\|r^{\mathrm{NTK}}(s)\right\|. \tag{50}$$

By Lemma E.13, we get:

$$\left\|r^{\mathrm{NTK}}(s)\right\| \leq \exp\left(-\frac{\lambda_0}{n}s\right)\sqrt{n}B,$$

where λ_0 is the smallest eigenvalue of the matrix $\boldsymbol{K}^{\mathrm{NTK}}$. Substituting this bound into Equation (50), we get:

$$\left|\frac{\mathrm{d}}{\mathrm{d}s}f_s^{\mathrm{NTK}}(x)\right| \leq \kappa^{\mathrm{NTK}} B \exp\left(-\frac{\lambda_0}{n}s\right).$$

Inserting this bound into Equation (49) gives:

$$\left|f_t^{\mathrm{NTK}}(x)\right| \leq \kappa^{\mathrm{NTK}} B\int_0^t \exp\left(-\frac{\lambda_0}{n}s\right)\mathrm{d}s = \frac{\kappa^{\mathrm{NTK}}Bn}{\lambda_0}\left(1 - \exp\left(-\frac{\lambda_0}{n}t\right)\right) \leq \kappa^{\mathrm{NTK}}Bt,$$

where we use the identity $1 - \exp(-a) \leq a$ for all $a \geq 0$. Since μ is a probability measure, the claimed $L^2(\mathcal{X})$ bound follows. $\square$

Lemma E.16 (High-probability entrywise upper-bound on the parameters at initialization)**.** *Let $\theta_0 = (\boldsymbol{v}_0, \boldsymbol{u}_0, \boldsymbol{b}_0, a_0)$ be the parameters at initialization, where $\boldsymbol{u}_{0,i} \sim \mathcal{N}(0, \mathbb{1}_{p\times p})$, $b_{0,i} \sim \mathcal{N}(0,1)$, and $v_{0,i} \sim \mathcal{N}(0,1)$ are drawn independently for $i = 1, \dots, m/2$, with mirrored copies $\boldsymbol{u}_{0,i+m/2} = \boldsymbol{u}_{0,i}$, $b_{0,i+m/2} = b_{0,i}$, $v_{0,i+m/2} = -v_{0,i}$, and $a_0 = 0$. Let $m \geq 2$. Then, the following holds with probability at least $1 - \frac{p+2}{\sqrt{2\pi}} m^{-1/2}$ with respect to random initialization:*

$$\max_{1\leq i\leq m} \max\left(|v_{0,i}|, |b_{0,i}|, \|\boldsymbol{u}_{0,i}\|_\infty\right) \leq \sqrt{3\ln m}.$$

Proof. By the mirrored initialization, the parameters of neuron $i+m/2$ coincide with those of neuron i, up to the sign of the outer weight. It therefore suffices to control the $m(p+2)/2$ independent standard Gaussian variables $(u_{0,i})_j$, $b_{0,i}$, and $v_{0,i}$, for $1 \leq i \leq m/2$ and $1 \leq j \leq p$. For $Z \sim \mathcal{N}(0,1)$ and $t > 0$, the classical Gaussian tail bound gives $\Pr(|Z| \geq t) \leq \frac{2}{\sqrt{2\pi}t} \exp(-t^2/2)$. Since $m \geq 2$, we have $\sqrt{3\ln m} \geq \sqrt{3\ln 2} > 1$, hence $\Pr(|Z| \geq \sqrt{3\ln m}) \leq \frac{2}{\sqrt{2\pi}} \exp(-3\ln m/2) = \frac{2}{\sqrt{2\pi}} m^{-3/2}$. A union bound over the $m(p+2)/2$ variables shows that the bound fails with probability at most $\frac{m(p+2)}{2} \times \frac{2}{\sqrt{2\pi}} m^{-3/2} = \frac{p+2}{\sqrt{2\pi}} m^{-1/2}$, which concludes the proof. □

E.7 Uniform NTK approximation bounds

Proposition E.17, borrowed from [62], establishes that the empirical NTK of a sufficiently wide fully connected ReLU neural network with mirrored initialization converges uniformly to the infinite-width NTK, both over the entire training trajectory and over all inputs in a bounded domain. This is the key result that allows us to approximate the dynamics of the neural network by those of its NTK surrogate throughout training, which we use in Lemma E.18.

Proposition E.17 (Kernel uniform convergence – Proposition 45 in [62])**.** *Consider n i.i.d. samples $\{(x_i, y_i)\}_{i=1}^n$ and their corresponding augmented inputs $\{\tilde{x}_i\}_{i=1}^n$, where $\tilde{x}_i = [x_i^\top, 1]^\top$ for $1 \leq i \leq n$. Let $\tilde{M}_X = \sum_{i=1}^n \|\tilde{x}_i\|$, $\mathcal{B}_r = \{x \in \mathcal{X} \mid 1 \leq \|\tilde{x}\| \leq r\}$ for $r \geq 1$, and λ_0 be the smallest eigenvalue of the infinite-width NTK kernel matrix. There exist a constant $C > 0$ and a polynomial* poly *such that, for any $\delta \in (0,1)$ and $k > 0$, when the width satisfies $m \geq \mathrm{poly}(n, \tilde{M}_X, \lambda_0^{-1}, |y_1|, \dots, |y_n|, \ln(1/\delta), k)$ and $m \geq r^k$, with probability at least $1-\delta$ with respect to random samples and random initialization, we have:*

$$\sup_{t\geq 0} \sup_{x,x'\in\mathcal{B}_r} \left|K_t^m(x,x') - K^{\mathrm{NTK}}(x,x')\right| \leq Cr^2 m^{-\frac{1}{12}} \sqrt{\ln m},$$

where the constant $C > 0$ is independent of δ and n.

Lemma E.18 (Kernel uniform convergence – Application of Proposition E.17 to our setting)**.** *Consider $n \geq 3$ i.i.d. samples $\{(x_i, y_i)\}_{i=1}^n \sim \rho$ and fix $p \geq 2$. Under Assumptions a, b and d, there exist a constant $C > 0$ and a polynomial* poly *such that, for any $\delta \in (0,1)$, when the width satisfies $m \geq \mathrm{poly}(n, 1/\delta, \ln n, \ln(1/\delta))$, with probability at least $1-\delta$ with respect to random samples and random initialization, we have:*

$$\sup_{t\geq 0} \sup_{\|x\|\leq A, \|x'\|\leq A} \left|K_t^m(x,x') - K^{\mathrm{NTK}}(x,x')\right| \leq Cm^{-\frac{1}{12}} \sqrt{\ln m},$$

where the constant $C > 0$ is independent of δ and n, but depends on A, and the coefficients of poly *depend on A, B, M_μ, and p.*

Proof. In this proof, we apply the statement of Proposition E.17 to our setting. Under Assumption a, we know that $\|x\| \leq A$, for any $x \in \mathcal{X}$, which implies that $\tilde{M}_X = \sum_{i=1}^n \|\tilde{x}_i\| \leq n\sqrt{A^2+1}$. Moreover, under Assumption b, we know that $|y_i| \leq B$, for $1 \leq i \leq n$. By taking $k = 1$ and $r = \sqrt{A^2+1}$ in Proposition E.17, we require that m should satisfy $m \geq \mathrm{poly}(n, \lambda_0^{-1}, \ln(1/\delta))$ and $m \geq \sqrt{A^2+1}$, where the coefficients of poly depend on A and B. Hence, we get that $m \geq \max(\mathrm{poly}(n, \lambda_0^{-1}, \ln(1/\delta)), \sqrt{A^2+1})$. After redefining $\mathrm{poly}(n, \lambda_0^{-1}, \ln(1/\delta))$ to be the polynomial that dominates both lower bounds on m, we simplify the requirement to $m \geq \mathrm{poly}(n, \lambda_0^{-1}, \ln(1/\delta))$. By Corollary E.10, since $n \geq 3$, the following holds with probability at least $1-\delta$ with respect to random samples: $\lambda_0^{-1} \leq P(n, 1/\delta, \ln n, \ln(1/\delta))$, where $P(n, 1/\delta, \ln n, \ln(1/\delta)) :=$

$C' n^{8/p} \delta^{-4/p} \left(\ln n + \ln(1/\delta)\right)^3$ is a polynomial whose coefficients depend on A, M_μ, and p. Consequently, it suffices to choose $m \geq \mathrm{poly}(n, P(n, 1/\delta, \ln n, \ln(1/\delta)), \ln(1/\delta))$. After redefining $\mathrm{poly}(n, 1/\delta, \ln n, \ln(1/\delta))$ to be the polynomial $\mathrm{poly}(n, P(n, 1/\delta, \ln n, \ln(1/\delta)), \ln(1/\delta))$, we obtain that $m \geq \mathrm{poly}(n, 1/\delta, \ln n, \ln(1/\delta))$ where the coefficients of poly depend on A, B, M_μ, and p. Since we want both events in Proposition E.17 and Corollary E.10 to hold simultaneously, then, using a union bound, they both hold with probability at least $1 - 2\delta$. Redefining δ to be 2δ concludes the proof. □

Lemma E.19 (Uniform NTK approximation of the neural network predictor)**.** *Consider $n \geq 3$ i.i.d. samples $\{(x_i, y_i)\}_{i=1}^n \sim \rho$ and fix $p \geq 2$. Under Assumptions a, b and d, there exist a constant $C > 0$ and a polynomial* poly *such that, for any $\delta \in (0, 1)$, when the width satisfies $m \geq \mathrm{poly}(n, 1/\delta, \ln n, \ln(1/\delta))$, with probability at least $1 - \delta$ with respect to random samples and random initialization, we have:*

$$\sup_{t \geq 0} \sup_{\|x\| \leq A} \left| f_t^{\mathrm{NN}}(x) - f_t^{\mathrm{NTK}}(x) \right| \leq C m^{-\frac{1}{24}} \sqrt{\ln m},$$

where the constant $C > 0$ is independent of δ and n, but depends on A, and the coefficients of poly *depend on A, B, M_μ, and p.*

Proof. We start by recalling that f_0^{NN} and f_0^{NTK} are assumed to be equal to 0. Using this, for any $x \in \mathcal{X}$, we have:

$$\begin{aligned}
\left| f_t^{\mathrm{NN}}(x) - f_t^{\mathrm{NTK}}(x) \right| &= \left| \int_0^t \left(\frac{\mathrm{d}}{\mathrm{d}s} f_s^{\mathrm{NN}}(x) - \frac{\mathrm{d}}{\mathrm{d}s} f_s^{\mathrm{NTK}}(x) \right) \mathrm{d}s \right| \\
&\leq \int_0^t \left| \frac{\mathrm{d}}{\mathrm{d}s} f_s^{\mathrm{NN}}(x) - \frac{\mathrm{d}}{\mathrm{d}s} f_s^{\mathrm{NTK}}(x) \right| \mathrm{d}s,
\end{aligned}$$

where we use the triangle inequality in the second line. We recall that the gradient flows of f_s^{NN} and f_s^{NTK} are the following:

$$\frac{\mathrm{d}}{\mathrm{d}s} f_s^{\mathrm{NN}}(x) = -\frac{1}{n} K_s^m(x, \boldsymbol{X})^\top r^{\mathrm{NN}}(s) \quad \text{and} \quad \frac{\mathrm{d}}{\mathrm{d}s} f_s^{\mathrm{NTK}}(x) = -\frac{1}{n} K^{\mathrm{NTK}}(x, \boldsymbol{X})^\top r^{\mathrm{NTK}}(s),$$

where:

$$\begin{aligned}
K_s^m(x, \boldsymbol{X}) &= [K_s^m(x, x_1), \ldots, K_s^m(x, x_n)]^\top \in \mathbb{R}^n, \\
K^{\mathrm{NTK}}(x, \boldsymbol{X}) &= [K^{\mathrm{NTK}}(x, x_1), \ldots, K^{\mathrm{NTK}}(x, x_n)]^\top \in \mathbb{R}^n, \\
r^{\mathrm{NN}}(s) &= f_s^{\mathrm{NN}}(\boldsymbol{X}) - \boldsymbol{Y} = [f_s^{\mathrm{NN}}(x_1) - y_1, \ldots, f_s^{\mathrm{NN}}(x_n) - y_n]^\top \in \mathbb{R}^n, \\
r^{\mathrm{NTK}}(s) &= f_s^{\mathrm{NTK}}(\boldsymbol{X}) - \boldsymbol{Y} = [f_s^{\mathrm{NTK}}(x_1) - y_1, \ldots, f_s^{\mathrm{NTK}}(x_n) - y_n]^\top \in \mathbb{R}^n.
\end{aligned}$$

Consequently, we get:

$$\begin{aligned}
&\left| f_t^{\mathrm{NN}}(x) - f_t^{\mathrm{NTK}}(x) \right| \\
\leq &\int_0^t \left| \frac{\mathrm{d}}{\mathrm{d}s} f_s^{\mathrm{NN}}(x) - \frac{\mathrm{d}}{\mathrm{d}s} f_s^{\mathrm{NTK}}(x) \right| \mathrm{d}s \\
= &\int_0^t \left| -\frac{1}{n} K_s^m(x, \boldsymbol{X})^\top r^{\mathrm{NN}}(s) + \frac{1}{n} K^{\mathrm{NTK}}(x, \boldsymbol{X})^\top r^{\mathrm{NTK}}(s) \right| \mathrm{d}s \\
= &\frac{1}{n} \int_0^t \left| K_s^m(x, \boldsymbol{X})^\top r^{\mathrm{NN}}(s) - \left(K^{\mathrm{NTK}}(x, \boldsymbol{X}) - K_s^m(x, \boldsymbol{X}) + K_s^m(x, \boldsymbol{X}) \right)^\top r^{\mathrm{NTK}}(s) \right| \mathrm{d}s \\
= &\frac{1}{n} \int_0^t \left| K_s^m(x, \boldsymbol{X})^\top (r^{\mathrm{NN}}(s) - r^{\mathrm{NTK}}(s)) + \left(K_s^m(x, \boldsymbol{X}) - K^{\mathrm{NTK}}(x, \boldsymbol{X}) \right)^\top r^{\mathrm{NTK}}(s) \right| \mathrm{d}s \\
\leq &\frac{1}{n} \int_0^t \left(\left\| K_s^m(x, \boldsymbol{X}) - K^{\mathrm{NTK}}(x, \boldsymbol{X}) \right\| + \left\| K^{\mathrm{NTK}}(x, \boldsymbol{X}) \right\| \right) \left\| r^{\mathrm{NN}}(s) - r^{\mathrm{NTK}}(s) \right\| \mathrm{d}s \\
&+ \frac{1}{n} \int_0^t \left\| K_s^m(x, \boldsymbol{X}) - K^{\mathrm{NTK}}(x, \boldsymbol{X}) \right\| \left\| r^{\mathrm{NTK}}(s) \right\| \mathrm{d}s,
\end{aligned} \tag{51}$$

where we use the triangle inequality in the last inequality. Let:

$$\Delta = \sup_{t\geq 0} \sup_{\|x\|\leq A, \|x'\|\leq A} \left|K_t^m(x,x') - K^{\mathrm{NTK}}(x,x')\right|.$$

We notice that:

$$\left\|K_s^m(x,\boldsymbol{X}) - K^{\mathrm{NTK}}(x,\boldsymbol{X})\right\| = \sqrt{\sum_{i=1}^n \left|K_s^m(x,x_i) - K^{\mathrm{NTK}}(x,x_i)\right|^2} \leq \sqrt{n}\Delta, \tag{52}$$

$$\left\|K^{\mathrm{NTK}}(x,\boldsymbol{X})\right\| = \sqrt{\sum_{i=1}^n \left|K^{\mathrm{NTK}}(x,x_i)\right|^2} \leq \sqrt{n}\kappa^{\mathrm{NTK}}, \tag{53}$$

where in the second line we use Lemma E.2 to bound $\left|K^{\mathrm{NTK}}(x,x_i)\right|$ by $\kappa^{\mathrm{NTK}} := (p+2)(A^2+1)+1$. Moreover, by Lemma E.13, we get:

$$\left\|r^{\mathrm{NTK}}(s)\right\| \leq \exp\left(-\frac{\lambda_0}{n}s\right)\sqrt{n}B, \tag{54}$$

$$\left\|r^{\mathrm{NN}}(s) - r^{\mathrm{NTK}}(s)\right\| \leq \Delta s \exp(\gamma s)\sqrt{n}B, \tag{55}$$

where λ_0 is the smallest eigenvalue of the matrix $\boldsymbol{K}^{\mathrm{NTK}}$ and $\gamma = \Delta - \frac{\lambda_0}{n}$. Substituting Equations (52) to (55) into Equation (51), we obtain:

$$\left|f_t^{\mathrm{NN}}(x) - f_t^{\mathrm{NTK}}(x)\right| \leq \left(\Delta + \kappa^{\mathrm{NTK}}\right)\Delta B \int_0^t s\exp(\gamma s)\,\mathrm{d}s + \Delta B \int_0^t \exp\left(-\frac{\lambda_0}{n}s\right)\mathrm{d}s.$$

Applying integration by parts gives:

$$\begin{aligned}\int_0^t s\exp(\gamma s)\,\mathrm{d}s =& \left[\frac{s}{\gamma}\exp(\gamma s)\right]_{s=0}^{s=t} - \frac{1}{\gamma}\int_0^t \exp(\gamma s)\,\mathrm{d}s = \frac{t}{\gamma}\exp(\gamma t) - \frac{1}{\gamma^2}\left(\exp(\gamma t) - 1\right)\\ =& \frac{\exp(\gamma t)}{\gamma^2}(\gamma t - 1) + \frac{1}{\gamma^2}.\end{aligned}$$

Also, we have:

$$\int_0^t \exp\left(-\frac{\lambda_0}{n}s\right)\mathrm{d}s = \left[-\frac{n}{\lambda_0}\exp\left(-\frac{\lambda_0}{n}s\right)\right]_{s=0}^{s=t} = \frac{n}{\lambda_0}\left[1 - \exp\left(-\frac{\lambda_0}{n}t\right)\right] \leq \frac{n}{\lambda_0}.$$

Thus:

$$\left|f_t^{\mathrm{NN}}(x) - f_t^{\mathrm{NTK}}(x)\right| \leq \Delta B\left[\left(\Delta + \kappa^{\mathrm{NTK}}\right)\left(\frac{\exp(\gamma t)}{\gamma^2}(\gamma t - 1) + \frac{1}{\gamma^2}\right) + \frac{n}{\lambda_0}\right]. \tag{56}$$

By Lemma E.18, we know that with probability at least $1-\delta$ with respect to random samples and random initialization, it holds that $\Delta \leq Cm^{-\frac{1}{12}}\sqrt{\ln m}$ when m is large enough such that $m \geq \mathrm{poly}(n, 1/\delta, \ln n, \ln(1/\delta))$, where the constant $C > 0$ is independent of δ and n, but depends on A, and the coefficients of poly depend on A, B, M_μ, and p. When $m \geq \max\left(\left(\frac{2C\sqrt{12}n}{\lambda_0}\right)^{24}, 1\right)$, we obtain by Lemma I.9 that $\Delta \leq \frac{\lambda_0}{2n}$, which implies that $\gamma \leq -\frac{\lambda_0}{2n} < 0$. Similarly, when $m \geq \max\left(\left(C\sqrt{12}\right)^{24}, 1\right)$, it holds by Lemma I.9 that $\Delta \leq 1$. Hence, when $m \geq \max\left(\mathrm{poly}(n, 1/\delta, \ln n, \ln(1/\delta)), \left(\frac{2C\sqrt{12}n}{\lambda_0}\right)^{24}, \left(C\sqrt{12}\right)^{24}, 1\right)$, it holds that $\Delta \leq \min\left(Cm^{-\frac{1}{12}}\sqrt{\ln m}, \frac{\lambda_0}{2n}, 1\right)$. Consequently, with probability at least $1-\delta$ with respect to random samples and random initialization, Equation (56) simplifies to:

$$\begin{aligned}\left|f_t^{\mathrm{NN}}(x) - f_t^{\mathrm{NTK}}(x)\right| \leq& Cm^{-\frac{1}{12}}\sqrt{\ln m}B\left[\left(1+\kappa^{\mathrm{NTK}}\right)\frac{4n^2}{\lambda_0^2} + \frac{n}{\lambda_0}\right]\\ \leq& Cm^{-\frac{1}{12}}\sqrt{\ln m}Bn^2\left[\left(1+\kappa^{\mathrm{NTK}}\right)\frac{4}{\lambda_0^2} + \frac{1}{\lambda_0}\right].\end{aligned} \tag{57}$$

By Corollary E.10, since $n \geq 3$, the following holds with probability at least $1-\delta$ with respect to random samples: $\lambda_0^{-1} \leq \mathrm{poly}_1(n, 1/\delta, \ln n, \ln(1/\delta))$, where $\mathrm{poly}_1(n, 1/\delta, \ln n, \ln(1/\delta)) := C' n^{8/p} \delta^{-4/p} \left(\ln n + \ln(1/\delta)\right)^3$ is a polynomial whose coefficients depend on A, M_μ, and p. Consequently, we obtain:

$$\lambda_0^{-1} \leq \mathrm{poly}_1(n, 1/\delta, \ln n, \ln(1/\delta)) \quad \text{and} \quad \lambda_0^{-2} \leq \left(\mathrm{poly}_1(n, 1/\delta, \ln n, \ln(1/\delta))\right)^2. \tag{58}$$

Using this, we require that $m \geq P_1(n, 1/\delta, \ln n, \ln(1/\delta))$ where P_1 is a polynomial that dominates:

$$\max\left(\mathrm{poly}(n, 1/\delta, \ln n, \ln(1/\delta)), \left(2C\sqrt{12} n \mathrm{poly}_1\left(n, \frac{1}{\delta}, \ln n, \ln\frac{1}{\delta}\right)\right)^{24}, \left(C\sqrt{12}\right)^{24}, 1\right).$$

Plugging Equation (58) in Equation (57), we get:

$$\left|f_t^{\mathrm{NN}}(x) - f_t^{\mathrm{NTK}}(x)\right| \leq C m^{-\frac{1}{12}} \sqrt{\ln m} P_2\left(n, \frac{1}{\delta}, \ln n, \ln\frac{1}{\delta}\right),$$

where P_2 is a polynomial expressed as:

$$P_2\left(n, \frac{1}{\delta}, \ln n, \ln\frac{1}{\delta}\right) = Bn^2\Bigg[4(1+\kappa^{\mathrm{NTK}})\left(\mathrm{poly}_1\left(n, \frac{1}{\delta}, \ln n, \ln\frac{1}{\delta}\right)\right)^2 + \mathrm{poly}_1\left(n, \frac{1}{\delta}, \ln n, \ln\frac{1}{\delta}\right)\Bigg].$$

When $m \geq \max\left(P_1(n, 1/\delta, \ln n, \ln(1/\delta)), \left(P_2(n, 1/\delta, \ln n, \ln(1/\delta))\right)^{24}\right)$, we obtain:

$$\left|f_t^{\mathrm{NN}}(x) - f_t^{\mathrm{NTK}}(x)\right| \leq C m^{-\frac{1}{24}} \sqrt{\ln m}.$$

We conclude the proof by redefining poly to be the polynomial that dominates:

$$\max\left(P_1(n, 1/\delta, \ln n, \ln(1/\delta)), \left(P_2(n, 1/\delta, \ln n, \ln(1/\delta))\right)^{24}\right). \qquad \square$$

E.8 Generalization guarantees

E.8.1 High-probability generalization bounds

Corollary E.20, taken from [63], provides a high-probability bound on the generalization error of a broad class of *spectral algorithms* [63, Algorithm 1] applied to kernel regression in an RKHS. This class includes gradient methods [63, Example 3.2] and ridge regression [63, Example 3.3] as special cases, unified through the notion of a *filter function* [63, Definition 3.1]. The result shows that the estimator produced by any of these algorithms converges to the best approximation of the regression function in the RKHS at a rate that depends on the smoothness of the target function through the *source condition*, and on the complexity of the hypothesis space through the *capacity condition*. This is the key result that we use in Lemma E.21 to control the estimation error of the NTK gradient flow, by identifying the RKHS with the one associated to the *infinite-width* NTK and verifying that the assumptions required by this corollary are satisfied in our setting.

Corollary E.20 (Probabilistic generalization bound of the NTK predictor – Corollary 4.4 in [63])**.** *Let $\mathcal{H}$ be an RKHS associated with a Mercer kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $\mathcal{T} : L^2(\mathcal{X}) \to L^2(\mathcal{X})$ be the integral operator, and $\Sigma : \mathcal{H} \to \mathcal{H}$ be the covariance operator, all defined in Appendix H. Denote by $f_{\mathcal{H}}$ the projection of the regression function $f^\star$ onto the closure of $\mathcal{H}$ in $L^2(\mathcal{X})$. Given i.i.d. samples $\{(x_i, y_i)\}_{i=1}^n$ drawn from a distribution ρ with marginal μ on $\mathcal{X}$, we consider the following empirical risk minimization problem:*

$$f_\lambda = \operatorname*{argmin}_{f \in \mathcal{H}} \frac{1}{2n} \sum_{i=1}^n \left(f(x_i) - y_i\right)^2,$$

which we want to solve using a spectral algorithm with filter function G_λ indexed with $\lambda > 0$. We make the following assumptions:

Assumption 1. *(Bounded kernel) There exists a constant $\kappa \geq 1$ such that $K(x, x') \leq \kappa^2$ for any $x, x' \in \mathcal{X}$.*
Assumption 2. *(Moment condition) There exist positive constants $Q, J > 0$ such that for all integers $l \geq 2$, we have $\mathbb{E}_{y\sim\rho}[|y|^l \mid x] \leq \frac{1}{2} l! J^{l-2} Q^2$.*
Assumption 3. *(Regularity/smoothness of $f_{\mathcal{H}}$) $f_{\mathcal{H}}$ satisfies $\mathbb{E}_{x\sim\mu}[|f_{\mathcal{H}}(x) - f^\star(x)|^2 K(x,\cdot) \otimes K(x,\cdot)] \preceq U^2\Sigma$ and the following source condition $f_{\mathcal{H}} = \phi(\mathcal{T}) g_0$ with $\|g_0\|_{L^2(\mathcal{X})} \leq S$, where $U, S \geq 0$, $\phi(u) = \kappa^{-2\max(0,\zeta-1)} u^\zeta$, and $\zeta \geq 0$.*
Assumption 4. *(Capacity condition) For some $\gamma \in]0, 1]$ and $c_\gamma > 0$, the effective dimension of $\mathcal{H}$ satisfies $d_{\text{eff}}(\lambda) \leq c_\gamma \lambda^{-\gamma}$ for all $\lambda > 0$, where $d_{\text{eff}}(\lambda) = \text{Trace}(\Sigma(\Sigma + \lambda \mathbb{1}_{\mathcal{H}})^{-1})$.*

Let $a \in [0, \min(\frac{1}{2}, \zeta)]$ and $\lambda_{\text{op}} = n^{-\frac{1}{2\zeta+\gamma}}$. There exists a constant $C > 0$ such that, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ with respect to random samples, we have:

$$\left\| \mathcal{T}^{-a}(f_{\lambda_{\text{op}}} - f_{\mathcal{H}}) \right\|_{L^2(\mathcal{X})} \leq C \left(\ln \frac{6}{\delta} \right)^{2-a} n^{-\frac{\zeta-a}{2\zeta+\gamma}} \quad \text{if } 2\zeta + \gamma > 1,$$

where the constant $C > 0$ is independent of δ, n, and λ_{op}, but depends on κ, c_γ, γ, ζ, U, J, Q, S, and a.

Lemma E.21 (Probabilistic generalization bound of the NTK predictor – Application of Corollary E.20 to our setting)**.** *Consider n i.i.d. samples $\{(x_i, y_i)\}_{i=1}^n \sim \rho$. Let Assumptions a to d hold. If we choose a stopping time $t_{\text{op}} = n^{\frac{p+1}{s(p+1)+p}}$ with $\frac{1}{p+1} < s \leq 2$, then there exists a constant $C > 0$ such that, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ with respect to random samples, we have:*

$$\left\| f^{\text{NTK}}_{t_{\text{op}}} - f^\star \right\|_{L^2(\mathcal{X})} \leq C \left(\ln \frac{6}{\delta} \right)^2 n^{-\frac{1}{2}\frac{s(p+1)}{s(p+1)+p}},$$

where the constant $C > 0$ is independent of δ, n, and t_{op}, but depends on s, p, A, B, M_μ, and R.

Proof. In this proof, we apply the result in Corollary E.20 to our setting using our assumptions. For that, we set $a = 0$, $s = 2\zeta$ to be the smoothness parameter of the interpolation space $[\mathcal{H}^{\text{NTK}}]^s$, and $\beta := \frac{p+1}{p} = \frac{1}{\gamma}$ to be the eigenvalue decay rate of K^{NTK} as shown in Lemma E.12 under Assumptions a and d, where a, ζ, and γ are the notations used in Corollary E.20. We verify that the assumptions in Corollary E.20 hold in our setting for K^{NTK}.

- Assumption (**1**) is satisfied in our setting under Assumption a, as shown in Lemma E.2, since it suffices to choose $\kappa = \sqrt{(p+2)(A^2+1)+1} \geq 1$.
- Assumption (**2**) holds in our setting under Assumption b since the fact that $|y| \leq B$ implies $\mathbb{E}_{y\sim\rho}[|y|^l \mid x] \leq B^l = B^{l-2}B^2 \leq \frac{1}{2} l! B^{l-2} B^2$ for all $l \geq 2$, so it suffices to choose $J = B$ and $Q = B$.
- The first part of Assumption (**3**) is satisfied in our setting under Assumptions a and c for $U = 0$, because $f^\star$ coincides with its projection onto the closure of $\mathcal{H}^{\text{NTK}}$ in $L^2(\mathcal{X})$. In fact, we show in Lemma E.4 that $f^\star$ belongs to the closure of $\mathcal{H}^{\text{NTK}}$ in $L^2(\mathcal{X})$, which means that its projection onto that closure is itself. The second part of Assumption (**3**) is satisfied for $S = R$ with $0 \leq \zeta \leq 1$ under Assumption c. The condition on ζ implies that $0 \leq s \leq 2$.
- Assumption (**4**) is a consequence of Lemma E.12. In fact, letting $(\lambda_i)_{i\geq 1}$ denote the eigenvalues of Σ, we have the following upper-bound on the effective dimension of $\mathcal{H}$:

$$\begin{aligned} d_{\text{eff}}(\lambda) =& \text{Trace}(\Sigma(\Sigma + \lambda \mathbb{1}_{\mathcal{H}})^{-1}) = \sum_{i\geq 1} \frac{\lambda_i}{\lambda_i + \lambda} \leq \sum_{i\geq 1} \min\left(1, \frac{k_2 i^{-\beta}}{\lambda}\right) \\ &\leq \int_0^\infty \min\left(1, \frac{k_2 x^{-\beta}}{\lambda}\right) \mathrm{d}x = \frac{\beta}{\beta - 1} k_2^{\frac{1}{\beta}} \lambda^{-\frac{1}{\beta}}, \end{aligned}$$

  where we use $\frac{\lambda_i}{\lambda_i + \lambda} \leq \min\left(1, \frac{\lambda_i}{\lambda}\right)$, the upper bound $\lambda_i \leq k_2 i^{-\beta}$ of Lemma E.12, and the fact that $x \mapsto \min(1, k_2 x^{-\beta}/\lambda)$ is non-increasing. The integral equals $x_0 + \frac{x_0}{\beta - 1}$ with

$x_0 = (k_2/\lambda)^{1/\beta}$. This is precisely Assumption (4) with $c_\gamma = \frac{\beta}{\beta-1} k_2^{1/\beta} = (p+1)k_2^{\frac{p}{p+1}}$, since $\beta = \frac{p+1}{p}$.

It remains to identify the filter function and the optimal stopping time. According to [58, Lemma D.1], for a given $\lambda > 0$, the filter function of the gradient flow is given by:

$$G_\lambda(u) = \frac{1 - \exp(-\frac{u}{\lambda})}{u},$$

where $0 \le u \le \kappa^{\mathrm{NTK}}$ with $\kappa^{\mathrm{NTK}} := (p+2)(A^2+1)+1$ by Lemma E.2. From [58, Proof of Lemma D.1], the continuous time parameter is $t = \frac{1}{\lambda}$. Using the fact that $\lambda_{\mathrm{op}}^{-1} = n^{\frac{1}{2\zeta+\gamma}} = n^{\frac{\beta}{s\beta+1}}$ allows us to obtain that $t_{\mathrm{op}} = n^{\frac{\beta}{s\beta+1}} = n^{\frac{p+1}{s(p+1)+p}}$, where t_{op} is the stopping time of the gradient flow algorithm. It is important to note that the result in Corollary E.20 only holds for $s > 1 - \frac{1}{\beta} = \frac{1}{p+1}$. Combining this with $0 \le s \le 2$ provided earlier gives $\frac{1}{p+1} < s \le 2$. □

Proposition E.22 (Probabilistic generalization bound)**.** *Consider $n \ge 3$ i.i.d. samples $\{(x_i, y_i)\}_{i=1}^n \sim \rho$ and fix $p \ge 2$. Let Assumptions a to d hold. If we choose a stopping time $t_{\mathrm{op}} = n^{\frac{p+1}{s(p+1)+p}}$ with $\frac{1}{p+1} < s \le 2$, then there exist a constant $C > 0$ and a polynomial* poly *such that, for any $\delta \in (0,1)$, when the width satisfies $m \ge \mathrm{poly}(n, 1/\delta, \ln n, \ln(1/\delta))$, with probability at least $1-\delta$ with respect to random samples and random initialization, we have:*

$$\left\| f_{t_{\mathrm{op}}}^{\mathrm{NN}} - f^\star \right\|_{L^2(\mathcal{X})} \le C \left(\ln \frac{6}{\delta} \right)^2 n^{-\frac{1}{2}\frac{s(p+1)}{s(p+1)+p}},$$

where the constant $C > 0$ is independent of δ, n and t_{op}, but depends on s, p, A, B, M_μ, and R, and the coefficients of poly *depend on A, B, M_μ, and p.*

Proof. By the triangle inequality and the identity $\|\cdot\|_{L^2(\mathcal{X})} \le \|\cdot\|_\infty$, we have:

$$\begin{aligned} \left\| f_{t_{\mathrm{op}}}^{\mathrm{NN}} - f^\star \right\|_{L^2(\mathcal{X})} &\le \left\| f_{t_{\mathrm{op}}}^{\mathrm{NN}} - f_{t_{\mathrm{op}}}^{\mathrm{NTK}} \right\|_{L^2(\mathcal{X})} + \left\| f_{t_{\mathrm{op}}}^{\mathrm{NTK}} - f^\star \right\|_{L^2(\mathcal{X})} \\ &\le \left\| f_{t_{\mathrm{op}}}^{\mathrm{NN}} - f_{t_{\mathrm{op}}}^{\mathrm{NTK}} \right\|_{\infty} + \left\| f_{t_{\mathrm{op}}}^{\mathrm{NTK}} - f^\star \right\|_{L^2(\mathcal{X})}. \end{aligned} \tag{59}$$

Applying Lemma E.21, we obtain under Assumptions a to d that there exists a constant $C_1 > 0$ such that, for any $\delta \in (0,1)$, the following holds with probability at least $1 - \frac{\delta}{2}$ with respect to random samples:

$$\left\| f_{t_{\mathrm{op}}}^{\mathrm{NTK}} - f^\star \right\|_{L^2(\mathcal{X})} \le C_1 \left(\ln \frac{6}{\delta} \right)^2 n^{-\frac{1}{2}\frac{s(p+1)}{s(p+1)+p}}, \tag{60}$$

where $C_1 > 0$ is a constant independent of δ, n, and t_{op}, and that depends on s, p, A, B, M_μ, and R. By Lemma E.19, under Assumptions a, b and d, we know that there exist a constant $C_2 > 0$ and a polynomial P such that, for any $\delta \in (0,1)$, when $m \ge P(n, 1/\delta, \ln n, \ln(1/\delta))$, with probability at least $1 - \frac{\delta}{2}$ with respect to random samples and random initialization, we have:

$$\left\| f_{t_{\mathrm{op}}}^{\mathrm{NN}} - f_{t_{\mathrm{op}}}^{\mathrm{NTK}} \right\|_\infty \le \sup_{\|x\| \le A} \left| f_{t_{\mathrm{op}}}^{\mathrm{NN}}(x) - f_{t_{\mathrm{op}}}^{\mathrm{NTK}}(x) \right| \le C_2 m^{-\frac{1}{24}} \sqrt{\ln m},$$

where the constant $C_2 > 0$ is independent of δ and n, but depends on A, and the coefficients of P depend on A, B, M_μ, and p. Choosing $m \ge \max\left(P(n, 1/\delta, \ln n, \ln(1/\delta)), n^{24}(\ln n)^{12}\right)$ yields $m^{-\frac{1}{24}}\sqrt{\ln m} \le 6n^{-1}$ by Lemma I.8 since $n \ge 3$. This implies that:

$$\left\| f_{t_{\mathrm{op}}}^{\mathrm{NN}} - f_{t_{\mathrm{op}}}^{\mathrm{NTK}} \right\|_\infty \le 6C_2 n^{-1}. \tag{61}$$

Substituting the bounds in Equations (60) and (61) into Equation (59), we obtain that:

$$\left\| f_{t_{\mathrm{op}}}^{\mathrm{NN}} - f^\star \right\|_{L^2(\mathcal{X})} \le 6C_2 n^{-1} + C_1 \left(\ln \frac{6}{\delta} \right)^2 n^{-\frac{1}{2}\frac{s(p+1)}{s(p+1)+p}}.$$

Noticing that $n^{-1} \leq n^{-\frac{1}{2}\frac{s(p+1)}{s(p+1)+p}}$, we get:

$$\left\| f_{t_{\mathrm{op}}}^{\mathrm{NN}} - f^\star \right\|_{L^2(\mathcal{X})} \leq \left(6C_2 + C_1 \left(\ln \frac{6}{\delta} \right)^2 \right) n^{-\frac{1}{2}\frac{s(p+1)}{s(p+1)+p}}.$$

Given that $\delta \in (0,1)$, it holds that $\ln \frac{6}{\delta} > \ln 6 > 1$. Hence, we get $\left(\ln \frac{6}{\delta}\right)^2 > 1$. Therefore, $6C_2 \leq 6C_2 \left(\ln \frac{6}{\delta}\right)^2$. As a result, combining the terms gives:

$$\left\| f_{t_{\mathrm{op}}}^{\mathrm{NN}} - f^\star \right\|_{L^2(\mathcal{X})} \leq (6C_2 + C_1) \left(\ln \frac{6}{\delta} \right)^2 n^{-\frac{1}{2}\frac{s(p+1)}{s(p+1)+p}}.$$

Defining the constant $C := 6C_2 + C_1$ establishes the desired convergence rate. Since we want both bounds in Equations (60) and (61) to hold simultaneously, then, using a union bound, they both hold with probability at least $1 - \delta$. We conclude the proof by choosing $m \geq \mathrm{poly}\left(n, 1/\delta, \ln n, \ln(1/\delta)\right)$, where poly is a polynomial that dominates $\max\left(P(n, 1/\delta, \ln n, \ln(1/\delta)), n^{24}(\ln n)^{12}\right)$. □

E.8.2 Expected generalization bounds

Deriving generalization bounds in expectation is essential to characterize the convergence rate of optimization algorithms, whose guarantees are naturally stated in expectation. However, the standard approach of converting high-probability bounds to expected bounds via $\mathbb{E}[X] = \int_0^\infty \Pr(X > t)\mathrm{d}t$ for $X \geq 0$ does not apply here, as the required network width m depends on δ (see Proposition E.22) in a way that prevents a direct integration over δ. To overcome this, we introduce truncated neural network predictors, obtained by setting the networks to zero on a carefully defined low-probability event, which allows us to pass from high-probability bounds to expected bounds while maintaining control over the contribution of the low-probability event. To this end, we introduce six events that will be used to control the generalization error of the early-stopped neural network in expectation.

High-probability events governing the generalization error. Recall that $f_{t_{\mathrm{op}}}^{\mathrm{NN}}$ denotes the neural network trained by gradient flow and stopped at time t_{op}. To analyze the generalization error of $f_{t_{\mathrm{op}}}^{\mathrm{NN}}$, we decompose it as:

$$\left\| f_{t_{\mathrm{op}}}^{\mathrm{NN}} - f^\star \right\|_{L^2(\mathcal{X})} \leq \left\| f_{t_{\mathrm{op}}}^{\mathrm{NN}} - f_{t_{\mathrm{op}}}^{\mathrm{NTK}} \right\|_{L^2(\mathcal{X})} + \left\| f_{t_{\mathrm{op}}}^{\mathrm{NTK}} - f^\star \right\|_{L^2(\mathcal{X})},$$

where $f_{t_{\mathrm{op}}}^{\mathrm{NTK}}$ denotes the corresponding NTK gradient flow stopped at the same time t_{op}. The first term measures how well the finite-width neural network remains uniformly close to its infinite-width NTK surrogate, while the second term is the estimation error of the NTK gradient flow itself. The following are the two events under which each of these two terms can be made small.

1. $\mathcal{E}_{\mathrm{NTK}} = \left\{ \left\| f_{t_{\mathrm{op}}}^{\mathrm{NTK}} - f^\star \right\|_{L^2(\mathcal{X})} \leq C_1 \left(\ln \frac{6}{\delta} \right)^2 n^{-\frac{1}{2}\frac{s(p+1)}{s(p+1)+p}} \right\}$, where $\delta \in (0,1)$ and $C_1 > 0$ is a constant that is independent of δ, n, and t_{op}, but depends on s, p, A, B, M_μ, and R. It is the event that the NTK gradient flow achieves the target estimation rate. By Lemma E.21, $\mathcal{E}_{\mathrm{NTK}}$ holds with probability at least $1 - \delta$ with respect to random samples.
2. $\mathcal{E}_{\mathrm{NN}} = \left\{ \left\| f_{t_{\mathrm{op}}}^{\mathrm{NN}} - f_{t_{\mathrm{op}}}^{\mathrm{NTK}} \right\|_{L^2(\mathcal{X})} \leq C_2 m^{-\frac{1}{24}} \sqrt{\ln m} \right\}$, where $C_2 > 0$ is a constant that is independent of δ, n and t_{op} with $\delta \in (0,1)$, but depends on A. By Lemma E.19, for $n \geq 3$, $\mathcal{E}_{\mathrm{NN}}$ holds with probability at least $1 - \delta$ with respect to random samples and random initialization, provided that $m \geq \mathrm{poly}(n, 1/\delta, \ln n, \ln(1/\delta))$, where the coefficients of poly depend on A, B, M_μ, and p.

Rather than bounding the probability of $\mathcal{E}_{\mathrm{NN}}$ directly, we instead exhibit the following events, each depending only on the random initialization or the empirical NTK at initialization, and whose intersection is included in $\mathcal{E}_{\mathrm{NN}}$, as we discuss in Remark E.23.

3. $\mathcal{E}_{\theta_0} = \left\{ \max_{1 \leq i \leq m} \max\left(|v_{0,i}|, |b_{0,i}|, \|\boldsymbol{u}_{0,i}\|_\infty \right) \leq \sqrt{3 \ln m} \right\}$ is the event that no parameter of the initialization θ_0 is not atypically large, ensuring that the parameters remain in a controlled neighborhood of θ_0 during training. By Lemma E.16, this event holds with probability at least $1 - \frac{p+2}{\sqrt{2\pi}} m^{-1/2}$ with respect to random initialization.

4. $\mathcal{R} = \left\{ \mathrm{Card}\left(\left\{ i \in \{1,\dots,m\} : \left| \tilde{\boldsymbol{u}}_{0,i}^\top \tilde{x} \right| \le \frac{\sqrt{2\pi}}{4} m^{-\eta} \right\} \right) \le \lfloor m^\kappa \rfloor,\ \forall x \in \mathcal{X} \right\}$, where $\tilde{\boldsymbol{u}}_{0,i} = [\boldsymbol{u}_{0,i}^\top, b_{0,i}]^\top$ are the augmented first-layer weights at initialization, $\tilde{x} = [x^\top, 1]^\top$ is the augmented input, $0 < \eta < \varrho \le 1$, $\tau > \frac{1}{2}$, and $\kappa > \max(1-\eta, \tau)$. It is the event that, uniformly over all inputs, at most $\lfloor m^\kappa \rfloor$ neurons have a *small pre-activation* value $\left| \tilde{\boldsymbol{u}}_{0,i}^\top \tilde{x} \right|$ at initialization. For the remaining neurons, the pre-activation value is large enough that changing their activation pattern would require the parameters to travel far from their initialization. Consequently, $\mathcal{R}$ guarantees that the activation pattern of the majority of neurons is effectively frozen throughout training, which is precisely the condition needed for the network to operate in the *lazy training NTK regime*. By [58, Lemma C.2], when the width m is sufficiently large such that:

$$m^{1-\eta} + m^\tau \le \lfloor m^\kappa \rfloor \qquad \text{and} \qquad \frac{2Ap\sqrt{\ln m}}{\lfloor m^\varrho \rfloor} \le \frac{\sqrt{2\pi}}{4} m^{-\eta}, \tag{62}$$

the event $\mathcal{R}$ holds with probability at least $1 - m^{p\varrho} \exp\left(-2m^{2\tau-1}\right)$ with respect to random initialization. By Lemma I.10, the conditions on m in Equation (62) are satisfied when m is sufficiently large such that $m \ge \max\left(2^{\frac{1}{\varrho}}, \left(\frac{128A^2p^2}{\pi(\varrho-\eta)} \right)^{\frac{1}{\varrho-\eta}}, 3^{\frac{1}{\kappa-\max(1-\eta,\tau)}} \right)$.

5. $\mathcal{C} = \left\{ \sup_{z,z' \in \mathcal{N}_\epsilon} \left| K_0^m(z,z') - K^{\mathrm{NTK}}(z,z') \right| \le C_3 \sqrt{\frac{\ln m}{m}} \right\}$, where $\mathcal{N}_\epsilon$ denotes a uniform grid on the compact $\mathcal{X}$ with mesh size $\epsilon = \frac{2A}{\lfloor m^\varrho \rfloor}$ where $0 < \varrho \le 1$, and $C_3 > 0$ is a constant that depends on p, A, and ϱ. It is the event that the empirical NTK at initialization K_0^m concentrates uniformly around the infinite-width NTK K^{NTK} over the grid $\mathcal{N}_\epsilon$ with a discrepancy of order $\sqrt{\frac{\ln m}{m}}$. By [58, Lemma C.3], when m is large enough such that $\frac{3p\varrho}{c_0} \ln m \le m$ with $c_0 > 0$ being an absolute positive constant, the event $\mathcal{C}$ holds with probability at least $1 - 4m^{-p\varrho}$ with respect to random initialization. By Lemma I.10, it is enough to require $m \ge \max\left(\frac{36p^2\varrho^2}{c_0^2}, 1 \right)$ to ensure that $\frac{3p\varrho}{c_0} \ln m \le m$.

Controlling the smallest eigenvalue λ_0 of the infinite-width NTK matrix $\boldsymbol{K}^{\mathrm{NTK}}$ is essential for the convergence of the gradient flow. To see this, recall from Lemma E.13 that the explicit solution of the NTK gradient flow at the training points $(\boldsymbol{X}, \boldsymbol{Y}) = \{(x_i, y_i)\}_{i=1}^n$ satisfies at $t \ge 0$:

$$\left\| f_t^{\mathrm{NTK}}(\boldsymbol{X}) - \boldsymbol{Y} \right\| \le \exp\left(-\frac{\lambda_0}{n} t \right) \sqrt{n} B.$$

A positive lower bound $\lambda_0 \ge \lambda_\star > 0$ therefore guarantees exponential convergence of the training error at rate $\frac{\lambda_\star}{n}$. If $\lambda_0 = 0$, the training error never decays, and the gradient flow fails to converge entirely. Since the samples $x_1, \dots, x_n$ are drawn randomly, the smallest eigenvalue λ_0 of the infinite-width NTK at initialization is itself a random quantity, and must therefore be lower-bounded with high probability, which motivates the following event.

6. $\mathcal{E}_{\lambda_0} = \{\lambda_0 \ge \lambda_\star(n,\delta)\}$, where $\lambda_\star(n,\delta) := C_4\big(1 + g(\epsilon_\star(n,\delta))\big)^{-3} \epsilon_\star(n,\delta)^4 > 0$, $\epsilon_\star(n,\delta) := \min\big(\epsilon_A,\ \epsilon_0(n,\delta)\big) \in (0,1]$, $\epsilon_A := \min\left(1, \frac{2}{\sqrt{A^2+1}}\right)$, $\epsilon_0(n,\delta) := \frac{2}{\pi}\left(\frac{2\delta}{M_\mu (A^2+1)^{(p+1)/2} v_p n^2} \right)^{\frac{1}{p}}$, $g(\epsilon) := \frac{(p+1)\ln(1/\epsilon)}{\ln(p+1)}$, $\delta \in (0,1)$, $v_p = \mathrm{Vol}(B(0,1))$ is the volume of the unit ball in $\mathbb{R}^p$, M_μ is the bound on the probability density function f_μ, and $C_4 > 0$ is an absolute positive constant. It is the event that the smallest eigenvalue of the infinite-width NTK at initialization is bounded away from zero, a necessary condition for the gradient flow to converge. By Lemma E.9, $\mathcal{E}_{\lambda_0}$ holds with probability at least $1 - \delta$ with respect to random samples.

Remark E.23 (Inclusion of events). According to [58, Proof of Proposition 3.3], we have the following inclusion:

$$\mathcal{E}_{\theta_0} \cap \mathcal{R} \cap \mathcal{C} \subset \left\{ \sup_{t \ge 0} \sup_{\|x\| \le A, \|x'\| \le A} \left| K_t^m(x,x') - K^{\mathrm{NTK}}(x,x') \right| \le C m^{-\frac{1}{12}} \sqrt{\ln m} \right\}.$$

Furthermore, in the proof of Lemma E.19, we show that:

$$\left\{\sup_{t\geq 0}\sup_{\|x\|\leq A,\|x'\|\leq A}\left|K_t^m(x,x')-K^{\mathrm{NTK}}(x,x')\right|\leq Cm^{-\frac{1}{12}}\sqrt{\ln m}\right\}\cap\mathcal{E}_{\lambda_0}\subset\mathcal{E}_{\mathrm{NN}}.$$

Consequently, we obtain $\mathcal{E}:=\mathcal{E}_{\theta_0}\cap\mathcal{R}\cap\mathcal{C}\cap\mathcal{E}_{\lambda_0}\subset\mathcal{E}_{\mathrm{NN}}$. This inclusion motivates the following definition, in which we truncate the neural network on $\mathcal{E}$ rather than on $\mathcal{E}_{\mathrm{NN}}$ directly, since $\mathcal{E}$ is contained in $\mathcal{E}_{\mathrm{NN}}$ and its probability can be bounded explicitly from its four constituent parts.

Definition E.24 (Truncated neural network)**.** Let $\mathcal{E}=\mathcal{E}_{\theta_0}\cap\mathcal{R}\cap\mathcal{C}\cap\mathcal{E}_{\lambda_0}$. For a chosen stopping time t_{op}, the truncated neural network associated with $f_{t_{\mathrm{op}}}^{\mathrm{NN}}$ is defined by $\widetilde{f}_{t_{\mathrm{op}}}^{\mathrm{NN}}:=f_{t_{\mathrm{op}}}^{\mathrm{NN}}\mathbb{1}_{\mathcal{E}}$.

The event $\mathcal{E}$ and the truncated predictor depend on a deterministic confidence parameter $\delta\in(0,1)$, through $\lambda_\star(n,\delta)$. The inner and outer estimators in Equation (7) choose δ to be the corresponding minimizer $\delta^\star$, specified in Equations (29) and (68) in the proofs of Propositions D.4 and E.25, respectively.

By truncating the neural network on $\mathcal{E}$, the truncated predictor $\widetilde{f}_{t_{\mathrm{op}}}^{\mathrm{NN}}$ vanishes on $\mathcal{E}^{\complement}$, which allows us to pass from the high-probability bound of Proposition E.22 to the expected bound of Proposition E.25 by decomposing the expected error into a contribution on $\mathcal{E}$, controlled by the high-probability bound, and a contribution on $\mathcal{E}^{\complement}$, which vanishes by definition of the truncation. The explicit probability bounds on each constituent event of $\mathcal{E}$ are then used to show that $\Pr(\mathcal{E}^{\complement})$ is small enough to be absorbed into the overall rate.

Proposition E.25 (Expected generalization bound)**.** *Consider n i.i.d. samples $\{(x_i,y_i)\}_{i=1}^n\sim\rho$ and fix $p\geq 2$. Let Assumptions a to d hold. If we choose a stopping time $t_{\mathrm{op}}=n^{\frac{p+1}{s(p+1)+p}}$ with $\frac{1}{p+1}<s\leq 2$, then, considering the early-stopped truncated neural network predictor $\widetilde{f}_{t_{\mathrm{op}}}^{\mathrm{NN}}$ for the target function $f^\star$, there exist a constant $C>0$ and a polynomial* poly *such that when $n\geq\max\left(3,\left\lceil\left(\frac{3C_1^2e^3}{2(\kappa^{\mathrm{NTK}})^2B^2}\right)^{\frac{s(p+1)+p}{(s+2)(p+1)}}\right\rceil,\left\lceil\left(\frac{1}{3(\kappa^{\mathrm{NTK}})^2}\right)^{\frac{s(p+1)+p}{2(p+1)}}\right\rceil,\left\lfloor\left(\frac{C_1^2(\ln 6)^3}{3(\kappa^{\mathrm{NTK}})^2B^2}\right)^{\frac{s(p+1)+p}{(s+2)(p+1)}}\right\rfloor+1\right)$ and the width satisfies $m\geq\mathrm{poly}(n,\ln n)$, the following holds:*

$$\mathbb{E}\left[\left\|\widetilde{f}_{t_{\mathrm{op}}}^{\mathrm{NN}}-f^\star\right\|_{L^2(\mathcal{X})}^2\right]\leq C\left(\ln n\right)^4 n^{-\frac{s(p+1)}{s(p+1)+p}},$$

where $\kappa^{\mathrm{NTK}}:=(p+2)(A^2+1)+1$, the constant $C>0$ is independent of n and t_{op}, but depends on s, p, A, B, M_μ, and R, and the coefficients of poly *depend on s, p, A, B, M_μ, and R. The expectation here is with respect to random samples and random initialization.*

Proof. By the triangle inequality and the identity $(a+b)^2\leq 2(a^2+b^2)$, we have:

$$\mathbb{E}\left[\left\|\widetilde{f}_{t_{\mathrm{op}}}^{\mathrm{NN}}-f^\star\right\|_{L^2(\mathcal{X})}^2\right]\leq 2\left(\mathbb{E}\left[\left\|\widetilde{f}_{t_{\mathrm{op}}}^{\mathrm{NN}}-f_{t_{\mathrm{op}}}^{\mathrm{NTK}}\right\|_{L^2(\mathcal{X})}^2\right]+\mathbb{E}\left[\left\|f_{t_{\mathrm{op}}}^{\mathrm{NTK}}-f^\star\right\|_{L^2(\mathcal{X})}^2\right]\right). \tag{63}$$

Next, we upper-bound each expectation on the right-hand side of the inequality, and then combine the resulting bounds.

Upper-bound on $\mathbb{E}\left[\left\|\widetilde{f}_{t_{\mathrm{op}}}^{\mathrm{NN}}-f_{t_{\mathrm{op}}}^{\mathrm{NTK}}\right\|_{L^2(\mathcal{X})}^2\right]$. Let $\mathcal{E}=\mathcal{E}_{\theta_0}\cap\mathcal{R}\cap\mathcal{C}\cap\mathcal{E}_{\lambda_0}$. By the law of total expectation, we decompose the error over the disjoint events $\mathcal{E}$ and $\mathcal{E}^{\complement}$ as follows:

$$\begin{aligned}
\mathbb{E}\left[\left\|\widetilde{f}_{t_{\mathrm{op}}}^{\mathrm{NN}}-f_{t_{\mathrm{op}}}^{\mathrm{NTK}}\right\|_{L^2(\mathcal{X})}^2\right]=&\mathbb{E}\left[\left\|\widetilde{f}_{t_{\mathrm{op}}}^{\mathrm{NN}}-f_{t_{\mathrm{op}}}^{\mathrm{NTK}}\right\|_{L^2(\mathcal{X})}^2\mathbb{1}_{\mathcal{E}}\right]+\mathbb{E}\left[\left\|\widetilde{f}_{t_{\mathrm{op}}}^{\mathrm{NN}}-f_{t_{\mathrm{op}}}^{\mathrm{NTK}}\right\|_{L^2(\mathcal{X})}^2\mathbb{1}_{\mathcal{E}^{\complement}}\right]\\
=&\mathbb{E}\left[\left\|f_{t_{\mathrm{op}}}^{\mathrm{NN}}-f_{t_{\mathrm{op}}}^{\mathrm{NTK}}\right\|_{L^2(\mathcal{X})}^2\mathbb{1}_{\mathcal{E}}\right]+\mathbb{E}\left[\left\|f_{t_{\mathrm{op}}}^{\mathrm{NTK}}\right\|_{L^2(\mathcal{X})}^2\mathbb{1}_{\mathcal{E}^{\complement}}\right]\\
\leq&\mathbb{E}\left[\left\|f_{t_{\mathrm{op}}}^{\mathrm{NN}}-f_{t_{\mathrm{op}}}^{\mathrm{NTK}}\right\|_{L^2(\mathcal{X})}^2\mathbb{1}_{\mathcal{E}_{\mathrm{NN}}}\right]+(\kappa^{\mathrm{NTK}})^2B^2t_{\mathrm{op}}^2\Pr\left(\mathcal{E}^{\complement}\right)\\
\leq&C_2^2m^{-\frac{1}{12}}\ln m+(\kappa^{\mathrm{NTK}})^2B^2n^{\frac{2(p+1)}{s(p+1)+p}}\Pr\left(\mathcal{E}^{\complement}\right),
\end{aligned}\tag{64}$$

where, in the second line, we use the definition of the truncated neural network in Definition E.24 which gives $\widetilde{f}^{\mathrm{NN}}_{t_{\mathrm{op}}} = 0$ on $\mathcal{E}^{\complement}$; in the third line, we use $\mathcal{E} \subset \mathcal{E}_{\mathrm{NN}}$ as explained in Remark E.23, and Lemma E.15 to upper-bound $\left\| f^{\mathrm{NTK}}_{t_{\mathrm{op}}} \right\|_{L^2(\mathcal{X})}$ by $\kappa^{\mathrm{NTK}} B t_{\mathrm{op}}$ with $\kappa^{\mathrm{NTK}} := (p+2)(A^2+1)+1$; and in the fourth line, we upper-bound $\left\| \widetilde{f}^{\mathrm{NN}}_{t_{\mathrm{op}}} - f^{\mathrm{NTK}}_{t_{\mathrm{op}}} \right\|^2_{L^2(\mathcal{X})}$ by $C_2^2 m^{-\frac{1}{12}} \ln m$ on the event $\mathcal{E}_{\mathrm{NN}}$ and use $t_{\mathrm{op}} = n^{\frac{p+1}{s(p+1)+p}}$. Furthermore, by De Morgan's law and a union bound, we have that:

$$
\begin{aligned}
\Pr\left(\mathcal{E}^{\complement}\right) =& \Pr\left((\mathcal{E}_{\theta_0} \cap \mathcal{R} \cap \mathcal{C} \cap \mathcal{E}_{\lambda_0})^{\complement}\right) \\
=& \Pr\left(\mathcal{E}^{\complement}_{\theta_0} \cup \mathcal{R}^{\complement} \cup \mathcal{C}^{\complement} \cup \mathcal{E}^{\complement}_{\lambda_0}\right) \\
\leq& \Pr\left(\mathcal{E}^{\complement}_{\theta_0}\right) + \Pr\left(\mathcal{R}^{\complement}\right) + \Pr\left(\mathcal{C}^{\complement}\right) + \Pr\left(\mathcal{E}^{\complement}_{\lambda_0}\right) \\
\leq& \frac{p+2}{\sqrt{2\pi}} m^{-1/2} + m^{p\varrho} \exp\left(-2m^{2\tau-1}\right) + 4m^{-p\varrho} + \delta.
\end{aligned}
$$

We note that $\frac{p+2}{\sqrt{2\pi}} m^{-1/2} \leq \delta$ and $4m^{-p\varrho} \leq \delta$ when we choose m large enough such that $m \geq \frac{(p+2)^2}{2\pi\delta^2}$ and $m \geq \left(\frac{4}{\delta}\right)^{\frac{1}{p\varrho}}$, respectively. Moreover, by Lemma I.7, $m^{p\varrho} \exp\left(-2m^{2\tau-1}\right) \leq \delta$ holds whenever m is large enough such that $m \geq \max(1, M_1(1/\delta))$ where:

$$
M_1\left(\frac{1}{\delta}\right) = \left(\left[\frac{p\varrho}{2\tau-1}\left(\ln\left(\frac{p\varrho}{2\tau-1}\right) - 1\right) + \ln\left(\frac{1}{\delta}\right)\right]_+\right)^{\frac{1}{2\tau-1}}, \qquad \text{with } [t]_+ := \max(t, 0).
$$

$\mathcal{R}$ and $\mathcal{C}$ hold with high-probability when $m \geq \max\left(2^{\frac{1}{\varrho}}, \left(\frac{128A^2p^2}{\pi(\varrho-\eta)}\right)^{\frac{1}{\varrho-\eta}}, 3^{\frac{1}{\kappa-\max(1-\eta,\tau)}}\right)$ and $m \geq \max\left(\frac{36p^2\varrho^2}{c_0^2}, 1\right)$, respectively, where c_0 is an absolute positive constant, as discussed in Items 4 and 5. Combining all these conditions, we conclude that when m is large enough such that $m \geq M_2(1/\delta)$ where:

$$
M_2\left(\frac{1}{\delta}\right) = \max\left(\frac{(p+2)^2}{2\pi\delta^2}, \left(\frac{4}{\delta}\right)^{\frac{1}{p\varrho}}, M_1\left(\frac{1}{\delta}\right), 2^{\frac{1}{\varrho}}, \left(\frac{128A^2p^2}{\pi(\varrho-\eta)}\right)^{\frac{1}{\varrho-\eta}}, 3^{\frac{1}{\kappa-\max(1-\eta,\tau)}}, \frac{36p^2\varrho^2}{c_0^2}\right),
$$

we get $\Pr(\mathcal{E}^{\complement}) \leq 4\delta$. Substituting this into Equation (64), we obtain:

$$
\mathbb{E}\left[\left\| \widetilde{f}^{\mathrm{NN}}_{t_{\mathrm{op}}} - f^{\mathrm{NTK}}_{t_{\mathrm{op}}} \right\|^2_{L^2(\mathcal{X})}\right] \leq C_2^2 m^{-\frac{1}{12}} \ln m + 4(\kappa^{\mathrm{NTK}})^2 B^2 n^{\frac{2(p+1)}{s(p+1)+p}} \delta. \tag{65}
$$

Since $m \geq \mathrm{poly}(n, 1/\delta, \ln n, \ln(1/\delta))$ should be satisfied for $\mathcal{E}_{\mathrm{NN}}$ to hold with high-probability, we require $m \geq M_3(1/\delta)$ where $M_3(1/\delta) = \max(\mathrm{poly}(n, 1/\delta, \ln n, \ln(1/\delta)), M_2(1/\delta))$. Redefining poly as the polynomial dominating $M_3(1/\delta)$, we obtain $m \geq \mathrm{poly}(n, 1/\delta, \ln n, \ln(1/\delta))$, which should be satisfied so that the bound on $\mathbb{E}\left[\left\| \widetilde{f}^{\mathrm{NN}}_{t_{\mathrm{op}}} - f^{\mathrm{NTK}}_{t_{\mathrm{op}}} \right\|^2_{L^2(\mathcal{X})}\right]$ holds.

Upper-bound on $\mathbb{E}\left[\left\|f_{t_{\text{op}}}^{\text{NTK}}-f^\star\right\|_{L^2(\mathcal{X})}^2\right]$. By the law of total expectation, we decompose the error over the disjoint events $\mathcal{E}_{\text{NTK}}$ and $\mathcal{E}_{\text{NTK}}^{\complement}$ as follows:

$$\begin{aligned}
\mathbb{E}\left[\left\|f_{t_{\text{op}}}^{\text{NTK}}-f^\star\right\|_{L^2(\mathcal{X})}^2\right] =& \mathbb{E}\left[\left\|f_{t_{\text{op}}}^{\text{NTK}}-f^\star\right\|_{L^2(\mathcal{X})}^2 \mathbb{1}_{\mathcal{E}_{\text{NTK}}}\right] + \mathbb{E}\left[\left\|f_{t_{\text{op}}}^{\text{NTK}}-f^\star\right\|_{L^2(\mathcal{X})}^2 \mathbb{1}_{\mathcal{E}_{\text{NTK}}^{\complement}}\right] \\
\leq & C_1^2\left(\ln\frac{6}{\delta}\right)^4 n^{-\frac{s(p+1)}{s(p+1)+p}} + \mathbb{E}\left[\left(\left\|f_{t_{\text{op}}}^{\text{NTK}}\right\|_{L^2(\mathcal{X})}+\|f^\star\|_\infty\right)^2 \mathbb{1}_{\mathcal{E}_{\text{NTK}}^{\complement}}\right] \\
\leq & C_1^2\left(\ln\frac{6}{\delta}\right)^4 n^{-\frac{s(p+1)}{s(p+1)+p}} + 2\mathbb{E}\left[\left(\left\|f_{t_{\text{op}}}^{\text{NTK}}\right\|_{L^2(\mathcal{X})}^2+\|f^\star\|_\infty^2\right) \mathbb{1}_{\mathcal{E}_{\text{NTK}}^{\complement}}\right] \\
\leq & C_1^2\left(\ln\frac{6}{\delta}\right)^4 n^{-\frac{s(p+1)}{s(p+1)+p}} + 2\left((\kappa^{\text{NTK}})^2B^2t_{\text{op}}^2+B^2\right)\Pr\left(\mathcal{E}_{\text{NTK}}^{\complement}\right) \\
\leq & C_1^2\left(\ln\frac{6}{\delta}\right)^4 n^{-\frac{s(p+1)}{s(p+1)+p}} + 2\left((\kappa^{\text{NTK}})^2B^2n^{\frac{2(p+1)}{s(p+1)+p}}+B^2\right)\delta, \qquad (66)
\end{aligned}$$

where, in the second line, we upper-bound $\left\|f_{t_{\text{op}}}^{\text{NTK}}-f^\star\right\|_{L^2(\mathcal{X})}^2$ by $C_1^2\left(\ln\frac{6}{\delta}\right)^4 n^{-\frac{s(p+1)}{s(p+1)+p}}$ on the event $\mathcal{E}_{\text{NTK}}$ and use the triangle inequality together with the identity $\|f^\star\|_{L^2(\mathcal{X})}\leq\|f^\star\|_\infty$; in the third line, we apply the identity $(a+b)^2\leq 2(a^2+b^2)$; in the fourth line, we use Lemmas E.1 and E.15 to upper-bound $\left\|f_{t_{\text{op}}}^{\text{NTK}}\right\|_{L^2(\mathcal{X})}$ by $\kappa^{\text{NTK}}Bt_{\text{op}}$ with $\kappa^{\text{NTK}}:=(p+2)(A^2+1)+1$ and $\|f^\star\|_\infty$ by B; and in the last line, we use $t_{\text{op}}=n^{\frac{p+1}{s(p+1)+p}}$.

Expected generalization bound. Plugging Equations (65) and (66) into Equation (63) yields:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\widetilde{f}_{t_{\text{op}}}^{\text{NN}}-f^\star\right\|_{L^2(\mathcal{X})}^2\right] \\
\leq & 2\left(C_2^2m^{-\frac{1}{12}}\ln m + C_1^2\left(\ln\frac{6}{\delta}\right)^4 n^{-\frac{s(p+1)}{s(p+1)+p}} + \left(6(\kappa^{\text{NTK}})^2B^2n^{\frac{2(p+1)}{s(p+1)+p}}+2B^2\right)\delta\right).
\end{aligned}$$

Since $n\geq\left\lceil\left(\frac{1}{3(\kappa^{\text{NTK}})^2}\right)^{\frac{s(p+1)+p}{2(p+1)}}\right\rceil$, we have $3(\kappa^{\text{NTK}})^2B^2n^{\frac{2(p+1)}{s(p+1)+p}}\geq B^2$, therefore:

$$\mathbb{E}\left[\left\|\widetilde{f}_{t_{\text{op}}}^{\text{NN}}-f^\star\right\|_{L^2(\mathcal{X})}^2\right] \leq 2\left(C_2^2m^{-\frac{1}{12}}\ln m+\varphi(\delta)\right), \qquad (67)$$

where $\varphi:(0,1)\to(0,+\infty)$ is the function defined for any $\delta\in(0,1)$ by:

$$\varphi(\delta)=C_1^2\left(\ln\frac{6}{\delta}\right)^4 n^{-\frac{s(p+1)}{s(p+1)+p}} + 12(\kappa^{\text{NTK}})^2B^2n^{\frac{2(p+1)}{s(p+1)+p}}\delta.$$

We minimize φ over $\delta\in(0,1)$. By Lemma I.1, since $n\geq\left\lceil\left(\frac{C_1^2(\ln 6)^3}{3(\kappa^{\text{NTK}})^2B^2}\right)^{\frac{s(p+1)+p}{(s+2)(p+1)}}\right\rceil+1$, the minimizer $\delta^\star\in(0,1)$ and the corresponding minimum value $\varphi(\delta^\star)$ are given by:

$$\delta^\star=\frac{9C_1^2}{(\kappa^{\text{NTK}})^2B^2}n^{-\frac{(s+2)(p+1)}{s(p+1)+p}}W_0^3\left(n^{\frac{1}{3}\frac{(s+2)(p+1)}{s(p+1)+p}}\left(\frac{2(\kappa^{\text{NTK}})^2B^2}{3C_1^2}\right)^{\frac{1}{3}}\right), \qquad (68)$$

and

$$\begin{aligned}
\varphi(\delta^\star) =& 27C_1^2W_0^3\left(n^{\frac{1}{3}\frac{(s+2)(p+1)}{s(p+1)+p}}\left(\frac{2(\kappa^{\text{NTK}})^2B^2}{3C_1^2}\right)^{\frac{1}{3}}\right) \\
&\times\left(3W_0\left(n^{\frac{1}{3}\frac{(s+2)(p+1)}{s(p+1)+p}}\left(\frac{2(\kappa^{\text{NTK}})^2B^2}{3C_1^2}\right)^{\frac{1}{3}}\right)+4\right)n^{-\frac{s(p+1)}{s(p+1)+p}},
\end{aligned}$$

where W_0 is the principal branch of the Lambert W function. Note that since $\frac{1}{p+1} < s \leq 2$, one can verify that $\frac{(s+2)(p+1)}{s(p+1)+p} \leq 4$, so that $\frac{1}{3}\frac{(s+2)(p+1)}{s(p+1)+p} \leq \frac{4}{3}$. Additionally, since:

$$n \geq \max\left(3, \left\lceil \left(\frac{3C_1^2 e^3}{2(\kappa^{\mathrm{NTK}})^2 B^2}\right)^{\frac{s(p+1)+p}{(s+2)(p+1)}} \right\rceil\right) \geq \max\left(e, \left(\frac{3C_1^2 e^3}{2(\kappa^{\mathrm{NTK}})^2 B^2}\right)^{\frac{s(p+1)+p}{(s+2)(p+1)}}\right),$$

applying Lemma I.3 yields:

$$W_0^3\left(n^{\frac{1}{3}\frac{(s+2)(p+1)}{s(p+1)+p}}\left(\frac{2(\kappa^{\mathrm{NTK}})^2 B^2}{3C_1^2}\right)^{\frac{1}{3}}\right)\left(3W_0\left(n^{\frac{1}{3}\frac{(s+2)(p+1)}{s(p+1)+p}}\left(\frac{2(\kappa^{\mathrm{NTK}})^2 B^2}{3C_1^2}\right)^{\frac{1}{3}}\right)+4\right)$$
$$\leq A_1\left(\left(\frac{2(\kappa^{\mathrm{NTK}})^2 B^2}{3C_1^2}\right)^{\frac{1}{3}}\right)(\ln n)^4,$$

where A_1 is defined for any $u > 0$ by:

$$A_1(u) = \frac{512}{27} + \frac{448}{9}|\ln u| + 48(\ln u)^2 + 20|\ln u|^3 + 3(\ln u)^4.$$

Substituting this bound into the expression of $\varphi(\delta^\star)$, we obtain:

$$\varphi(\delta^\star) \leq C'(\ln n)^4 n^{-\frac{s(p+1)}{s(p+1)+p}},$$

where $C' = 27C_1^2 A_1\left(\left(\frac{2(\kappa^{\mathrm{NTK}})^2 B^2}{3C_1^2}\right)^{\frac{1}{3}}\right)$. Plugging this bound in Equation (67) yields:

$$\mathbb{E}\left[\left\|\widehat{f}_{t_{\mathrm{op}}}^{\mathrm{NN}} - f^\star\right\|_{L^2(\mathcal{X})}^2\right] \leq 2\left(C_2^2 m^{-\frac{1}{12}}\ln m + C'(\ln n)^4 n^{-\frac{s(p+1)}{s(p+1)+p}}\right).$$

Choosing $m \geq \max\left(\mathrm{poly}(n, 1/\delta, \ln n, \ln(1/\delta)), n^{12}(\ln n)^{12}\right)$ gives $m^{-\frac{1}{12}}\ln m \leq 24n^{-1}$ by Lemma I.8. Since $s > 0$ and $p > 0$, we have $0 < \frac{s(p+1)}{s(p+1)+p} < 1$, so $n^{-1} \leq n^{-\frac{s(p+1)}{s(p+1)+p}}$. Moreover, $n \geq 3 > e$ implies $1 \leq (\ln n)^4$. Combining these observations leads to:

$$\mathbb{E}\left[\left\|\widehat{f}_{t_{\mathrm{op}}}^{\mathrm{NN}} - f^\star\right\|_{L^2(\mathcal{X})}^2\right] \leq 2\left(24C_2^2 + C'\right)(\ln n)^4 n^{-\frac{s(p+1)}{s(p+1)+p}}.$$

The desired bound follows by setting $C = 2\left(24C_2^2 + C'\right)$.

We now derive an explicit lower-bound on $\delta^\star$ to characterize the precise quantities on which m depends. Recall from Equation (68) that:

$$\delta^\star = \frac{9C_1^2}{(\kappa^{\mathrm{NTK}})^2 B^2} n^{-\frac{(s+2)(p+1)}{s(p+1)+p}} W_0^3(z), \qquad z := n^{\frac{1}{3}\frac{(s+2)(p+1)}{s(p+1)+p}}\left(\frac{2(\kappa^{\mathrm{NTK}})^2 B^2}{3C_1^2}\right)^{\frac{1}{3}}.$$

Since $n \geq \left\lceil \left(\frac{3C_1^2 e^3}{2(\kappa^{\mathrm{NTK}})^2 B^2}\right)^{\frac{s(p+1)+p}{(s+2)(p+1)}} \right\rceil$, we have $z \geq e$. As W_0 is increasing on $[0, +\infty)$ with $W_0(e) = 1$, this gives $W_0(z) \geq 1$, hence:

$$\delta^\star \geq \frac{9C_1^2}{(\kappa^{\mathrm{NTK}})^2 B^2} n^{-\frac{(s+2)(p+1)}{s(p+1)+p}}.$$

Using $\frac{(s+2)(p+1)}{s(p+1)+p} \leq 4$, we obtain:

$$\frac{1}{\delta^\star} \leq \frac{(\kappa^{\mathrm{NTK}})^2 B^2}{9C_1^2} n^{\frac{(s+2)(p+1)}{s(p+1)+p}} \leq \frac{(\kappa^{\mathrm{NTK}})^2 B^2}{9C_1^2} n^4,$$

and, taking logarithms and using $\ln n \geq 1$ for $n \geq 3$, we get:

$$\ln\frac{1}{\delta^\star} \leq \left|\ln\frac{(\kappa^{\mathrm{NTK}})^2 B^2}{9C_1^2}\right| + 4\ln n \leq \left(\left|\ln\frac{(\kappa^{\mathrm{NTK}})^2 B^2}{9C_1^2}\right| + 4\right)\ln n.$$

Both $1/\delta^\star$ and $\ln(1/\delta^\star)$ are therefore dominated by a polynomial in $(n, \ln n)$ whose coefficients depend only on s, p, A, B, M_μ and R. Redefining poly accordingly, the condition $m \geq \mathrm{poly}(n, 1/\delta^\star, \ln n, \ln(1/\delta^\star))$ is implied by $m \geq \mathrm{poly}(n, \ln n)$, which concludes the proof. □

F Differentiability of the Objectives and Properties of the Optimal Solutions

The next two Lemmas follow directly from Lemma D.3 and Lemma D.4 in [78], respectively. Here, differentiability is understood in the Fréchet sense. We write $\partial_h L_{out}(h) \in L^2(\mathcal{X})$ and $\partial_h L_{in}(\omega, h) \in L^2(\mathcal{X})$ for the partial derivatives with respect to h, $\partial_h^2 L_{in}(\omega, h) : L^2(\mathcal{X}) \to L^2(\mathcal{X})$ and $\partial_{\omega,h}^2 L_{in}(\omega, h) : L^2(\mathcal{X}) \to \mathbb{R}^d$ for the second-order partial derivatives of L_{in} with respect to h and to (ω, h).

Lemma F.1 (Differentiability of L_{out})**.** *Under Assumption B, L_{out} is differentiable on $L^2(\mathcal{X})$, with partial derivative given, for any $h \in L^2(\mathcal{X})$, by:*

$$\partial_h L_{out}(h) \in L^2(\mathcal{X}) : x \mapsto h(x) - \mathbb{E}_{\mathbb{P}}\left[y \mid x\right].$$

Lemma F.2 (Differentiability of $\partial_h L_{in}$)**.** *Under Assumption A, the function $(\omega, h) \mapsto \partial_h L_{in}(\omega, h)$ is differentiable on $U \times L^2(\mathcal{X})$, with partial derivatives given, for any $\omega \in U$ and $h \in L^2(\mathcal{X})$, by:*

$$\partial_h^2 L_{in}(\omega, h) = \mathrm{Id}_{L^2(\mathcal{X})}, \quad \partial_{\omega,h}^2 L_{in}(\omega, h) f = -\mathbb{E}_{\mathbb{P}}\left[\partial_\omega g_\omega(z) f(x)\right], \text{ for any } f \in L^2(\mathcal{X}).$$

Lemma F.3 (Uniform boundedness of target functions)**.** *Under Assumptions A and B, we have:*

- $h_\omega^\star \in L^\infty(\mathcal{X})$ *for any* $\omega \in U$, *and* $\sup_{\omega \in U} \|h_\omega^\star\|_\infty \leq D$;
- $b^\star \in L^\infty(\mathcal{X})$ *and* $\|b^\star\|_\infty \leq B$;
- $a_\omega^\star \in L^\infty(\mathcal{X})$ *for any* $\omega \in U$, *and* $\sup_{\omega \in U} \|a_\omega^\star\|_\infty \leq D + B$.

Proof. **Uniform boundedness of $h_\omega^\star$.** Let $\omega \in U$. By Assumption A, $|g_\omega(z)| \leq D$ for any $z \in \mathcal{Z}$, so that, by Jensen's inequality, for $\mathbb{P}_X$-almost every $x \in \mathcal{X}$, $|h_\omega^\star(x)| = |\mathbb{E}_{\mathbb{P}}\left[g_\omega(z) \mid x\right]| \leq \mathbb{E}_{\mathbb{P}}\left[|g_\omega(z)| \mid x\right] \leq D$. Taking the essential supremum over x and then the supremum over $\omega \in U$ gives $\sup_{\omega \in U} \|h_\omega^\star\|_\infty \leq D$.

Boundedness of $b^\star$. Similarly, by Assumption B, $|y| \leq B$, so that $|b^\star(x)| = |\mathbb{E}_{\mathbb{P}}[y \mid x]| \leq \mathbb{E}_{\mathbb{P}}[|y| \mid x] \leq B$ for $\mathbb{P}_X$-almost every $x \in \mathcal{X}$, hence $\|b^\star\|_\infty \leq B$.

Uniform boundedness of $a_\omega^\star$. Let $\omega \in U$. By the triangle inequality, $\|a_\omega^\star\|_\infty = \| - h_\omega^\star + b^\star\|_\infty \leq \|h_\omega^\star\|_\infty + \|b^\star\|_\infty$. Taking the supremum over $\omega \in U$ together with the two previous bounds gives $\sup_{\omega \in U} \|a_\omega^\star\|_\infty \leq D + B$. □

Proposition F.4 (Lipschitz continuity of $\omega \mapsto h_\omega^\star$)**.** *Under Assumption A, the function $\omega \mapsto h_\omega^\star$ is G-Lipschitz continuous on U.*

Proof. Let $\omega_1, \omega_2 \in U$. Since U is convex, the segment $\omega_s := \omega_2 + s(\omega_1 - \omega_2)$, $s \in [0, 1]$, lies in U, so that applying the Cauchy-Schwarz inequality and Assumption A gives for any $z \in \mathcal{Z}$:

$$|g_{\omega_1}(z) - g_{\omega_2}(z)| \leq \left| \int_0^1 \langle \partial_\omega g_{\omega_s}(z), \omega_1 - \omega_2 \rangle \, \mathrm{d}s \right| \leq G \|\omega_1 - \omega_2\|.$$

Hence, by Jensen's inequality, for $\mathbb{P}_X$-almost every $x \in \mathcal{X}$, we have:

$$\left|h_{\omega_1}^\star(x) - h_{\omega_2}^\star(x)\right| = |\mathbb{E}_{\mathbb{P}}\left[g_{\omega_1}(z) - g_{\omega_2}(z) \mid x\right]| \leq \mathbb{E}_{\mathbb{P}}\left[|g_{\omega_1}(z) - g_{\omega_2}(z)| \mid x\right] \leq G \|\omega_1 - \omega_2\|.$$

Since $\mathbb{P}_X$ is a probability measure, $\|\cdot\|_{L^2(\mathcal{X})} \leq \|\cdot\|_\infty$, and therefore:

$$\left\|h_{\omega_1}^\star - h_{\omega_2}^\star\right\|_{L^2(\mathcal{X})} \leq G \|\omega_1 - \omega_2\|,$$

which concludes the proof. □

G A Sobolev Space Characterization of the Infinite-Width NTK's RKHS

In this section, our goal is to show that the RKHS $\mathcal{H}^{\mathrm{NTK}}$ of the infinite-width NTK K^{NTK} of the two-layer ReLU network of the form in Equation (31), and its corresponding interpolation spaces, are Sobolev spaces when the input space $\mathcal{X}$ is a bounded Lipschitz domain (Definition G.1).

G.1 Preliminaries

For a positive semi-definite kernel k on a set E, we write $\mathcal{H}(k)$ for its RKHS. Standard facts about $\mathcal{H}(k)$ are used repeatedly: *(restriction)* for $E_0 \subset E$, $\mathcal{H}(k|_{E_0}) = \{f|_{E_0} : f \in \mathcal{H}(k)\}$ with $\|g\|_{\mathcal{H}(k|_{E_0})} = \min\{\|f\|_{\mathcal{H}(k)} : f|_{E_0} = g\}$; *(sums)* $\mathcal{H}(k_1 + k_2) = \mathcal{H}(k_1) + \mathcal{H}(k_2)$ with $\|f\|^2_{\mathcal{H}(k_1+k_2)} = \min\{\|f_1\|^2_{\mathcal{H}(k_1)} + \|f_2\|^2_{\mathcal{H}(k_2)} : f = f_1 + f_2\}$; and *(scaling)* for $c > 0$, $\mathcal{H}(ck) = \mathcal{H}(k)$ as sets, with $\|f\|_{\mathcal{H}(ck)} = c^{-1/2}\|f\|_{\mathcal{H}(k)}$ [8]. For a subset F of a normed space $\mathcal{H}$, we say F *embeds continuously* into a normed space $\mathcal{H}'$ if $F \subset \mathcal{H}'$ and there is a constant $C > 0$ such that $\|f\|_{\mathcal{H}'} \leq C\|f\|_{\mathcal{H}}$ for every $f \in F$; we write this $F \hookrightarrow \mathcal{H}'$. For a function f on a topological space E, the *support* of f, written $\operatorname{supp} f$, is the closure of $\{x \in E : f(x) \neq 0\}$. f is said to be *compactly supported* in an open set $U \subset E$ if $\operatorname{supp} f$ is a compact subset of U. Let $\phi : \mathbb{R}^p \to \mathbb{S}^p \subset \mathbb{R}^{p+1}$ be the map defined for any $x \in \mathbb{R}^p$ by $\phi(x) = \tilde{x}/\|\tilde{x}\|$ with $\tilde{x} = [x^\top, 1]^\top$. We write $\mathbb{S}^p_+ := \{y \in \mathbb{S}^p : y_{p+1} > 0\}$ and $\mathbb{S}^p_- := \{y \in \mathbb{S}^p : y_{p+1} < 0\}$ for the open upper and lower hemispheres of $\mathbb{S}^p$. For any $u \in [-1, 1]$, we define $\kappa(u) := u\kappa_0(u) + \kappa_1(u)$ where $\kappa_0(u) := \frac{1}{\pi}(\pi - \arccos(u))$ and $\kappa_1(u) := \frac{1}{\pi}\big(u(\pi - \arccos(u)) + \sqrt{1-u^2}\big)$. Using Equation (38):

$$K^{\mathrm{NTK}}(x, x') = \frac{1}{2}\|\tilde{x}\|\|\tilde{x}'\|\kappa\left(\langle\phi(x), \phi(x')\rangle\right) + 1,$$

Setting $K_0(y, y') := \kappa(\langle y, y'\rangle)$ for $y, y' \in \mathbb{S}^p$, and using the notation $M_g[k](x, x') := g(x)k(x, x')g(x')$ for a function g and kernel k and $\phi^* K_0(x, x') := K_0(\phi(x), \phi(x'))$ for $x, x' \in \mathcal{X}$:

$$K^{\mathrm{NTK}} = \frac{1}{2}M_q[\phi^* K_0] + \mathcal{I}_{\mathcal{X}}, \quad \text{where } q(x) := \|\tilde{x}\| \text{ and } \mathcal{I}_{\mathcal{X}}(x, x') = 1 \text{ for } x, x' \in \mathcal{X}. \tag{69}$$

In this section, $\mathcal{X}$ is a *bounded Lipschitz domain*, in the sense of Definition G.1 below. We write $L^2(\mathcal{X}, \mathrm{d}x)$ for the space of square-integrable functions on $\mathcal{X}$ with respect to the Lebesgue measure $\mathrm{d}x$, and recall $L^2(\mathcal{X}) := L^2(\mathcal{X}, \mu)$. Let $L^1_{\mathrm{loc}}(\mathcal{X}, \mathrm{d}x)$ denote the space of *locally integrable* functions on $\mathcal{X}$, *i.e.*, functions $f : \mathcal{X} \to \mathbb{R}$ such that $\int_K |f(x)|\mathrm{d}x < \infty$ for every compact $K \subset \mathcal{X}$, and let $C^\infty_c(\mathcal{X})$ denote the space of *smooth, compactly supported* functions on $\mathcal{X}$. Write $\mathbb{N}_0 := \{0, 1, 2, \dots\}$ for the nonnegative integers.

Definition G.1 (Bounded Lipschitz domain [44, Definition 1.2.1.1])**.** $\mathcal{X}$ is *bounded Lipschitz* if it is open, bounded, and nonempty, and its boundary $\partial\mathcal{X}$ is locally the graph of a Lipschitz function: for every $x_0 \in \partial\mathcal{X}$, there exist $r > 0$, an orthonormal coordinate system $(z_1, \dots, z_{p-1}, z_p)$ centered at x_0, and a Lipschitz function $\gamma : \mathbb{R}^{p-1} \to \mathbb{R}$ such that:

$$\begin{aligned}\mathcal{X} \cap B(x_0, r) =& \{(z_1, \dots, z_p) \in B(x_0, r) : z_p > \gamma(z_1, \dots, z_{p-1})\},\\ \text{and } \partial\mathcal{X} \cap B(x_0, r) =& \{(z_1, \dots, z_p) \in B(x_0, r) : z_p = \gamma(z_1, \dots, z_{p-1})\},\end{aligned}$$

where $B(x_0, r)$ is the open ball centered at x_0 of radius r.

Definition G.2 (Weak derivative)**.** For $f \in L^1_{\mathrm{loc}}(\mathcal{X}, \mathrm{d}x)$ and a multi-index $\alpha = (\alpha_1, \dots, \alpha_p) \in \mathbb{N}^p_0$ with $|\alpha| = \alpha_1 + \ldots + \alpha_p$, $v \in L^1_{\mathrm{loc}}(\mathcal{X}, \mathrm{d}x)$ is the *weak α-th derivative* of f, written $v = D^\alpha f$, if:

$$\int_{\mathcal{X}} f(x) D^\alpha \phi(x)\mathrm{d}x = (-1)^{|\alpha|}\int_{\mathcal{X}} v(x)\phi(x)\mathrm{d}x \qquad \text{for all } \phi \in C^\infty_c(\mathcal{X}).$$

The following are the definitions of Sobolev spaces used throughout this section, following the standard treatment in, *e.g.*, [3, Section 5.1].

Definition G.3 (Sobolev space, integer order)**.** For $m \in \mathbb{N}_0$, the Sobolev space of integer order m is:

$$H^m(\mathcal{X}) := \left\{f \in L^2(\mathcal{X}, \mathrm{d}x) : D^\alpha f \in L^2(\mathcal{X}, \mathrm{d}x) \text{ for every } \alpha \text{ with } |\alpha| \leq m\right\},$$

with $\|f\|^2_{H^m(\mathcal{X})} := \sum_{|\alpha|\leq m}\|D^\alpha f\|^2_{L^2(\mathcal{X}, \mathrm{d}x)}$.

Definition G.4 (Sobolev space, fractional order)**.** For $\tau = m + \vartheta$ with $m \in \mathbb{N}_0$ and $\vartheta \in (0, 1)$, define the *Gagliardo (Sobolev–Slobodeckij) seminorm* as:

$$[f]^2_{\tau,\mathcal{X}} := \sum_{|\alpha|=m}\iint_{\mathcal{X}\times\mathcal{X}} \frac{|D^\alpha f(x) - D^\alpha f(y)|^2}{\|x - y\|^{p+2\vartheta}}\mathrm{d}x\mathrm{d}y.$$

The Sobolev space of fractional order τ is $H^\tau(\mathcal{X}) := \left\{f \in H^m(\mathcal{X}) : [f]_{\tau,\mathcal{X}} < \infty\right\}$, with $\|f\|^2_{H^\tau(\mathcal{X})} := \|f\|^2_{H^m(\mathcal{X})} + [f]^2_{\tau,\mathcal{X}}$. For $\tau \in \mathbb{N}_0$, *i.e.*, $\vartheta = 0$, $H^\tau(\mathcal{X})$ is understood as in Definition G.3.

Corollary G.5 (Domination absorbs a summand)**.** *Let k_1, k_2 be positive semi-definite kernels on the same set. If $\mathcal{H}(k_1) \hookrightarrow \mathcal{H}(k_2)$ continuously, then $\mathcal{H}(k_1 + k_2) \cong \mathcal{H}(k_2)$.*

Proof. By the sum property of RKHSs, $\mathcal{H}(k_1 + k_2) = \mathcal{H}(k_1) + \mathcal{H}(k_2)$. Since $\mathcal{H}(k_1) \subset \mathcal{H}(k_2)$, this equals $\mathcal{H}(k_2)$ as a set. Let C be the embedding constant, then $\|f_1\|_{\mathcal{H}(k_2)} \leq C\|f_1\|_{\mathcal{H}(k_1)}$. For $f = f_1 + f_2 \in \mathcal{H}(k_1) + \mathcal{H}(k_2)$, we have:

$$\|f\|_{\mathcal{H}(k_2)} \leq \|f_1\|_{\mathcal{H}(k_2)} + \|f_2\|_{\mathcal{H}(k_2)} \leq C\|f_1\|_{\mathcal{H}(k_1)} + \|f_2\|_{\mathcal{H}(k_2)} \leq (C+1)\sqrt{\|f_1\|^2_{\mathcal{H}(k_1)} + \|f_2\|^2_{\mathcal{H}(k_2)}}.$$

Taking the infimum over decompositions gives $\|f\|_{\mathcal{H}(k_2)} \leq (C+1)\|f\|_{\mathcal{H}(k_1+k_2)}$, and the trivial decomposition $f_1 = 0$ gives the reverse. □

Lemma G.6. *Let k be a positive semi-definite kernel on a set E_2, let $\varphi_0 : E_1 \to E_2$ be a surjection, and let $g : E_1 \to \mathbb{R}$ be nowhere zero. Then the map $f \mapsto g \cdot (f \circ \varphi_0)$ is an isometric isomorphism from $\mathcal{H}(k)$ onto $\mathcal{H}\big(M_g[\varphi_0^* k]\big)$.*

Proof. If $\Psi : E_2 \to W$ is a feature map for k, *i.e.*, $k(y, y') = \langle \Psi(y), \Psi(y') \rangle_W$, then $\Psi'(x) := g(x)\Psi(\varphi_0(x))$ is a feature map for $M_g[\varphi_0^* k]$, since $\langle \Psi'(x), \Psi'(x') \rangle_W = g(x)g(x')k(\varphi_0(x), \varphi_0(x')) = M_g[\varphi_0^* k](x, x')$. Write $\Phi(w) := \langle w, \Psi(\cdot) \rangle_W$ and $\Phi'(w) := \langle w, \Psi'(\cdot) \rangle_W$ for the corresponding representation maps, so that $\mathcal{H}(k) = \{\Phi(w) : w \in W\}$, $\mathcal{H}(M_g[\varphi_0^* k]) = \{\Phi'(w) : w \in W\}$, and, for $f \in \mathcal{H}(k)$, $\|f\|_{\mathcal{H}(k)} = \min\{\|w\|_W : \Phi(w) = f\}$, with the analogous expression for $\mathcal{H}(M_g[\varphi_0^* k])$ via Φ'. Since $\Phi'(w)(x) = \langle w, \Psi'(x) \rangle_W = g(x)\langle w, \Psi(\varphi_0(x)) \rangle_W = g(x)\Phi(w)(\varphi_0(x))$ for every $w \in W$, we have $\Phi' = T \circ \Phi$, where $T(f) := g \cdot (f \circ \varphi_0)$. In particular, any w_0 representing f on the left (*i.e.*, $\Phi(w_0) = f$) automatically represents $T(f)$ on the right (*i.e.*, $\Phi'(w_0) = T(f)$). We claim that, for such w_0, the set of *all* $w \in W$ representing f on the left coincides with the set of all $w \in W$ representing $T(f)$ on the right. Since Φ and Φ' are linear, the former set is the coset $w_0 + N$ and the latter is $w_0 + N'$, where $N := \ker \Phi$ and $N' := \ker \Phi'$; so it suffices to show $N = N'$. Indeed, $w \in N$ means $\langle w, \Psi(y) \rangle_W = 0$ for every $y \in E_2$, which, since φ_0 is surjective, holds if and only if $\langle w, \Psi(\varphi_0(x)) \rangle_W = 0$ for every $x \in E_1$; and since g is nowhere zero, this holds if and only if $g(x)\langle w, \Psi(\varphi_0(x)) \rangle_W = \langle w, \Psi'(x) \rangle_W = 0$ for every $x \in E_1$, *i.e.*, $w \in N'$. Hence $N = N'$, and the two cosets $w_0 + N$, $w_0 + N'$ are identical as subsets of W. Consequently, for every $f \in \mathcal{H}(k)$, the set of $w \in W$ representing f on the left, namely $w_0 + N$, is exactly the set of $w \in W$ representing $T(f)$ on the right, namely $w_0 + N' = w_0 + N$. Minimizing $\|w\|_W$ over this common set gives:

$$\|f\|_{\mathcal{H}(k)} = \min_{w \in w_0 + N} \|w\|_W = \min_{w \in w_0 + N'} \|w\|_W = \|T(f)\|_{\mathcal{H}(M_g[\varphi_0^* k])} = \|g \cdot (f \circ \varphi_0)\|_{\mathcal{H}(M_g[\varphi_0^* k])}.$$

In addition, T is a bijection from $\mathcal{H}(k)$ onto $\mathcal{H}(M_g[\varphi_0^* k])$. Indeed, if $T(f_1) = T(f_2)$, then $g(x)f_1(\varphi_0(x)) = g(x)f_2(\varphi_0(x))$ for all $x \in E_1$. Since g is nowhere zero, $f_1(\varphi_0(x)) = f_2(\varphi_0(x))$ and since φ_0 is surjective, $f_1 = f_2$. Thus T is injective. Surjectivity holds because every $h \in \mathcal{H}(M_g[\varphi_0^* k])$ equals $\Phi'(w)$ for some w, hence $h = T(\Phi(w))$. Together with the norm identity above, T is an isometric isomorphism. □

Lemma G.7. *The map ϕ is a real-analytic diffeomorphism of $\mathbb{R}^p$ onto the open upper hemisphere $\mathbb{S}^p_+ := \{y \in \mathbb{S}^p : y_{p+1} > 0\}$, with inverse $\phi^{-1}(y) = (y_1, \ldots, y_p)/y_{p+1}$. Moreover, under the assumption that there exists $A > 0$ such that $\|x\| \leq A$ for $x \in \mathcal{X}$, it holds that:*

$$\phi(x)_{p+1} = \big(1 + \|x\|^2\big)^{-1/2} \;\geq\; \big(1 + A^2\big)^{-1/2} \;>\; 0 \qquad \textit{for all } x \in \mathcal{X}, \tag{70}$$

so $\widehat{\mathcal{X}} := \phi(\mathcal{X})$ has closure contained in the compact cap $\{y \in \mathbb{S}^p : y_{p+1} \geq (1 + A^2)^{-1/2}\} \subset \mathbb{S}^p_+$. Consequently, there exists $\chi \in C^\infty(\mathbb{S}^p)$ with $0 \leq \chi \leq 1$, $\chi \equiv 1$ on $\widehat{\mathcal{X}}$, and $\operatorname{supp} \chi \subset \mathbb{S}^p_+$.

Proof. The map ϕ is real-analytic on $\mathbb{R}^p$, since its coordinate functions are compositions of polynomials with the real-analytic function $t \mapsto t^{-1/2}$ on $(0, \infty)$. Likewise, ϕ^{-1} is real-analytic on $\mathbb{S}^p_+$, since $y_{p+1} > 0$ there, and hence $t \mapsto t^{-1}$ is real-analytic on $(0, \infty)$. A direct calculation shows that $\phi^{-1} \circ \phi = \mathrm{Id}_{\mathbb{R}^p}$ and $\phi \circ \phi^{-1} = \mathrm{Id}_{\mathbb{S}^p_+}$, so ϕ is a real-analytic diffeomorphism. Equation (70) follows from $\|x\| \leq A$ since $t \mapsto (1 + t)^{-1/2}$ is decreasing. Let $C_A := \{y \in \mathbb{S}^p : y_{p+1} \geq (1 + A^2)^{-1/2}\}$. By Equation (70), $\widehat{\mathcal{X}} \subset C_A$, and since C_A is closed (the preimage of a closed half-line under the

continuous map $y \mapsto y_{p+1}$), $\overline{\widehat{\mathcal{X}}} \subset C_A$ as well. In particular, $\overline{\widehat{\mathcal{X}}}$ is compact being a closed subset of the compact C_A. Moreover, $C_A \subset \mathbb{S}^p_+$ since $(1+A^2)^{-1/2} > 0$. A smooth bump function for the pair $\big(\overline{\widehat{\mathcal{X}}}, \mathbb{S}^p_+\big)$, which exists since $\overline{\widehat{\mathcal{X}}}$ is compact and contained in the open set $\mathbb{S}^p_+$, gives χ [60, Proposition 2.26]. □

Corollary G.8. *Define $\widehat{\mathcal{X}} \coloneqq \phi(\mathcal{X})$. It holds that:*

$$\mathcal{H}^{\mathrm{NTK}} = q \cdot \Big[\mathcal{H}\big(K_0|_{\widehat{\mathcal{X}}}\big) \circ \phi\Big] + \mathbb{R} \cdot \mathbb{1}_{\mathcal{X}}, \tag{71}$$

where $\mathbb{1}_{\mathcal{X}} : \mathcal{X} \to \mathbb{R}$ denotes the constant-one function, i.e., $\mathbb{1}_{\mathcal{X}}(x) = 1$ for all $x \in \mathcal{X}$.

Proof. Since $\widehat{\mathcal{X}} \coloneqq \phi(\mathcal{X})$, the restriction $\phi|_{\mathcal{X}}$ is a surjection onto $\widehat{\mathcal{X}}$, and $q \geq 1 > 0$ on $\mathcal{X}$. Lemma G.6 therefore applies to the kernel $M_q[\phi^* K_0]$ appearing in Equation (69), giving $\mathcal{H}\big(M_q[\phi^* K_0]\big) = q \cdot [\mathcal{H}(K_0|_{\widehat{\mathcal{X}}}) \circ \phi]$ isometrically. The factor $\frac{1}{2}$ rescales the norm only, and the sum property of $\mathcal{H}(\cdot)$ recalled above assembles the two terms of Equation (69) into Equation (71). □

G.2 The sphere kernel: parity gap and the hemisphere lemma

Denote by $L^2(\mathbb{S}^p)$ the space of square-integrable functions on $\mathbb{S}^p$ with respect to its surface measure. Let $\{Y_{k,j}\}_{k\geq 0,\, 1\leq j\leq a_k}$ denote an orthonormal basis of $L^2(\mathbb{S}^p)$, with $Y_{k,j}$ a spherical harmonic of degree k and a_k the dimension of the degree-k harmonics $\mathcal{Y}_k$. Every $f \in L^2(\mathbb{S}^p)$ decomposes uniquely as $f = \sum_{k\geq 0} f_k$, $f_k \coloneqq \sum_{j=1}^{a_k} \langle f, Y_{k,j}\rangle Y_{k,j} \in \mathcal{Y}_k$. For $\tau \geq 0$, the Sobolev space $H^\tau(\mathbb{S}^p)$ is defined spectrally by $\|f\|^2_{H^\tau(\mathbb{S}^p)} \coloneqq \sum_{k\geq 0}(1 + k(k+p-1))^\tau \|f_k\|^2_{L^2(\mathbb{S}^p)} \asymp \sum_{k\geq 0}(1+k)^{2\tau}\|f_k\|^2_{L^2(\mathbb{S}^p)}$. Given a dot-product kernel $\kappa(\langle y, y'\rangle)$ on $\mathbb{S}^p$ with Mercer decomposition $\kappa(\langle y, y'\rangle) = \sum_{k\geq 0} \mu_k \sum_{j=1}^{a_k} Y_{k,j}(y) Y_{k,j}(y')$, its RKHS is:

$$\mathcal{H}(\kappa) \coloneqq \Big\{ f = \sum_{k\geq 0}\sum_{j=1}^{a_k} c_{k,j} Y_{k,j} \;:\; \|f\|^2_{\mathcal{H}(\kappa)} \coloneqq \sum_{k\geq 0}\sum_{j=1}^{a_k} \frac{c_{k,j}^2}{\mu_k} < \infty \Big\},$$

with the convention, forced by finiteness, that $c_{k,j} = 0$ whenever $\mu_k = 0$. Moreover, for an open subset $\mathcal{V} \subset \mathbb{S}^p$, the Sobolev space $H^\tau(\mathcal{V})$ is defined by restriction $H^\tau(\mathcal{V}) \coloneqq \big\{U|_{\mathcal{V}} \;:\; U \in H^\tau(\mathbb{S}^p)\big\}$ with norm $\|g\|_{H^\tau(\mathcal{V})} \coloneqq \inf\big\{\|U\|_{H^\tau(\mathbb{S}^p)} \;:\; U \in H^\tau(\mathbb{S}^p),\; U|_{\mathcal{V}} = g\big\}$. For more details, we refer the reader to [10, 12, 49].

Corollary G.9 (Parity gap on $\mathbb{S}^p$)**.** *Recall $K_0(y, y') \coloneqq \kappa(\langle y, y'\rangle)$ for $y, y' \in \mathbb{S}^p$, with Mercer eigenvalues $(\mu_k)_{k\geq 0}$. Then, we have:*

$$\mu_0, \mu_1 > 0, \qquad \mu_k = 0 \text{ for every odd } k \geq 3, \qquad \mu_k \sim C(p+1)\, k^{-(p+1)} \text{ otherwise as } k \to \infty,$$

for a constant $C(p+1) > 0$. Consequently, writing $\mathcal{E} \coloneqq \overline{\bigoplus_{k \text{ even}} \mathcal{Y}_k} \oplus \mathcal{Y}_1 \subset L^2(\mathbb{S}^p)$, it holds that:

$$\mathcal{H}(K_0) \cong \mathcal{E} \cap H^{\frac{p+1}{2}}(\mathbb{S}^p) \;\subsetneq\; H^{\frac{p+1}{2}}(\mathbb{S}^p).$$

Proof. The first part of this corollary comes from [18, Proposition 5].

Identifying $\mathcal{E}$. Since $L^2(\mathbb{S}^p) = \bigoplus_{k\geq 0} \mathcal{Y}_k$ is an orthogonal decomposition, the closed subspace $\mathcal{E}$ consists exactly of those $f = \sum_{k\geq 0} f_k \in L^2(\mathbb{S}^p)$ with $f_k = 0$ for every odd $k \geq 3$. We note that the spherical-harmonic expansion already converges in $L^2(\mathbb{S}^p)$, since we have by Parseval's identity that $\|f\|^2_{L^2(\mathbb{S}^p)} = \sum_k \|f_k\|^2_{L^2(\mathbb{S}^p)} < \infty$ for any $f \in L^2(\mathbb{S}^p)$.

A two-sided bound on μ_k^{-1}. Inverting $\mu_k \sim C(p+1)k^{-(p+1)}$ gives $\mu_k^{-1} \sim C(p+1)^{-1}k^{p+1}$ as $k \to \infty$. By the definition of asymptotic equivalence, there exist constants $c_0, C_0 > 0$ depending only on p and $K \geq 1$ such that $c_0 k^{p+1} \leq \mu_k^{-1} \leq C_0 k^{p+1}$, for $k \geq K$. It remains to control the finitely many indices $k < K$. Since $\mu_k > 0$ for every $k \in \{1\} \cup 2\mathbb{N}_0$ with $2\mathbb{N}_0 = \{0, 2, 4, 6, \ldots\}$, the quantities $\frac{\mu_k^{-1}}{(1+k)^{p+1}}$, with $k \in \{1\} \cup 2\mathbb{N}_0$ and $k < K$, form a finite collection of strictly positive finite numbers. Thus, their minimum is positive and their maximum is finite. Combining these bounds

with the preceding estimate for $k \geq K$, and using $k^{p+1} \asymp (1+k)^{p+1}$ for $k \geq 1$, yields constants $c, C > 0$, depending only on p, such that $c(1+k)^{p+1} \leq \mu_k^{-1} \leq C(1+k)^{p+1}$ for $k \in \{1\} \cup 2\mathbb{N}_0$.

(⊆). Let $f \in \mathcal{H}(K_0)$, so $f = \sum_{k,j} c_{k,j} Y_{k,j}$ with $\sum_{k,j} c_{k,j}^2/\mu_k < \infty$ and $c_{k,j} = 0$ forced wherever $\mu_k = 0$, *i.e.*, at every odd $k \geq 3$. By the identification above this already gives $f \in \mathcal{E}$. Summing the norm condition over j (Parseval within each degree, $\sum_j c_{k,j}^2 = \|f_k\|_{L^2(\mathbb{S}^p)}^2$) and using the lower-bound on μ_k^{-1}, we get:

$$\begin{aligned}\|f\|_{H^{\frac{p+1}{2}}(\mathbb{S}^p)}^2 =& \sum_{k\geq 0}(1+k)^{p+1}\|f_k\|_{L^2(\mathbb{S}^p)}^2 = \sum_{k\in\{1\}\cup 2\mathbb{N}_0}(1+k)^{p+1}\|f_k\|_{L^2(\mathbb{S}^p)}^2 \\ \leq& \frac{1}{c}\sum_{k\in\{1\}\cup 2\mathbb{N}_0}\mu_k^{-1}\|f_k\|_{L^2(\mathbb{S}^p)}^2 = \frac{1}{c}\|f\|_{\mathcal{H}(K_0)}^2 < \infty,\end{aligned}$$

where the first equality holds because $f_k = 0$ for $k \notin \{1\} \cup 2\mathbb{N}_0$; so $f \in H^{\frac{p+1}{2}}(\mathbb{S}^p)$. Hence $f \in \mathcal{E} \cap H^{\frac{p+1}{2}}(\mathbb{S}^p)$.

(⊇). Let $f \in \mathcal{E} \cap H^{\frac{p+1}{2}}(\mathbb{S}^p)$. Then, $f_k = 0$ for every odd $k \geq 3$, which is exactly the forcing condition defining $\mathcal{H}(K_0)$ above, so f is an admissible candidate for $\mathcal{H}(K_0)$. Since $f \in H^{\frac{p+1}{2}}(\mathbb{S}^p)$, $\sum_{k\geq 0}(1+k)^{p+1}\|f_k\|_{L^2(\mathbb{S}^p)}^2 < \infty$. As $f \in \mathcal{E}$, this sum is already supported on $\{1\} \cup 2\mathbb{N}_0$, so the upper-bound on μ_k^{-1} gives:

$$\|f\|_{\mathcal{H}(K_0)}^2 = \sum_{k\in\{1\}\cup 2\mathbb{N}_0}\mu_k^{-1}\|f_k\|_{L^2(\mathbb{S}^p)}^2 \leq C\sum_{k\in\{1\}\cup 2\mathbb{N}_0}(1+k)^{p+1}\|f_k\|_{L^2(\mathbb{S}^p)}^2 = C\|f\|_{H^{\frac{p+1}{2}}(\mathbb{S}^p)}^2 < \infty,$$

where we use the fact that $f \in H^{\frac{p+1}{2}}(\mathbb{S}^p)$. Hence, $f \in \mathcal{H}(K_0)$.

The two inclusions give $\mathcal{H}(K_0) = \mathcal{E} \cap H^{\frac{p+1}{2}}(\mathbb{S}^p)$ as *sets*, and the same constants c, C, one used in each direction, give the equivalence of norms. Together, this gives $\mathcal{H}(K_0) \cong \mathcal{E} \cap H^{\frac{p+1}{2}}(\mathbb{S}^p)$.

Properness. A nonzero $Y \in \mathcal{Y}_3$ is a polynomial, hence lies in every $H^\tau(\mathbb{S}^p)$ for $\tau \geq 0$. Its only nonzero component is at degree 3, which is odd and ≥ 3, so $Y \notin \mathcal{E}$ by the identification above, hence $Y \notin \mathcal{H}(K_0)$. So the containment $\mathcal{H}(K_0) \subsetneq H^{\frac{p+1}{2}}(\mathbb{S}^p)$ is strict. □

For $y \in \mathbb{S}^p$, write $\mathfrak{a}(y) := -y$ for the antipodal map. We first record how composition with $\mathfrak{a}$ acts on the spherical-harmonic expansion of a function.

Lemma G.10. *For any $h \in L^2(\mathbb{S}^p)$ with expansion $h = \sum_{k\geq 0} h_k$, $h_k \in \mathcal{Y}_k$, the function $h \circ \mathfrak{a}$ has expansion $(h \circ \mathfrak{a})_k = (-1)^k h_k$. Consequently, $\|h \circ \mathfrak{a}\|_{H^\tau(\mathbb{S}^p)} = \|h\|_{H^\tau(\mathbb{S}^p)}$ for every $\tau \geq 0$.*

Proof. Every $Y \in \mathcal{Y}_k$ is the restriction to $\mathbb{S}^p$ of a polynomial on $\mathbb{R}^{p+1}$ that is homogeneous of degree k, *i.e.*, $Y(\lambda x) = \lambda^k Y(x)$ for every scalar λ. Taking $\lambda = -1$ gives $Y(-y) = (-1)^k Y(y)$ for every $y \in \mathbb{S}^p$. Applying this to each term of $h = \sum_{k\geq 0} h_k$ yields:

$$h(\mathfrak{a}(y)) = h(-y) = \sum_{k\geq 0} h_k(-y) = \sum_{k\geq 0}(-1)^k h_k(y),$$

so $h \circ \mathfrak{a} = \sum_{k\geq 0}(-1)^k h_k$. Since $(-1)^k h_k \in \mathcal{Y}_k$, this is precisely the spherical-harmonic expansion of $h \circ \mathfrak{a}$, giving $(h \circ \mathfrak{a})_k = (-1)^k h_k$. Since $\|(-1)^k h_k\|_{L^2(\mathbb{S}^p)} = \|h_k\|_{L^2(\mathbb{S}^p)}$ for every k, we obtain:

$$\begin{aligned}\|h \circ \mathfrak{a}\|_{H^\tau(\mathbb{S}^p)}^2 =& \sum_{k\geq 0}\big(1 + k(k+p-1)\big)^\tau \|(h \circ \mathfrak{a})_k\|_{L^2(\mathbb{S}^p)}^2 \\ =& \sum_{k\geq 0}\big(1 + k(k+p-1)\big)^\tau \|h_k\|_{L^2(\mathbb{S}^p)}^2 = \|h\|_{H^\tau(\mathbb{S}^p)}^2.\end{aligned}$$ □

Lemma G.11 (Hemisphere lemma)**.** *Let $\mathcal{V} \subset \mathbb{S}^p$ be open, bounded, with Lipschitz boundary, with $\overline{\mathcal{V}}$ compactly contained in $\mathbb{S}_+^p$. Then, it holds that $\mathcal{H}\big(K_0|_{\mathcal{V}}\big) \cong H^{\frac{p+1}{2}}(\mathcal{V})$.*

Proof. *($\subseteq$).* Let $g \in \mathcal{H}(K_0|_{\mathcal{V}})$. By the restriction property of RKHSs, there exists $F \in \mathcal{H}(K_0)$ with $F|_{\mathcal{V}} = g$ and $\|F\|_{\mathcal{H}(K_0)} = \|g\|_{\mathcal{H}(K_0|_{\mathcal{V}})}$. By Corollary G.9, it holds that $\mathcal{H}(K_0) \subsetneq H^{\frac{p+1}{2}}(\mathbb{S}^p)$ and there is a constant $C_1 > 0$, depending only on p, with $\|F\|_{H^{\frac{p+1}{2}}(\mathbb{S}^p)} \leq C_1 \|F\|_{\mathcal{H}(K_0)}$. Since F is in particular one extension of g to $\mathbb{S}^p$, the definition of the restriction norm on $\mathcal{V}$ as an infimum over all such extensions gives $\|g\|_{H^{\frac{p+1}{2}}(\mathcal{V})} \leq \|F\|_{H^{\frac{p+1}{2}}(\mathbb{S}^p)}$. Chaining the three displays, we get:

$$\|g\|_{H^{\frac{p+1}{2}}(\mathcal{V})} \leq \|F\|_{H^{\frac{p+1}{2}}(\mathbb{S}^p)} \leq C_1 \|F\|_{\mathcal{H}(K_0)} = C_1 \|g\|_{\mathcal{H}(K_0|_{\mathcal{V}})} < \infty,$$

so $g \in H^{\frac{p+1}{2}}(\mathcal{V})$ with $\|g\|_{H^{\frac{p+1}{2}}(\mathcal{V})} \leq C_1 \|g\|_{\mathcal{H}(K_0|_{\mathcal{V}})}$.

($\supseteq$). Let $g \in H^{\frac{p+1}{2}}(\mathcal{V})$. Our goal is to produce a function on the whole sphere, in $\mathcal{H}(K_0)$, that restricts to g on $\mathcal{V}$.

Step 1: extend g to the sphere, then confine the extension to the upper hemisphere. Since $\mathcal{V}$ is bounded Lipschitz, there is a bounded linear extension operator [85, Chapter VI, Theorem 5], giving $\widetilde{G} \in H^{\frac{p+1}{2}}(\mathbb{S}^p)$ with $\widetilde{G}|_{\mathcal{V}} = g$ and $\|\widetilde{G}\|_{H^{\frac{p+1}{2}}(\mathbb{S}^p)} \leq C_2 \|g\|_{H^{\frac{p+1}{2}}(\mathcal{V})}$ for a constant $C_2 > 0$ depending only on p and the Lipschitz character of $\mathcal{V}$. Let $\chi \in C^\infty(\mathbb{S}^p)$ be the smooth bump function from Lemma G.7, equal to 1 on $\mathcal{V}$ and supported in $\mathbb{S}^p_+$, and set $G := \chi \widetilde{G}$. By the multiplier property of Sobolev spaces [88, §4.2], there is a constant $C_3 > 0$, depending only on χ, such that:

$$\|G\|_{H^{\frac{p+1}{2}}(\mathbb{S}^p)} \leq C_3 \|\widetilde{G}\|_{H^{\frac{p+1}{2}}(\mathbb{S}^p)} \leq C_2 C_3 \|g\|_{H^{\frac{p+1}{2}}(\mathcal{V})}.$$

Moreover, $\operatorname{supp} G \subset \operatorname{supp} \chi \subset \mathbb{S}^p_+$, and since $\chi \equiv 1$ on $\mathcal{V}$, we have $G|_{\mathcal{V}} = \chi|_{\mathcal{V}} \cdot \widetilde{G}|_{\mathcal{V}} = \widetilde{G}|_{\mathcal{V}} = g$.

Step 2: symmetrize across the equator. Define $F := G + G \circ \mathfrak{a}$. We verify three properties of F.

(i) F restricts to g on $\mathcal{V}$. For $y \in \mathcal{V} \subset \mathbb{S}^p_+$, the antipodal point $-y$ lies in the open lower hemisphere $\mathbb{S}^p_- := \{y : y_{p+1} < 0\}$, which is disjoint from $\operatorname{supp} G \subset \mathbb{S}^p_+$. Hence, $G(-y) = 0$ and $F(y) = G(y) + G(-y) = G(y) = g(y)$.

(ii) $F \in H^{\frac{p+1}{2}}(\mathbb{S}^p)$, with $\|F\|_{H^{\frac{p+1}{2}}(\mathbb{S}^p)} \leq 2\|G\|_{H^{\frac{p+1}{2}}(\mathbb{S}^p)}$. By Lemma G.10, $\|G \circ \mathfrak{a}\|_{H^{\frac{p+1}{2}}(\mathbb{S}^p)} = \|G\|_{H^{\frac{p+1}{2}}(\mathbb{S}^p)}$. The triangle inequality gives the bound on $F = G + G \circ \mathfrak{a}$.

(iii) F has no spherical-harmonic component of odd degree. By Lemma G.10, $F_k = G_k + (G \circ \mathfrak{a})_k = G_k + (-1)^k G_k = \big(1 + (-1)^k\big) G_k$. For odd k, $1 + (-1)^k = 0$, so $F_k = 0$.

By (ii) and (iii), $F \in \mathcal{E} \cap H^{\frac{p+1}{2}}(\mathbb{S}^p)$. By Corollary G.9, $F \in \mathcal{H}(K_0)$ with $\|F\|_{\mathcal{H}(K_0)} \leq C\|F\|_{H^{\frac{p+1}{2}}(\mathbb{S}^p)} \leq 2CC_2C_3\|g\|_{H^{\frac{p+1}{2}}(\mathcal{V})}$, where $C > 0$ depends only on p. By (i), F is an admissible extension of g, so by the restriction property of RKHSs, $g \in \mathcal{H}(K_0|_{\mathcal{V}})$ with $\|g\|_{\mathcal{H}(K_0|_{\mathcal{V}})} \leq \|F\|_{\mathcal{H}(K_0)} \leq 2CC_2C_3\|g\|_{H^{\frac{p+1}{2}}(\mathcal{V})}$. □

G.3 The RKHS $\mathcal{H}^{\mathrm{NTK}}$ is a Sobolev space

Theorem G.12 (The RKHS of K^{NTK} is a Sobolev space)**.** *Let $p \geq 1$ and let $\mathcal{X} \subset \mathbb{R}^p$ be a nonempty bounded Lipschitz domain, with $\|x\| \leq A$ for any $x \in \mathcal{X}$. It holds that $\mathcal{H}^{\mathrm{NTK}} \cong H^{\frac{p+1}{2}}(\mathcal{X})$ with equivalence constants depending only on p, A, and the Lipschitz character of $\mathcal{X}$.*

Proof. By Corollary G.8, we have $\mathcal{H}^{\mathrm{NTK}} = q \cdot \big[\mathcal{H}(K_0|_{\widehat{\mathcal{X}}}) \circ \phi\big] + \mathbb{R} \cdot \mathbb{1}_{\mathcal{X}}$.

Step 1: identify the spherical factor. We check the three hypotheses of Lemma G.11 for $\widehat{\mathcal{X}} = \phi(\mathcal{X})$.

$\widehat{\mathcal{X}}$ is open. By Lemma G.7, ϕ is a diffeomorphism, hence an open map. Since $\mathcal{X}$ is open, so is $\widehat{\mathcal{X}}$.

$\widehat{\mathcal{X}}$ is bounded with Lipschitz boundary. Boundedness is immediate, since $\mathbb{S}^p$ is compact and $\widehat{\mathcal{X}} \subset \mathbb{S}^p$. For the boundary, let $\mathcal{A}$ be the maximal real-analytic atlas of $\mathbb{S}^p$ (for its standard structure as a submanifold of $\mathbb{R}^{p+1}$), and declare an open $\widehat{\Omega} \subset \mathbb{S}^p$ to have Lipschitz boundary if there is a chart $(\psi, U) \in \mathcal{A}$ with $\overline{\widehat{\Omega}} \subset U$ such that $\psi(\widehat{\Omega})$ is a bounded Lipschitz domain in $\mathbb{R}^p$. This does not depend

on the chart. Indeed, let $(\psi_1, U_1), (\psi_2, U_2) \in \mathcal{A}$ both satisfy $\overline{\widehat{\Omega}} \subset U_1$ and $\overline{\widehat{\Omega}} \subset U_2$. Since $\mathbb{S}^p$ is compact, $\overline{\widehat{\Omega}}$ is compact. Given that $\psi_1 : U_1 \to \mathbb{R}^p$ and $\psi_2 : U_2 \to \mathbb{R}^p$ are homeomorphisms, we have $\overline{\psi_1(\widehat{\Omega})} = \psi_1(\overline{\widehat{\Omega}})$ and $\overline{\psi_2(\widehat{\Omega})} = \psi_2(\overline{\widehat{\Omega}})$. Since $\overline{\widehat{\Omega}} \subset U_1 \cap U_2$, we get $\overline{\psi_1(\widehat{\Omega})} \subset \psi_1(U_1 \cap U_2)$ and $\overline{\psi_2(\widehat{\Omega})} \subset \psi_2(U_1 \cap U_2)$. The transition map $\tau_0 = \psi_2 \circ \psi_1^{-1} \colon \psi_1(U_1 \cap U_2) \to \psi_2(U_1 \cap U_2)$ is a real-analytic C^1 diffeomorphism, with inverse $\tau_0^{-1} = \psi_1 \circ \psi_2^{-1}$. By the two displayed inclusions, τ_0 is defined on a neighborhood of $\overline{\psi_1(\widehat{\Omega})}$ and τ_0^{-1} is defined on a neighborhood of $\overline{\psi_2(\widehat{\Omega})}$, so [50, Theorem 4.1] applies to each: it shows $\psi_1(\widehat{\Omega})$ is a bounded Lipschitz domain if and only if $\tau_0(\psi_1(\widehat{\Omega})) = \psi_2(\widehat{\Omega})$ is, which proves the chart-independence property.

We now specialize to ϕ^{-1}. The map $\phi^{-1} \colon \mathbb{S}^p_+ \to \mathbb{R}^p$, $\phi^{-1}(y) = (y_1, \ldots, y_p)/y_{p+1}$, is a real-analytic diffeomorphism with real-analytic inverse $\phi(x) = \tilde{x}/\sqrt{1 + \|x\|^2}$, so $(\phi^{-1}, \mathbb{S}^p_+) \in \mathcal{A}$. Since $\mathcal{X}$ is bounded, $\overline{\mathcal{X}}$ is compact. The fact that $\phi : \mathbb{R}^p \to \mathbb{S}^p_+$ is a homeomorphism gives $\overline{\widehat{\mathcal{X}}} = \overline{\phi(\mathcal{X})} = \phi(\overline{\mathcal{X}})$. Boundedness of $\overline{\mathcal{X}}$ also keeps $\phi(x)_{p+1} = 1/\sqrt{1 + \|x\|^2}$ bounded away from 0 on $\overline{\mathcal{X}}$, so $\overline{\widehat{\mathcal{X}}} = \phi(\overline{\mathcal{X}}) \subset \mathbb{S}^p_+$. Thus $(\phi^{-1}, \mathbb{S}^p_+)$ is a chart with $\overline{\widehat{\mathcal{X}}} \subset \mathbb{S}^p_+$, and $\phi^{-1}(\widehat{\mathcal{X}}) = \mathcal{X}$ is a bounded Lipschitz domain by hypothesis. Hence, $\widehat{\mathcal{X}}$ has Lipschitz boundary.

$\overline{\widehat{\mathcal{X}}}$ is compactly contained in $\mathbb{S}^p_+$. By Lemma G.7, $\overline{\widehat{\mathcal{X}}}$ lies in the closed cap $\{y \in \mathbb{S}^p : y_{p+1} \geq (1 + A^2)^{-1/2}\}$, which is compact and contained in $\mathbb{S}^p_+$. Hence, $\overline{\widehat{\mathcal{X}}}$ is compactly contained in $\mathbb{S}^p_+$.

All three hypotheses hold, so Lemma G.11 applies and gives $\mathcal{H}\big(K_0|_{\widehat{\mathcal{X}}}\big) \cong H^{\frac{p+1}{2}}(\widehat{\mathcal{X}})$.

Step 2: transport back to $\mathcal{X}$. We prove in this step that $q \cdot \big[\mathcal{H}(K_0|_{\widehat{\mathcal{X}}}) \circ \phi\big] \cong H^{\frac{p+1}{2}}(\mathcal{X})$.

Undoing ϕ. Let $\mathcal{A}$ be the maximal real-analytic atlas of $\mathbb{S}^p$, fixed in Step 1. For any two charts $(\psi_1, U_1), (\psi_2, U_2) \in \mathcal{A}$, the transition map $\tau_0 := \psi_2 \circ \psi_1^{-1}$ is a bijection from $\psi_1(U_1 \cap U_2)$ onto $\psi_2(U_1 \cap U_2)$. Therefore, τ_0 and its inverse $\tau_0^{-1} = \psi_1 \circ \psi_2^{-1}$ are real-analytic, since ψ_1, ψ_2 are. Thus, τ_0 is a real-analytic diffeomorphism between these two open subsets of $\mathbb{R}^p$, and in particular, being real-analytic, τ_0 is C^∞, matching the hypothesis of [15, Theorem 80]. For a bounded open $\widehat{\Omega} \subset \mathbb{S}^p$ with $\overline{\widehat{\Omega}} \subset U$ for some chart $(\psi, U) \in \mathcal{A}$, define $\|g\|_{H^{\frac{p+1}{2}}(\widehat{\Omega}),\psi} := \|g \circ \psi^{-1}\|_{H^{\frac{p+1}{2}}(\psi(\widehat{\Omega}))}$. We claim there is a constant $C_1 > 0$, depending only on p and the pair of charts $(\psi_1, U_1), (\psi_2, U_2)$, such that:

$$\frac{1}{C_1}\|g\|_{H^{\frac{p+1}{2}}(\widehat{\Omega}),\psi_1} \leq \|g\|_{H^{\frac{p+1}{2}}(\widehat{\Omega}),\psi_2} \leq C_1\|g\|_{H^{\frac{p+1}{2}}(\widehat{\Omega}),\psi_1}, \tag{72}$$

whenever $(\psi_1, U_1), (\psi_2, U_2) \in \mathcal{A}$ both contain $\overline{\widehat{\Omega}}$. Indeed, since $\widehat{\Omega} \subset U_1 \cap U_2$, and a diffeomorphism restricted to an open subset of its domain remains a diffeomorphism onto the corresponding open subset of its codomain, τ_0 restricts to a real-analytic diffeomorphism $\tau_0\big|_{\psi_1(\widehat{\Omega})} : \psi_1(\widehat{\Omega}) \to \psi_2(\widehat{\Omega})$, with real-analytic inverse $\tau_0^{-1}\big|_{\psi_2(\widehat{\Omega})} : \psi_2(\widehat{\Omega}) \to \psi_1(\widehat{\Omega})$. We retain the notation τ_0, τ_0^{-1} for these restrictions. We have $g \circ \psi_2^{-1} = (g \circ \psi_1^{-1}) \circ \tau_0^{-1}$, and [15, Theorem 80] applied to τ_0^{-1} between $\psi_2(\widehat{\Omega})$ and $\psi_1(\widehat{\Omega})$ gives a constant $C_1' > 0$ with:

$$\|(g \circ \psi_1^{-1}) \circ \tau_0^{-1}\|_{H^{\frac{p+1}{2}}(\psi_2(\widehat{\Omega}))} \leq C_1'\|g \circ \psi_1^{-1}\|_{H^{\frac{p+1}{2}}(\psi_1(\widehat{\Omega}))},$$

i.e., $\|g\|_{H^{\frac{p+1}{2}}(\widehat{\Omega}),\psi_2} \leq C_1'\|g\|_{H^{\frac{p+1}{2}}(\widehat{\Omega}),\psi_1}$. Conversely, $g \circ \psi_1^{-1} = (g \circ \psi_2^{-1}) \circ \tau_0$, and applying [15, Theorem 80] directly to τ_0 between $\psi_1(\widehat{\Omega})$ and $\psi_2(\widehat{\Omega})$ gives a constant $C_1'' > 0$ with:

$$\|(g \circ \psi_2^{-1}) \circ \tau_0\|_{H^{\frac{p+1}{2}}(\psi_1(\widehat{\Omega}))} \leq C_1''\,\|g \circ \psi_2^{-1}\|_{H^{\frac{p+1}{2}}(\psi_2(\widehat{\Omega}))},$$

i.e., $\|g\|_{H^{\frac{p+1}{2}}(\widehat{\Omega}),\psi_1} \leq C_1''\|g\|_{H^{\frac{p+1}{2}}(\widehat{\Omega}),\psi_2}$. Taking $C_1 := \max(C_1', C_1'')$ gives Equation (72).

Specialization to ϕ^{-1}. The map $\phi^{-1} \colon \mathbb{S}^p_+ \to \mathbb{R}^p$, $\phi^{-1}(y) = (y_1, \ldots, y_p)/y_{p+1}$, is a real-analytic diffeomorphism with real-analytic inverse $\phi(x) = \tilde{x}/\sqrt{1 + \|x\|^2}$, so $(\phi^{-1}, \mathbb{S}^p_+) \in \mathcal{A}$ by maximality of $\mathcal{A}$. As shown above, $\overline{\widehat{\mathcal{X}}} = \phi(\overline{\mathcal{X}}) \subset \mathbb{S}^p_+$, so $(\phi^{-1}, \mathbb{S}^p_+)$ is a chart of $\mathcal{A}$ containing $\overline{\widehat{\mathcal{X}}}$ in its domain,

exactly as required by Equation (72). Fix any anchor chart $(\psi_0, U_0) \in \mathcal{A}$ with $\overline{\widehat{\mathcal{X}}} \subset U_0$, and set $\|g\|_{H^{\frac{p+1}{2}}(\widehat{\mathcal{X}})} := \|g\|_{H^{\frac{p+1}{2}}(\widehat{\mathcal{X}}),\psi_0}$, which by Step 1 is independent of this choice of chart. Applying Equation (72) directly with $\widehat{\Omega} := \widehat{\mathcal{X}}$, $\psi_1 := \psi_0$, $\psi_2 := \phi^{-1}$ gives a constant $C_2 > 0$ with:

$$\frac{1}{C_2}\|g\|_{H^{\frac{p+1}{2}}(\widehat{\mathcal{X}})} \leq \|g\|_{H^{\frac{p+1}{2}}(\widehat{\mathcal{X}}),\phi^{-1}} \leq C_2\|g\|_{H^{\frac{p+1}{2}}(\widehat{\mathcal{X}})}. \tag{73}$$

Since $\|g\|_{H^{\frac{p+1}{2}}(\widehat{\mathcal{X}}),\phi^{-1}} = \|g \circ \phi\|_{H^{\frac{p+1}{2}}(\mathcal{X})}$, Equation (73) gives $\|g\|_{H^{\frac{p+1}{2}}(\widehat{\mathcal{X}})} \asymp \|g \circ \phi\|_{H^{\frac{p+1}{2}}(\mathcal{X})}$, *i.e.*, $H^{\frac{p+1}{2}}(\widehat{\mathcal{X}}) \cong H^{\frac{p+1}{2}}(\mathcal{X})$ via the pullback $g \mapsto g \circ \phi$.

Undoing q. The function $q(x) = \|\tilde{x}\|$ is real-analytic with $1 \leq q \leq \sqrt{1+A^2}$ on $\mathcal{X}$. Since $\mathcal{X}$ is a bounded Lipschitz domain, the multiplication theorem for Sobolev spaces on Lipschitz domains [15, Corollary 3] shows $H^{\frac{p+1}{2}}(\mathcal{X})$ is a Banach algebra under pointwise multiplication. Applying this to q and to $1/q$ (likewise real-analytic and bounded on $\mathcal{X}$, since $q \geq 1$), there is a constant $C_3 > 0$, depending only on A and p, such that multiplication by q is an isomorphism of $H^{\frac{p+1}{2}}(\mathcal{X})$, with:

$$\frac{1}{C_3}\|f\|_{H^{\frac{p+1}{2}}(\mathcal{X})} \leq \|qf\|_{H^{\frac{p+1}{2}}(\mathcal{X})} \leq C_3\|f\|_{H^{\frac{p+1}{2}}(\mathcal{X})}.$$

Composing. Combining the two isomorphisms above with the one from Step 1 ($\mathcal{H}(K_0|_{\widehat{\mathcal{X}}}) \cong H^{\frac{p+1}{2}}(\widehat{\mathcal{X}})$), we get $q \cdot \left[\mathcal{H}(K_0|_{\widehat{\mathcal{X}}}) \circ \phi\right] \cong H^{\frac{p+1}{2}}(\mathcal{X})$.

Step 3: absorb the outer bias. Recall from Corollary G.8 that $\mathcal{H}^{\mathrm{NTK}} = q \cdot \left[\mathcal{H}(K_0|_{\widehat{\mathcal{X}}}) \circ \phi\right] + \mathbb{R} \cdot \mathbb{1}_{\mathcal{X}}$. Write $k_1 := \mathbb{1}_{\mathcal{X}}$ for the constant kernel and k_2 for the kernel underlying the first summand, so that $\mathcal{H}(k_2) = q \cdot [\mathcal{H}(K_0|_{\widehat{\mathcal{X}}}) \circ \phi]$. By Step 2, $\mathcal{H}(k_2) \cong H^{\frac{p+1}{2}}(\mathcal{X})$. Since $\mathbb{1}_{\mathcal{X}}$ has vanishing weak derivatives of every order, only the L^2 term of the Sobolev norm survives, and $\|\mathbb{1}_{\mathcal{X}}\|^2_{H^{\frac{p+1}{2}}(\mathcal{X})} = \|\mathbb{1}_{\mathcal{X}}\|^2_{L^2(\mathcal{X})} = \int_{\mathcal{X}} 1\, dx = \mathrm{Vol}(\mathcal{X}) < \infty$, where $\mathrm{Vol}(\cdot)$ denotes the volume. Finiteness follows from the boundedness of $\mathcal{X}$. Thus $\mathbb{1}_{\mathcal{X}} \in H^{\frac{p+1}{2}}(\mathcal{X})$. Since $\mathcal{H}(k_1) = \mathbb{R} \cdot \mathbb{1}_{\mathcal{X}}$ with $\|c\mathbb{1}_{\mathcal{X}}\|_{\mathcal{H}(k_1)} = |c|$ for $c \in \mathbb{R}$, the homogeneity of the $H^{\frac{p+1}{2}}(\mathcal{X})$ norm and the previous display give $\|c\mathbb{1}_{\mathcal{X}}\|_{H^{\frac{p+1}{2}}(\mathcal{X})} = |c|\sqrt{\mathrm{Vol}(\mathcal{X})} = \sqrt{\mathrm{Vol}(\mathcal{X})}\|c\mathbb{1}_{\mathcal{X}}\|_{\mathcal{H}(k_1)}$. Hence, $\mathcal{H}(k_1)$ embeds continuously into $H^{\frac{p+1}{2}}(\mathcal{X}) \cong \mathcal{H}(k_2)$, with embedding constant $\sqrt{\mathrm{Vol}(\mathcal{X})}$. Applying Corollary G.5 with this k_1 and k_2 yields $\mathcal{H}^{\mathrm{NTK}} = \mathcal{H}(k_1) + \mathcal{H}(k_2) \cong \mathcal{H}(k_2) \cong H^{\frac{p+1}{2}}(\mathcal{X})$, which is the claimed identification. □

G.4 The interpolation spaces of $\mathcal{H}^{\mathrm{NTK}}$ are Sobolev spaces

For two Hilbert spaces E_0, E_1 continuously embedded in a common space, we denote the *real interpolation space* by $(E_0, E_1)_{s,2}$, $s \in (0,1)$, defined via the standard K-method [17, Chapter 3].

Corollary G.13 (A Sobolev characterization of $[\mathcal{H}^{\mathrm{NTK}}]^s$)**.** *Let $p \geq 1$ and $\mathcal{X} \subset \mathbb{R}^p$ be a nonempty bounded Lipschitz domain, with $\|x\| \leq A$ for any $x \in \mathcal{X}$. Assume that μ, the data distribution on $\mathcal{X}$, has a density f_μ bounded away from 0 and ∞, i.e., there exist $c_\mu, M_\mu > 0$ such that $c_\mu \leq f_\mu \leq M_\mu$. Then, for every $s \in [0,1]$, $[\mathcal{H}^{\mathrm{NTK}}]^s \cong H^{\frac{s(p+1)}{2}}(\mathcal{X})$, with equivalence constants depending on p, s, A, the Lipschitz character of $\mathcal{X}$, and the density bounds c_μ, M_μ.*

Proof. By Theorem G.12, we have $\mathcal{H}^{\mathrm{NTK}} \cong H^{\frac{p+1}{2}}(\mathcal{X})$. Since $0 < c_\mu \leq f_\mu \leq M_\mu$, the norms $\|\cdot\|_{L^2(\mathcal{X},\mathrm{d}x)}$ and $\|\cdot\|_{L^2(\mathcal{X})}$ are equivalent, so $L^2(\mathcal{X},\mathrm{d}x) \cong L^2(\mathcal{X})$. We use this identification below.

Interior case $s \in (0,1)$**.** Let $C_c^\infty(\mathcal{X})$ be the space of smooth, compactly supported functions on $\mathcal{X}$. It is standard that $C_c^\infty(\mathcal{X})$ is dense in $L^2(\mathcal{X},\mathrm{d}x)$. By the identification above, $C_c^\infty(\mathcal{X})$ is therefore dense in $L^2(\mathcal{X})$. Since $H^{\frac{p+1}{2}}(\mathcal{X}) \supset C_c^\infty(\mathcal{X})$, it follows that $H^{\frac{p+1}{2}}(\mathcal{X})$ is dense in $L^2(\mathcal{X})$.

The embedding $H^{\frac{p+1}{2}}(\mathcal{X}) \hookrightarrow L^2(\mathcal{X})$ is compact. Write $\kappa := \lfloor \frac{p+1}{2} \rfloor$. If p is odd, $\frac{p+1}{2} = \kappa$ is already an integer, and the Rellich–Kondrachov theorem [2, Theorem. 6.3, Part II, Eq. (6)] gives directly that $H^{\frac{p+1}{2}}(\mathcal{X}) = H^\kappa(\mathcal{X}) \hookrightarrow L^2(\mathcal{X},\mathrm{d}x)$ is compact. If p is even, $\kappa = \frac{p}{2} < \frac{p+1}{2}$, and the Rellich–Kondrachov theorem [2, Theorem. 6.3, Part I, Eq. (4)] shows that $H^\kappa(\mathcal{X}) \hookrightarrow L^2(\mathcal{X},\mathrm{d}x)$

is compact. By definition, $H^{\frac{p+1}{2}}(\mathcal{X}) \hookrightarrow H^{\kappa}(\mathcal{X})$ is continuous, and composing the two gives $H^{\frac{p+1}{2}}(\mathcal{X}) \hookrightarrow L^2(\mathcal{X}, \mathrm{d}x)$ compact. In either case, by the identification $L^2(\mathcal{X}, \mathrm{d}x) \cong L^2(\mathcal{X})$, $H^{\frac{p+1}{2}}(\mathcal{X}) \hookrightarrow L^2(\mathcal{X})$ is compact.

By [87, Theorem 4.6], $[\mathcal{H}^{\mathrm{NTK}}]^s \cong (L^2(\mathcal{X}), \mathcal{H}^{\mathrm{NTK}})_{s,2}$. Since $\mathcal{H}^{\mathrm{NTK}} \cong H^{\frac{p+1}{2}}(\mathcal{X})$, [24, Lemma H.5] gives $(L^2(\mathcal{X}), \mathcal{H}^{\mathrm{NTK}})_{s,2} \cong \left(L^2(\mathcal{X}), H^{\frac{p+1}{2}}(\mathcal{X})\right)_{s,2}$, so $[\mathcal{H}^{\mathrm{NTK}}]^s \cong \left(L^2(\mathcal{X}), H^{\frac{p+1}{2}}(\mathcal{X})\right)_{s,2}$. Using the identification $L^2(\mathcal{X}, \mathrm{d}x) \cong L^2(\mathcal{X})$ and the fact that $L^2(\mathcal{X}, \mathrm{d}x) \cong H^0(\mathcal{X})$, we get $L^2(\mathcal{X}) \cong H^0(\mathcal{X})$. [24, Lemma H.6] then gives $\left(L^2(\mathcal{X}), H^{\frac{p+1}{2}}(\mathcal{X})\right)_{s,2} \cong \left(H^0(\mathcal{X}), H^{\frac{p+1}{2}}(\mathcal{X})\right)_{s,2}$. Therefore, by [17, Theorem 6.4.5], we obtain $[\mathcal{H}^{\mathrm{NTK}}]^s \cong \left(H^0(\mathcal{X}), H^{\frac{p+1}{2}}(\mathcal{X})\right)_{s,2} = H^{\frac{s(p+1)}{2}}(\mathcal{X})$.

Endpoints $s \in \{0, 1\}$**.** Directly from the definition: $[\mathcal{H}^{\mathrm{NTK}}]^0 = L^2(\mathcal{X}) \cong H^0(\mathcal{X})$, and $[\mathcal{H}^{\mathrm{NTK}}]^1 = \mathcal{H}^{\mathrm{NTK}} \cong H^{\frac{p+1}{2}}(\mathcal{X})$ by Theorem G.12. □

H Background on Integral/Covariance Operators and Interpolation Spaces

Let $\mathcal{X} \subset \mathbb{R}^p$ be the input space, $\mathcal{Y} \subset \mathbb{R}$ be the target space, and ρ be an unknown probability distribution supported on $\mathcal{X} \times \mathcal{Y}$ with marginal distribution μ on $\mathcal{X}$. Let $L^2(\mathcal{X})$ denote the space of square-integrable functions with respect to μ defined on the input space $\mathcal{X}$ and taking values in $\mathcal{Y}$. Let $\mathcal{H}$ be an RKHS defined on $\mathcal{X}$ and associated with a kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. We assume that K is a *Mercer* kernel, *i.e.*, a continuous symmetric positive semi-definite kernel defined on a compact space $\mathcal{X}$. This implies that K is bounded, *i.e.*, there exists $\kappa \geq 0$ such that $K(x, x) \leq \kappa^2$ for any $x \in \mathcal{X}$, and that $\mathcal{H}$ is separable, *i.e.*, it contains a countable dense subset [86, Lemma 4.33]. Under this boundedness condition, it holds that $\mathcal{H} \subset L^2(\mathcal{X})$. Indeed, for any $f \in \mathcal{H}$ and $x \in \mathcal{X}$, we can write by the reproducing property that $f(x) = \langle f, K(x, \cdot)\rangle_{\mathcal{H}}$, which implies by Cauchy-Schwarz inequality that $|f(x)| \leq \|f\|_{\mathcal{H}}\sqrt{K(x, x)} \leq \kappa\|f\|_{\mathcal{H}}$. This gives that $\|f\|^2_{L^2(\mathcal{X})} \leq \kappa^2\|f\|^2_{\mathcal{H}} < +\infty$, and hence that $f \in L^2(\mathcal{X})$.

Integral and covariance operators. Let $\mathcal{I} : \mathcal{H} \to L^2(\mathcal{X})$ be the natural embedding inclusion operator defined for any $f \in \mathcal{H}$ by $\mathcal{I}(f) = f$. Its adjoint $\mathcal{I}^\star : L^2(\mathcal{X}) \to \mathcal{H}$ is defined for any $f \in L^2(\mathcal{X})$ by $\mathcal{I}^\star(f) = \int_{\mathcal{X}} K(x, \cdot)f(x)\mathrm{d}\mu(x) \in \mathcal{H}$. Both $\mathcal{I}$ and $\mathcal{I}^\star$ are Hilbert-Schmidt operators, with Hilbert-Schmidt norms:

$$\|\mathcal{I}\|_{\mathrm{HS}} = \|\mathcal{I}^\star\|_{\mathrm{HS}} = \sqrt{\int_{\mathcal{X}} K(x, x)\mathrm{d}\mu(x)} \leq \kappa,$$

and are in particular compact. Let $\Sigma = \mathcal{I}^\star\mathcal{I} : \mathcal{H} \to \mathcal{H}$ be the covariance operator defined by $\Sigma = \mathbb{E}_{x\sim\mu}[K(x, \cdot) \otimes K(x, \cdot)]$, where $(a \otimes b)c = \langle b, c\rangle_{\mathcal{H}} a$ for $a, b, c \in \mathcal{H}$. Let $\mathcal{T} = \mathcal{I}\mathcal{I}^\star : L^2(\mathcal{X}) \to L^2(\mathcal{X})$ be the integral operator defined for any $f \in L^2(\mathcal{X})$ by $\mathcal{T}f = \int_{\mathcal{X}} K(\cdot, t)f(t)\mathrm{d}\mu(t)$. Since K is a Mercer kernel, it follows that $\mathcal{T}$ is a compact, self-adjoint, and positive operator on $L^2(\mathcal{X})$ [86, Theorem 4.27]. By the *spectral theorem*, there exists an at most countable index set N, a non-increasing summable sequence $(\lambda_i)_{i\in N} \subset (0, +\infty)$, and an orthonormal family $(e_i)_{i\in N} \subset L^2(\mathcal{X})$ forming an orthonormal basis of $\overline{\mathrm{Range}(\mathcal{I})} \subset L^2(\mathcal{X})$, such that $\{\sqrt{\lambda_i}e_i\}_{i\in N}$ is an orthonormal basis of $\mathcal{H}$ [86, Section 4.5]. The operators $\mathcal{T}$ and Σ admit the spectral decompositions:

$$\mathcal{T} = \sum_{i\in N} \lambda_i\langle \cdot, e_i\rangle_{L^2(\mathcal{X})}e_i \quad \text{and} \quad \Sigma = \sum_{i\in N} \lambda_i\langle \cdot, \sqrt{\lambda_i}e_i\rangle_{\mathcal{H}}\sqrt{\lambda_i}e_i,$$

and share the same non-zero eigenvalues $(\lambda_i)_{i\in N}$, since $\mathcal{T} = \mathcal{I}\mathcal{I}^\star$ and $\Sigma = \mathcal{I}^\star\mathcal{I}$ have the same non-zero spectrum. Moreover, by Mercer's theorem [86, Theorem 4.49], we have for any $x, x' \in \mathcal{X}$:

$$K(x, x') = \sum_{i\in N} \lambda_i e_i(x)e_i(x'),$$

where the convergence is absolute and uniform. Since N is at most countable, there exists a bijection $\phi : \mathbb{N} \to N$; hence, by reindexing via ϕ, we may write the eigenpairs as $(\lambda_{\phi(i)}, e_{\phi(i)})_{i\geq 1}$, and, by a slight abuse of notation, simply denote them by $(\lambda_i, e_i)_{i\geq 1}$.

Interpolation space. For $s \geq 0$, we define the *interpolation space* $[\mathcal{H}]^s$ by [39, Section 2]:

$$[\mathcal{H}]^s = \mathrm{Range}(\mathcal{T}^{\frac{s}{2}}) = \left\{ \sum_{i\geq 1} a_i \lambda_i^{\frac{s}{2}} e_i \;\middle|\; \sum_{i\geq 1} a_i^2 < +\infty \right\} \subset L^2(\mathcal{X})$$

equipped with the norm $\left\| \sum_{i\geq 1} a_i \lambda_i^{\frac{s}{2}} e_i \right\|_{[\mathcal{H}]^s} = \left(\sum_{i\geq 1} a_i^2 \right)^{\frac{1}{2}}$. Note that $[\mathcal{H}]^0 \subset L^2(\mathcal{X})$ and $[\mathcal{H}]^1 = \mathcal{H}$. The functions in $[\mathcal{H}]^s$ with smaller values of s exhibit *lower smoothness* and are therefore more difficult for kernel regression algorithms to estimate.

I Auxiliary Technical Lemmas

In Lemmas I.1 to I.3, let W_0 be the principal branch of the Lambert W function, *i.e.*, the real-valued solution $v = W_0(z) \geq 0$ of an equation of the form $v \exp(v) = z$ for $z \geq 0$.

Lemma I.1. *Let $C_1, C_2 > 0$, $0 < \alpha < 1$, $0 < \gamma < 2$, and $x \geq \left\lfloor \left(\frac{4C_1 (\ln 6)^3}{C_2} \right)^{\frac{1}{\alpha+\gamma}} \right\rfloor + 1$. Let $\varphi : (0,1) \to (0, +\infty)$ be the function defined for any $\delta \in (0,1)$ by:*

$$\varphi(\delta) = C_1 \left(\ln \frac{6}{\delta} \right)^4 x^{-\alpha} + C_2 x^\gamma \delta.$$

Then, φ has a unique minimizer $\delta^\star \in (0,1)$, given by:

$$\delta^\star = \frac{108 C_1}{C_2} x^{-(\alpha+\gamma)} W_0^3 \left(x^{\frac{\alpha+\gamma}{3}} \left(\frac{C_2}{18 C_1} \right)^{\frac{1}{3}} \right),$$

and the corresponding minimum value is:

$$\varphi(\delta^\star) = 27 C_1 W_0^3 \left(x^{\frac{\alpha+\gamma}{3}} \left(\frac{C_2}{18 C_1} \right)^{\frac{1}{3}} \right) \left(3 W_0 \left(x^{\frac{\alpha+\gamma}{3}} \left(\frac{C_2}{18 C_1} \right)^{\frac{1}{3}} \right) + 4 \right) x^{-\alpha}.$$

Proof. **Existence and uniqueness of $\delta^\star \in (0,1)$.** The derivative of φ for any $\delta \in (0,1)$ is:

$$\varphi'(\delta) = -\frac{4 C_1 \left(\ln \frac{6}{\delta} \right)^3 x^{-\alpha}}{\delta} + C_2 x^\gamma.$$

The first term is strictly increasing in δ on $(0,1)$, since $\ln(6/\delta)$ and $1/\delta$ are positive and decrease as δ increases, while the second term is constant. Hence, φ' is strictly increasing on $(0,1)$. Moreover, $\varphi'(\delta) \to -\infty$ as $\delta \to 0^+$, and $\varphi'(\delta) \to -4C_1 (\ln 6)^3 x^{-\alpha} + C_2 x^\gamma$ as $\delta \to 1^-$. Since $x \geq \left\lfloor \left(\frac{4C_1 (\ln 6)^3}{C_2} \right)^{\frac{1}{\alpha+\gamma}} \right\rfloor + 1 > \left(\frac{4C_1 (\ln 6)^3}{C_2} \right)^{\frac{1}{\alpha+\gamma}}$, we have $x^{\alpha+\gamma} > \frac{4C_1 (\ln 6)^3}{C_2}$, so that the limit $-4C_1 (\ln 6)^3 x^{-\alpha} + C_2 x^\gamma$ is positive, and $\varphi'(\delta) > 0$ for every δ close enough to 1. Therefore, by the intermediate value theorem applied to the continuous function φ', there exists $\delta^\star \in (0,1)$ such that $\varphi'(\delta^\star) = 0$. Since φ' is strictly increasing, this zero is unique, and since φ' changes sign from negative to positive at $\delta^\star$, it is the unique minimizer of φ on $(0,1)$.

Formula for $\delta^\star$. The minimizer $\delta^\star$ satisfies $\varphi'(\delta^\star) = 0$, that is:

$$\frac{4 C_1 x^{-\alpha}}{\delta^\star} \left(\ln \frac{6}{\delta^\star} \right)^3 = C_2 x^\gamma.$$

Setting $y = \ln \frac{6}{\delta^\star}$, so that $\delta^\star = 6 \exp(-y)$, we obtain:

$$y^3 \exp(y) = \frac{3 C_2}{2 C_1} x^{\alpha+\gamma} \iff y \exp\left(\frac{y}{3} \right) = \left(\frac{3 C_2}{2 C_1} x^{\alpha+\gamma} \right)^{\frac{1}{3}}.$$

Equivalently, setting $z = x^{\frac{\alpha+\gamma}{3}} \left(\frac{C_2}{18 C_1} \right)^{\frac{1}{3}}$, $\frac{y}{3} \exp\left(\frac{y}{3} \right) = z$, since $\frac{1}{3} \left(\frac{3C_2}{2C_1} \right)^{\frac{1}{3}} = \left(\frac{C_2}{18 C_1} \right)^{\frac{1}{3}}$. Using the identity $W_0(w) \exp(W_0(w)) = w$, we get $y = 3 W_0(z)$. Since $\delta^\star = 6 \exp(-y)$, the identity

$\exp(-W_0(w)) = \frac{W_0(w)}{w}$ gives $\delta^\star = 6\left[\exp\left(-W_0(z)\right)\right]^3 = \frac{6W_0^3(z)}{z^3} = \frac{108C_1}{C_2}x^{-(\alpha+\gamma)}W_0^3(z)$, where we used $z^3 = \frac{C_2}{18C_1}x^{\alpha+\gamma}$ in the last equality.

Formula for $\varphi(\delta^\star)$. From the optimality condition, $C_2x^\gamma\delta^\star = 4C_1y^3x^{-\alpha}$. As a consequence:

$$\begin{aligned}\varphi(\delta^\star) &= C_1\left(\ln\frac{6}{\delta^\star}\right)^4 x^{-\alpha} + C_2x^\gamma\delta^\star = C_1y^4x^{-\alpha} + 4C_1y^3x^{-\alpha}\\ &= C_1y^3(y+4)x^{-\alpha} = 27C_1W_0^3(z)\left(3W_0(z)+4\right)x^{-\alpha}.\end{aligned}$$ □

Lemma I.2 (Upper-bound on W_0)**.** *For any $x \geq e$, it holds that $W_0(x) \leq \ln x$.*

Proof. Define $F(x) = \ln x - W_0(x)$ for any $x \geq e$. We have that:

$$\frac{\mathrm{d}}{\mathrm{d}x}F(x) = \frac{1}{x} - \frac{\mathrm{d}}{\mathrm{d}x}W_0(x). \tag{74}$$

To compute the derivative of $W_0(x)$ with respect to x, we differentiate the identity $W_0(x)\exp(W_0(x)) = x$. Using the product and chain rules yields:

$$\left(\frac{\mathrm{d}}{\mathrm{d}x}W_0(x)\right)\exp(W_0(x)) + W_0(x)\exp(W_0(x))\left(\frac{\mathrm{d}}{\mathrm{d}x}W_0(x)\right) = 1.$$

After rearranging the terms and using the identity $W_0(x)\exp(W_0(x)) = x$, we obtain:

$$\frac{\mathrm{d}}{\mathrm{d}x}W_0(x) = \frac{W_0(x)}{x(1+W_0(x))}.$$

Note that this derivative is defined for any $x \geq e$ because $x(1+W_0(x)) \neq 0$, which is due to the fact that $W_0(x) \geq 0$ for any $x \geq e$. Substituting the obtained derivative into Equation (74) gives:

$$\frac{\mathrm{d}}{\mathrm{d}x}F(x) = \frac{1}{x} - \frac{W_0(x)}{x(1+W_0(x))} = \frac{1}{x(1+W_0(x))} \geq 0.$$

Thus, F is increasing on $[e, +\infty)$. Since $F(e) = \ln e - W_0(e) = 0$, we get $F(x) \geq 0$ for any $x \geq e$, which implies that $W_0(x) \leq \ln x$. This concludes the proof. □

Lemma I.3. *Let $0 < \alpha < \gamma$ and $\kappa > 0$. When $x \geq \max(e, (\frac{e}{\kappa})^{\frac{1}{\alpha}})$, the following holds:*

$$\begin{aligned}W_0^3\left(x^\alpha\kappa\right)\left(3W_0\left(x^\alpha\kappa\right)+4\right) \leq &\Big(4\gamma^3 + 3\gamma^4 + 12(\gamma^2+\gamma^3)\left|\ln\kappa\right| + (12\gamma+18\gamma^2)(\ln\kappa)^2\\ &+ (4+12\gamma)\left|\ln\kappa\right|^3 + 3(\ln\kappa)^4\Big)(\ln x)^4.\end{aligned}$$

Proof. Fix $0 < \alpha < \gamma$ and $\kappa > 0$. We have $x^\alpha\kappa \geq e$ since it is assumed that $x \geq \left(\frac{e}{\kappa}\right)^{\frac{1}{\alpha}}$. Therefore, by Lemma I.2, we get $W_0\left(x^\alpha\kappa\right) \leq \ln\left(x^\alpha\kappa\right) = \alpha\ln x + \ln\kappa$. Using this, we have:

$$W_0^3(x^\alpha\kappa) \leq \left(\alpha\ln x + \ln\kappa\right)^3 = \alpha^3(\ln x)^3 + 3\alpha^2(\ln x)^2\ln\kappa + 3\alpha\ln x(\ln\kappa)^2 + (\ln\kappa)^3,$$

and

$$\begin{aligned}&W_0^3\left(x^\alpha\kappa\right)\left(3W_0\left(x^\alpha\kappa\right)+4\right) \leq \left(\alpha\ln x + \ln\kappa\right)^3\left(3\alpha\ln x + 3\ln\kappa + 4\right)\\ =&\left(\alpha^3(\ln x)^3 + 3\alpha^2(\ln x)^2\ln\kappa + 3\alpha\ln x(\ln\kappa)^2 + (\ln\kappa)^3\right)\left(3\alpha\ln x + 3\ln\kappa + 4\right)\\ =&3\alpha^4(\ln x)^4 + 3\alpha^3\ln\kappa(\ln x)^3 + 4\alpha^3(\ln x)^3 + 9\alpha^3(\ln x)^3\ln\kappa + 9\alpha^2(\ln x)^2(\ln\kappa)^2\\ &+ 12\alpha^2(\ln x)^2\ln\kappa + 9\alpha^2(\ln x)^2(\ln\kappa)^2 + 9\alpha\ln x(\ln\kappa)^3 + 12\alpha\ln x(\ln\kappa)^2\\ &+ 3\alpha\ln x(\ln\kappa)^3 + 3(\ln\kappa)^4 + 4(\ln\kappa)^3\\ =&3\alpha^4(\ln x)^4 + 12\alpha^3\ln\kappa(\ln x)^3 + 4\alpha^3(\ln x)^3 + 18\alpha^2(\ln x)^2(\ln\kappa)^2 + 12\alpha^2(\ln x)^2\ln\kappa\\ &+ 12\alpha\ln x(\ln\kappa)^3 + 12\alpha\ln x(\ln\kappa)^2 + 3(\ln\kappa)^4 + 4(\ln\kappa)^3.\end{aligned}$$

It is also assumed that $x \geq e$, hence $(\ln x)^i \leq (\ln x)^j$ for $0 \leq i < j \leq 4$. Also, we know that $\alpha < \gamma$. Consequently, we get $W_0^3(x^\alpha \kappa) \leq \left(\gamma^3 + 3\gamma^2 |\ln \kappa| + 3\gamma (\ln \kappa)^2 + |\ln \kappa|^3\right) (\ln x)^3$ and

$$
\begin{aligned}
& W_0^3 \left(x^\alpha \kappa\right) \left(3 W_0 \left(x^\alpha \kappa\right) + 4\right) \\
\leq & \Big(3\gamma^4 + 12\gamma^3 |\ln \kappa| + 4\gamma^3 + 18\gamma^2 (\ln \kappa)^2 + 12\gamma^2 |\ln \kappa| + 12\gamma |\ln \kappa|^3 + 12\gamma (\ln \kappa)^2 + 3(\ln \kappa)^4 \\
& \quad + 4 |\ln \kappa|^3 \Big) (\ln x)^4 \\
= & \left(4\gamma^3 + 3\gamma^4 + 12(\gamma^2 + \gamma^3) |\ln \kappa| + (12\gamma + 18\gamma^2)(\ln \kappa)^2 + (4 + 12\gamma) |\ln \kappa|^3 + 3(\ln \kappa)^4\right) (\ln x)^4.
\end{aligned}
$$

This concludes the proof. □

Lemma I.4 (Surface measure of Euclidean spherical caps)**.** *Let $\hat{y} \in \mathbb{S}^p$ and $\epsilon > 0$. Define the Euclidean spherical cap of radius ϵ centered at $\hat{y}$ as the Euclidean ball of radius ϵ centered at $\hat{y}$ intersected with the sphere: $B_{\mathbb{S}^p}(\hat{y}, \epsilon) := \{\hat{x} \in \mathbb{S}^p \mid \|\hat{x} - \hat{y}\| \leq \epsilon\}$. Then, it holds that $\sigma\left(B_{\mathbb{S}^p}(\hat{y}, \epsilon)\right) \leq v_p \left(\frac{\pi}{2}\epsilon\right)^p$, where σ is the surface measure on $\mathbb{S}^p$ and $v_p = \mathrm{Vol}(B(0,1))$ is the volume of the unit ball in $\mathbb{R}^p$.*

Proof. We denote by d_g the geodesic distance and fix $\hat{x}, \hat{y} \in \mathbb{S}^p$. On the unit sphere $\mathbb{S}^p$, we have $d_g(\hat{x}, \hat{y}) = \theta$, where $\theta \in [0, \pi]$ is the angle between $\hat{x}$ and $\hat{y}$. Since $\|\hat{x}\| = \|\hat{y}\| = 1$, it follows that:

$$
\|\hat{x} - \hat{y}\|^2 = \|\hat{x}\|^2 + \|\hat{y}\|^2 - 2\langle \hat{x}, \hat{y} \rangle = 2 - 2\cos\theta = 4\sin^2\left(\frac{\theta}{2}\right),
$$

where we use the identity $1 - \cos\theta = 2\sin^2\left(\frac{\theta}{2}\right)$ in the last equality. Hence, we obtain $\|\hat{x} - \hat{y}\| = 2\sin\left(\frac{\theta}{2}\right)$. Solving for θ gives $d_g(\hat{x}, \hat{y}) := \theta = 2\arcsin\left(\frac{\|\hat{x} - \hat{y}\|}{2}\right)$. Jordan's inequality states that $\sin\theta \geq \frac{2\theta}{\pi}$ for $\theta \in [0, \frac{\pi}{2}]$, or equivalently $\arcsin(t) \leq \frac{\pi t}{2}$ for $t \in [0, 1]$. Since $\frac{\|\hat{x} - \hat{y}\|}{2} \in [0, 1]$ for any $\hat{x}, \hat{y} \in \mathbb{S}^p$, we obtain $d_g(\hat{x}, \hat{y}) \leq \frac{\pi}{2}\|\hat{x} - \hat{y}\|$. Therefore, we get $B_{\mathbb{S}^p}(\hat{y}, \epsilon) \subset B_g(\hat{y}, \frac{\pi}{2}\epsilon)$, where $B_g(\hat{y}, \frac{\pi}{2}\epsilon) := \{\hat{x} \in \mathbb{S}^p \mid d_g(\hat{x}, \hat{y}) \leq \frac{\pi}{2}\epsilon\}$ is the geodesic ball on the sphere of radius $\frac{\pi}{2}\epsilon$ and centered at $\hat{y}$. The surface measure of a geodesic ball of radius r on $\mathbb{S}^p$ satisfies [61]:

$$
\sigma\left(B_g(\hat{y}, r)\right) = w_{p-1} \int_0^r \sin^{p-1} t \, \mathrm{d}t \leq w_{p-1} \int_0^r t^{p-1} \, \mathrm{d}t = \frac{w_{p-1}}{p} r^p,
$$

where we use the fact that $\sin t \leq t$ for $t \geq 0$, with w_{p-1} being the surface area of $\mathbb{S}^{p-1}$. Additionally, we have that $w_{p-1} = \frac{2\pi^{\frac{p}{2}}}{\Gamma\left(\frac{p}{2}\right)}$ and $v_p = \frac{\pi^{\frac{p}{2}}}{\Gamma\left(\frac{p}{2}+1\right)}$, where v_p is the volume of the unit ball in $\mathbb{R}^p$ and Γ is the gamma function. Using the identity $\Gamma(x+1) = x\Gamma(x)$, we obtain:

$$
v_p = \frac{\pi^{\frac{p}{2}}}{\left(\frac{p}{2}\right)\Gamma\left(\frac{p}{2}\right)} = \frac{w_{p-1}}{p},
$$

which yields $\sigma\left(B_g(\hat{y}, r)\right) \leq v_p \, r^p$. Using this inequality and the already established inclusion $B_{\mathbb{S}^p}(\hat{y}, \epsilon) \subset B_g(\hat{y}, \frac{\pi}{2}\epsilon)$, we get:

$$
\sigma\left(B_{\mathbb{S}^p}(\hat{y}, \epsilon)\right) \leq \sigma\left(B_g\left(\hat{y}, \frac{\pi}{2}\epsilon\right)\right) \leq v_p \left(\frac{\pi}{2}\epsilon\right)^p.
$$

This concludes the proof. □

Lemma I.5. *Let $\boldsymbol{K} \in \mathbb{R}^{n \times n}$ be a symmetric positive definite matrix. For any $t \geq 0$, it holds that:*

$$
\left\| \exp\left(-\frac{1}{n} t \boldsymbol{K}\right) \right\|_{\mathrm{op}} = \exp\left(-\frac{1}{n}\lambda_0 t\right),
$$

where λ_0 is the smallest eigenvalue of $\boldsymbol{K}$.

Proof. Since $\boldsymbol{K}$ is a symmetric positive definite matrix, it can be diagonalized as follows $\boldsymbol{K} = Q\Lambda Q^\top$, where Q is orthogonal (*i.e.*, $Q^{-1} = Q^\top$), $\Lambda = \mathbf{diag}(\lambda_1, \dots, \lambda_n)$, and $\lambda_1, \dots, \lambda_n > 0$ are the eigenvalues of $\boldsymbol{K}$. Then, we have:

$$\exp\left(-\frac{1}{n}t\boldsymbol{K}\right) = \exp\left(-\frac{1}{n}tQ\Lambda Q^\top\right) = \sum_{k=0}^{+\infty}\frac{(-\frac{1}{n}tQ\Lambda Q^\top)^k}{k!} = Q\left(\sum_{k=0}^{+\infty}\frac{(-\frac{1}{n}t\Lambda)^k}{k!}\right)Q^\top$$
$$= Q\exp\left(-\frac{1}{n}t\Lambda\right)Q^\top = Q\mathbf{diag}\left(\exp\left(-\frac{1}{n}\lambda_1 t\right), \dots, \exp\left(-\frac{1}{n}\lambda_n t\right)\right)Q^\top.$$

Consequently, we get:

$$\left\|\exp\left(-\frac{1}{n}t\boldsymbol{K}\right)\right\|_{\mathrm{op}} = \max_{1\le i\le n}\exp\left(-\frac{1}{n}\lambda_i t\right) = \exp\left(-\frac{1}{n}\lambda_0 t\right),$$

where $\lambda_0 \coloneqq \min(\lambda_1, \dots, \lambda_n)$ is the smallest eigenvalue of $\boldsymbol{K}$. This concludes the proof. □

Lemma I.6. *For any $m \ge 1$ and $\alpha, \beta > 0$, it holds that:*

$$\ln m \le \frac{m^\alpha}{\alpha}, \qquad \beta\ln m - m^\alpha \le \frac{\beta}{\alpha}\left(\ln\left(\frac{\beta}{\alpha}\right) - 1\right).$$

Proof. **First inequality.** Let $g(m) = \frac{m^\alpha}{\alpha} - \ln m$ for $m > 0$. We have:

$$g'(m) = m^{\alpha-1} - \frac{1}{m} = \frac{m^\alpha - 1}{m}.$$

Since $\frac{1}{m} > 0$, the sign of g' is determined by $m^\alpha - 1$: $g'(m) < 0$ for $m < 1$ and $g'(m) > 0$ for $m > 1$, so $m^\star = 1$ is a global minimum of g on $(0, +\infty)$. Since $g(1) = \frac{1}{\alpha} \ge 0$, we have $g(m) \ge \frac{1}{\alpha}$ for all $m > 0$, which gives:

$$\ln m \le \frac{m^\alpha - 1}{\alpha} \le \frac{m^\alpha}{\alpha}.$$

In particular, this holds for all $m \ge 1$.

Second inequality. Let $h(m) = \beta\ln m - m^\alpha$ for $m > 0$. We have:

$$h'(m) = \frac{\beta}{m} - \alpha m^{\alpha-1} = \frac{\beta - \alpha m^\alpha}{m}.$$

Since $\frac{1}{m} > 0$, the sign of h' is determined by $\beta - \alpha m^\alpha$: $h'(m) > 0$ for $m < m^\star$ and $h'(m) < 0$ for $m > m^\star$, where $m^\star = \left(\frac{\beta}{\alpha}\right)^{\frac{1}{\alpha}}$. Hence $m^\star$ is a global maximum of h on $(0, +\infty)$, so for all $m \ge 1$:

$$\beta\ln m - m^\alpha \le h(m^\star) = \beta\cdot\frac{1}{\alpha}\ln\frac{\beta}{\alpha} - \frac{\beta}{\alpha} = \frac{\beta}{\alpha}\left(\ln\left(\frac{\beta}{\alpha}\right) - 1\right).$$ □

Lemma I.7. *Let $p \ge 1$, $0 < \varrho \le 1$, $\tau > \frac{1}{2}$, and $\delta \in (0, 1)$. Then, it holds that $m^{p\varrho}\exp\left(-2m^{2\tau-1}\right) \le \delta$, for every $m \ge \max(1, M)$, where:*

$$M \coloneqq \left(\left[\frac{p\varrho}{2\tau-1}\left(\ln\left(\frac{p\varrho}{2\tau-1}\right) - 1\right) + \ln\left(\frac{1}{\delta}\right)\right]_+\right)^{\frac{1}{2\tau-1}} \quad \textit{with} \quad [t]_+ \coloneqq \max(t, 0).$$

Proof. Set $\alpha \coloneqq 2\tau - 1 > 0$, $\beta \coloneqq p\varrho > 0$ and $u \coloneqq \frac{\beta}{\alpha}$. Taking logarithms, the desired inequality is equivalent to:

$$p\varrho\ln m - 2m^{2\tau-1} \le \ln\delta. \tag{75}$$

By Lemma I.6, we have $\beta\ln m - m^\alpha \le u(\ln u - 1)$ for any $m \ge 1$, hence:

$$p\varrho\ln m - 2m^{2\tau-1} = (\beta\ln m - m^\alpha) - m^{2\tau-1} \le u(\ln u - 1) - m^{2\tau-1}.$$

It is therefore sufficient that:

$$m^{2\tau-1} \ge u(\ln u - 1) + \ln\left(\frac{1}{\delta}\right). \tag{76}$$

If the right-hand side of Equation (76) is nonpositive, then Equation (76) holds for every $m \ge 1$, since $m^{2\tau-1} \ge 1$. Otherwise, Equation (76) is equivalent to $m \ge M$. In both situations, $m \ge \max(1, M)$ implies Equation (76), hence Equation (75), which concludes the proof. □

Lemma I.8. *Let $n \geq 3$, $a \geq 1$, $0 < b \leq 1$, and define the function $f : (1, +\infty) \to \mathbb{R}$ by $f(x) = x^{-\frac{1}{a}}(\ln x)^b$. Then, for all $x \geq n^a(\ln n)^{ab}$, it holds that $f(x) \leq a^b(1+b)^b n^{-1}$.*

Proof. We first study the monotonicity of f. For $x > 1$, we have by the product rule:

$$f'(x) = -\frac{1}{a}x^{-\frac{1}{a}-1}(\ln x)^b + x^{-\frac{1}{a}}b(\ln x)^{b-1}\frac{1}{x} = x^{-\frac{1}{a}-1}(\ln x)^{b-1}\left(-\frac{1}{a}\ln x + b\right).$$

Since $x^{-\frac{1}{a}-1}(\ln x)^{b-1} > 0$ for $x > 1$, the sign of f' is entirely determined by $-\frac{1}{a}\ln x + b$. Hence, $f'(x) < 0$ whenever $x > e^{ab}$. Therefore, f is decreasing on $(e^{ab}, +\infty)$. Let $c = n^a(\ln n)^{ab}$ where $n \geq 3$. We have $\ln c = a\ln n + ab\ln\ln n > ab$, so $c > e^{ab}$. Thus, for all $x \geq c$, we have $f(x) \leq f(c)$. We compute:

$$\begin{aligned} f(c) &= n^{-1}(\ln n)^{-b}\left(a\ln n + ab\ln\ln n\right)^b = n^{-1}(\ln n)^{-b}a^b\left(\ln n + b\ln\ln n\right)^b \\ &\leq n^{-1}(\ln n)^{-b}a^b\left(\ln n + b(\ln n - 1)\right)^b \leq n^{-1}(\ln n)^{-b}a^b\left((1+b)\ln n\right)^b \\ &= n^{-1}a^b(1+b)^b, \end{aligned}$$

where in the second line we use $\ln\ln n \leq \ln n - 1$ for $n > 1$. Therefore, $f(x) \leq a^b(1+b)^b n^{-1}$ for all $x \geq n^a(\ln n)^{ab}$, which concludes the proof. □

Lemma I.9. *Let $C_1, C_2 > 0$. When $m \geq \max\left(\left(\frac{C_1\sqrt{12}}{C_2}\right)^{24}, 1\right)$, we have $C_1 m^{-\frac{1}{12}}\sqrt{\ln m} \leq C_2$.*

Proof. When $m \geq \left(\frac{C_1\sqrt{12}}{C_2}\right)^{24}$, we get $C_1\sqrt{12}m^{-\frac{1}{24}} \leq C_2$. By Lemma I.6, we have $\ln m \leq 12m^{\frac{1}{12}}$ when $m \geq 1$. Thus, combining both inequalities, when $m \geq \max\left(\left(\frac{C_1\sqrt{12}}{C_2}\right)^{24}, 1\right)$, we obtain $C_1 m^{-\frac{1}{12}}\sqrt{\ln m} \leq C_1 m^{-\frac{1}{12}}\sqrt{12}m^{\frac{1}{24}} = C_1\sqrt{12}m^{-\frac{1}{24}} \leq C_2$. This concludes the proof. □

Lemma I.10. *Let $0 < \alpha < \beta \leq 1$, $\delta > \frac{1}{2}$, $\gamma > \max(1-\alpha, \delta)$, $B > 0$, $d \geq 1$, and $c_0 > 0$ be an absolute constant. The following statements hold:*

(i) When $m \geq 3^{\frac{1}{\gamma - \max(1-\alpha,\delta)}}$, it holds that $m^{1-\alpha} + m^\delta \leq \lfloor m^\gamma \rfloor$.

(ii) When $m \geq \max\left(2^{\frac{1}{\beta}}, \left(\frac{128B^2d^2}{\pi(\beta-\alpha)}\right)^{\frac{1}{\beta-\alpha}}\right)$, it holds that $\frac{2Bd\sqrt{\ln m}}{\lfloor m^\beta \rfloor} \leq \frac{\sqrt{2\pi}}{4}m^{-\alpha}$.

(iii) When $m \geq \max\left(\frac{36d^2\beta^2}{c_0^2}, 1\right)$, it holds that $\frac{3d\beta}{c_0}\ln m \leq m$.

Proof. **Item (i).** Since $m \geq 1$, we have $1 \leq m^{\max(1-\alpha,\delta)}$, and therefore:

$$m^{1-\alpha} + m^\delta \leq 2m^{\max(1-\alpha,\delta)} = (2m^{\max(1-\alpha,\delta)} + 1) - 1 \leq 3m^{\max(1-\alpha,\delta)} - 1.$$

Moreover, since $m \geq 3^{\frac{1}{\gamma-\max(1-\alpha,\delta)}}$, we have $m^{\gamma-\max(1-\alpha,\delta)} \geq 3$, which gives $3m^{\max(1-\alpha,\delta)} \leq m^\gamma$. Combining with the standard inequality $\lfloor m^\gamma \rfloor \geq m^\gamma - 1$, we conclude:

$$m^{1-\alpha} + m^\delta \leq 3m^{\max(1-\alpha,\delta)} - 1 \leq m^\gamma - 1 \leq \lfloor m^\gamma \rfloor.$$

Item (ii). Since $m \geq 2^{\frac{1}{\beta}}$, we have $\frac{m^\beta}{2} \geq 1$. This implies that $\frac{m^\beta}{2} \leq m^\beta - 1 \leq \lfloor m^\beta \rfloor$. Consequently, we obtain:

$$\frac{2Bd\sqrt{\ln m}}{\lfloor m^\beta \rfloor} \leq \frac{4Bd\sqrt{\ln m}}{m^\beta} \leq \frac{4Bdm^{-\frac{\beta+\alpha}{2}}}{\sqrt{\beta-\alpha}},$$

where, in the second inequality, we use the identity $\ln m \leq \frac{m^{\beta-\alpha}}{\beta-\alpha}$ provided in Lemma I.6. So, it suffices to have:

$$\frac{4Bdm^{-\frac{\beta+\alpha}{2}}}{\sqrt{\beta-\alpha}} \leq \frac{\sqrt{2\pi}}{4}m^{-\alpha} \iff \frac{16Bd}{\sqrt{2\pi(\beta-\alpha)}} \leq m^{\frac{\beta-\alpha}{2}},$$

which is satisfied since $m \geq \left(\frac{128B^2d^2}{\pi(\beta-\alpha)}\right)^{\frac{1}{\beta-\alpha}}$.

Item (iii). By Lemma I.6, for any $m \geq 1$, we have $\ln m \leq 2\sqrt{m}$. Using this, we have:

$$\frac{3d\beta}{c_0}\ln m \leq \frac{6d\beta}{c_0}\sqrt{m}.$$

The result then follows since $m \geq \frac{36d^2\beta^2}{c_0^2}$. □

J Further Details on Experiments and Additional Numerical Results

Our code is available at https://github.com/fareselkhoury/OBO.

J.1 Empirical validation of the theory in the NTK regime on synthetic data

Statistical model. We generate synthetic data as follows:

$$x \sim \mathrm{Unif}([0,1]^2), \qquad z_j = \|x - c_j\|^{q_h} + \xi_j \ \ (j = 1, \dots, 4), \qquad y = \|x - x_0\|^{q_b} + \xi_y,$$

where $\xi_j, \xi_y \sim \mathcal{N}(0, \tau^2)$ *i.i.d.* with $\tau = 0.01$, and the centers $c_1, \dots, c_4$ and the point x_0 are as defined in the main text. x, ξ_j, and ξ_y are all redrawn independently at every sample. The feature map g_ω is given by $g_\omega(z) = \omega^\top z$, where $z := [z_1, z_2, z_3, z_4]^\top$. Since ξ_j, ξ_y are zero-mean and independent of x, we obtain $h^\star_\omega(x) = \mathbb{E}[g_\omega(z) \mid x] = \psi(x)^\top\omega$ and $b^\star(x) = \mathbb{E}[y \mid x] = \|x - x_0\|^{q_b}$, where $\psi(x) := [\|x - c_1\|^{q_h}, \dots, \|x - c_4\|^{q_h}]^\top$.

Setup. We use a width of $m = 8{,}000$ for both networks approximating $h^\star_\omega$ and $b^\star$, with input dimension $p = 2$ and outer dimension $d = 4$. Both networks are trained by full-batch gradient descent at stepsize 1, for $n_{\text{steps}} = \max(1, \mathrm{round}(t_{\text{op}}/1.0))$ steps, where t_{op} is the continuous-time stopping rule of Proposition 3.1. Because $h^\star_\omega$ and $b^\star$ are available in closed form here, we also compute the *true* hypergradient $\nabla\mathcal{F}(\omega_k)$ by a Monte Carlo estimate over 20,000 samples at every outer iteration, giving an accurate near-stationarity measure $\|\mathcal{G}_{1/\mathcal{L}}(\omega_k)\|$. We estimate $G = \sup_z \|z\|$ by the empirical maximum over 5,000 Monte Carlo samples, and use $\mathcal{L} = G^2$ exactly, since $g_\omega(z) = \omega^\top z$ is linear in ω and $\partial_\omega g_\omega(z) = z$ does not depend on ω. The projection set is $\Omega = \{\omega \in \mathbb{R}^d : \|\omega\| \leq 2\}$, the outer stepsize is $\eta = 0.05$, the outer loop runs for $K = 200$ iterations, and the hypergradient batch size is $n_g = 200$. The outer parameter is initialized as $\omega_0 \sim \mathcal{N}(0, \frac{1}{d}\mathbb{1}_{d\times d})$, so that $\mathbb{E}[\|\omega_0\|^2] = 1$ regardless of d, well inside the projection set Ω. For each smoothness level and each sample size, we repeat the estimation of $h^\star_\omega$ and $b^\star$ over 5 independent seeds and a held-out test set of 4,000 points, reporting the mean and using the standard error of the mean for the confidence intervals. The experiment is implemented using the PyTorch framework [75] and runs on a single NVIDIA RTX A5000 GPU. It takes approximately 3 hours to complete.

J.2 Insights into feature learning beyond the NTK regime

The NTK analysis is theoretically sound but often *conservative*, as it assumes fixed features and therefore precludes *feature learning*. Following [91], a single network can transition from the NTK to a feature learning regime depending on initialization scale. The bilevel structure, however, introduces an *additional degree of freedom*: the interplay between inner and outer initializations may jointly govern whether the network operates in the NTK or feature learning regime, potentially inducing feature learning even when a single network would remain in the NTK regime. This raises a natural question: *Could the bilevel structure exhibit a feature learning regime where the network adapts its features and exceeds NTK predictions?* Our experiments explore this possibility.

Statistical model. We generate synthetic data as follows:

$$x \sim \mathcal{N}\left(0, \frac{1}{p}\mathbb{1}_{p\times p}\right), \quad z = 2(\mathbb{1}_p^\top x + \varepsilon)\mathbb{1}_r + \eta, \quad y = g^\star(z) + \varepsilon,$$

where $\varepsilon \sim \mathcal{N}(0, 0.01^2)$, $\eta \sim \mathcal{N}(0, \frac{0.01^2}{4}\mathbb{1}_{r\times r})$, $g^\star(z) = \phi(z)^\top\omega^\star$, $\phi(z) = \mathrm{SiLU}(Az+b) \in \mathbb{R}^d$ with $A \in \mathbb{R}^{d\times r}$ having *i.i.d.* entries drawn from $\mathcal{N}(0, \frac{1}{r})$, $b \sim \mathcal{N}(0, 10^{-4}\mathbb{1}_{d\times d})$, and $\omega^\star \sim \mathcal{N}(0, \frac{1}{d}\mathbb{1}_{d\times d})$, all drawn once at initialization and fixed throughout.

Setup. We consider the empirical inner and outer objectives defined in Equation (3) with $g_\omega : z \mapsto \phi(z)^\top \omega$, where $\phi(z) = \mathrm{SiLU}(Az + b) \in \mathbb{R}^d$ is a fixed random feature map, $\mathrm{SiLU}(x) = x/(1 + \exp(-x))$ is applied entrywise, $A \in \mathbb{R}^{d\times r}$ and $b \in \mathbb{R}^d$ are drawn at initialization. We generate synthetic data with 100 training samples and 20 validation and test samples at each level, that remain fixed across all runs. Both $h^\star_\omega$ and $b^\star$ are approximated by neural networks h^{NN}_ω and b^{NN} of width 500, of the form in Equation (5) without second-layer bias, normalized by m rather than $\sqrt{m}$ to be consistent with the feature learning limit as m increases [26]. The first-layer parameters are initialized from a centered Gaussian of variance σ_0^2 for h^{NN}_ω and τ_0^2 for b^{NN}, while the second-layer weights are initialized to 0. Since b^{NN} is independent of the outer variable, it is trained once at the beginning; h^{NN}_ω is retrained at every outer iteration with a warm-start strategy [90]. Both networks are trained with full-batch gradient descent until the training error falls below 10^{-15}. Early stopping is applied based on validation loss, selecting the iterate with the lowest validation error. At the outer level, we use full-batch gradient descent for 5000 iterations, which is appropriate here because the sample size is small. We use the PyTorch framework [75] on an NVIDIA RTX 6000 ADA GPU. The experiments take around 10 hours to complete for each pair (σ_0^2, τ_0^2). We set $d = 30$, $p = 20$, and $r = 10$. The stepsize to train the two neural networks is fixed at 10, whereas the stepsize for the outer optimization is 0.01. We vary the initialization variances σ_0^2 and τ_0^2 in $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100\}$ independently, across training sizes and random seeds, to study the transition between the NTK and feature learning regimes.

Additional numerical results. Figure 3 shows the test outer loss over outer iterations for all combinations of (σ_0^2, τ_0^2). When both initializations are small, the network operates in the feature learning regime and converges to a low test loss. As either σ_0^2 or τ_0^2 increases, the network transitions toward the NTK regime, with curves plateauing at higher loss values. This is most visible in the bottom-right panels, where both initializations are large and all curves converge to similarly high loss regardless of σ_0^2, suggesting that large τ_0^2 alone is sufficient to push the network into the NTK regime.

Figure 4 shows the final test outer loss as a function of the sample size for all combinations of (σ_0^2, τ_0^2). Three observations emerge. First, all curves decrease as the sample size grows, reflecting improved generalization with more data regardless of the regime. Second, the ranking between regimes is preserved at every sample size: when both initializations are small (feature learning), the network consistently achieves lower loss than when either initialization is large (NTK), confirming that the regime effect observed in Figure 2 is not a finite-sample artifact. Third, when τ_0^2 is large, the curves collapse and become insensitive to σ_0^2, further confirming that large τ_0^2 alone is sufficient to push the network into the NTK regime, regardless of σ_0^2.

Figure 5 presents the same results as in Figure 2 in line plot form, showing the test inner and outer losses at the last outer iteration as a function of σ_0^2 for varying τ_0^2. The confidence intervals confirm that the ranking between regimes is stable across runs.

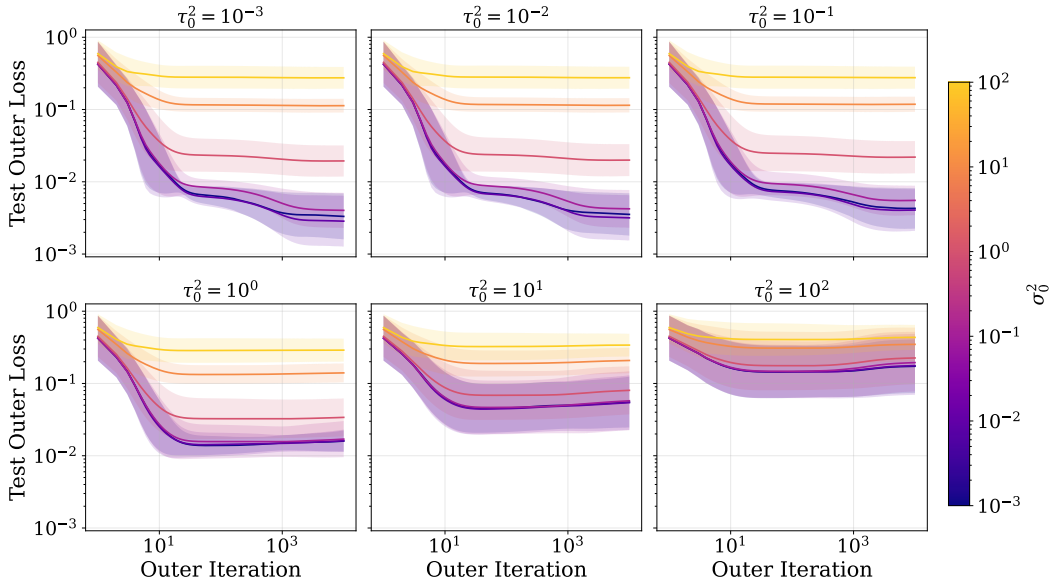


Figure 3: Test outer loss as a function of outer iterations for varying σ_0^2 and τ_0^2: small initializations (feature learning) converge to lower loss, while large initializations (NTK) plateau at higher values. The plots are averaged over 5 runs and displayed on a log scale. The line represents the mean across all runs, and the shaded region indicates the 95% confidence interval.

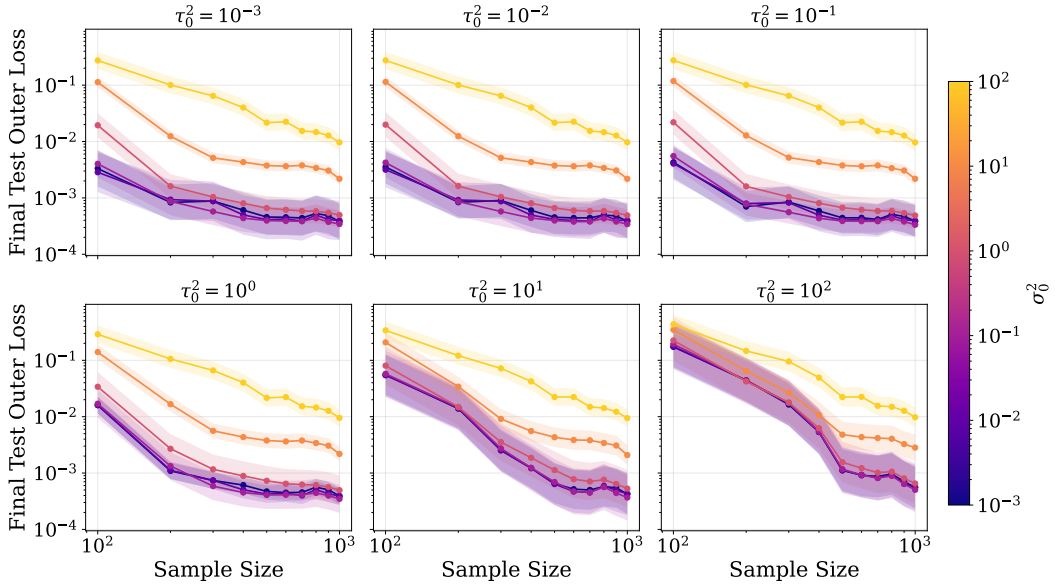


Figure 4: Test outer loss at the last iteration as a function of the sample size for varying σ_0^2 and τ_0^2: all curves decrease with the sample size, with the ranking between feature learning and NTK regimes preserved throughout. The plots are averaged over 5 runs and displayed on a log scale. The line represents the mean across all runs, and the shaded region indicates the 95% confidence interval.

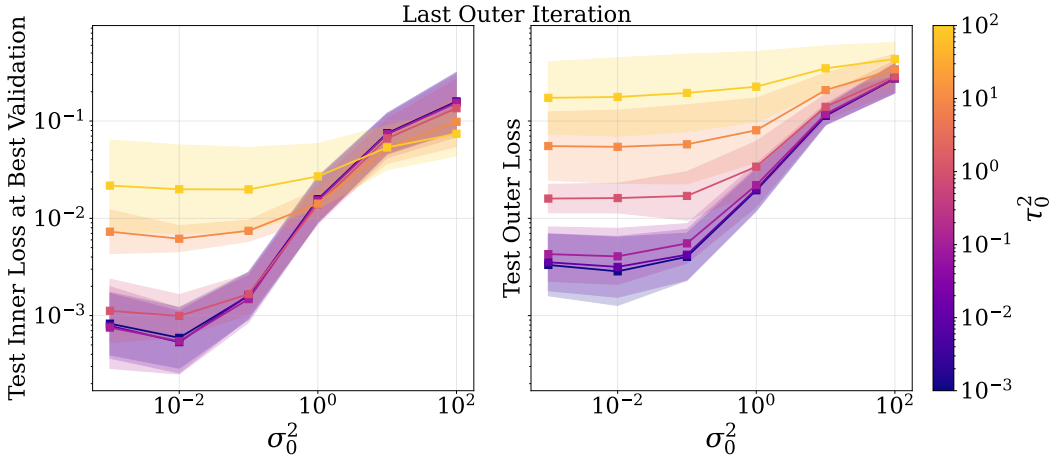


Figure 5: Test inner (left) and outer (right) losses at the last outer iteration as a function of σ_0^2 for varying τ_0^2: both losses are lowest when both initializations are small (feature learning regime) and degrade as either grows large (NTK regime). The plots are averaged over 5 runs and displayed on a log scale. The line represents the mean across all runs, and the shaded region indicates the 95% confidence interval.