# QUADRATIC WEAK-TO-STRONG GENERALIZATION IN RANDOM FEATURE NETWORKS VIA RANDOM MATRIX THEORY

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Weak-to-strong generalization is the phenomenon where a strong student model trained with labels produced by a weak teacher model is able to generalize better than the teacher. In this paper, we study this phenomenon in two-layer random feature networks where the model strength is determined by its width. Using tools from random matrix theory, we derive deterministic equivalents for the population errors of an optimally trained teacher and a student trained with gradient flow. For ReLU activation and a pure spherical harmonic target, we obtain sharp asymptotics under a Gaussian universality assumption, showing a quadratic improvement: the student error scales as the square of the teacher error. These results attain the general lower bound of Medvedev et al. (2025). We also analyze how the student behaves under more general stopping times and targets supported on multiple harmonic degrees, characterizing the regimes in which weak-to-strong generalization occurs and identifying the transition between quadratic, non-quadratic, and no improvement.

## 1 INTRODUCTION

Can a student trained only on a teacher's predictions surpass the teacher? A small model trained on the true labels still makes errors because of its limited capacity, and a larger student that sees only these flawed labels might be expected to reproduce them. *Weak-to-strong generalization* (WTS), introduced in Burns et al. (2024), is the phenomenon where it does not: the strong student generalizes better than the weak teacher that supervises it.

We study this phenomenon theoretically in a random feature neural network, in a regression problem with squared loss, where the strength of a model is its capacity, that is, its number of neurons. We choose this model for three reasons. Its strength has a single unambiguous meaning, the width. Its population loss is governed by the spectrum of the feature kernel, so the teacher's error and the student's training dynamics can be written explicitly, and, for data on the sphere, the expansion of the target in spherical harmonics gives a natural notion of its complexity, its harmonic degree. Finally, random feature models, introduced in Rahimi & Recht (2007) as a way of reducing computation in kernel methods, are conceptually simple yet exhibit phenomena also present in deep networks, and are a standard surrogate for them (Mei & Montanari, 2022; Bombari & Mondelli, 2024; Dohmatob & Bietti, 2022; Chizat et al., 2019). Our aim is to show that WTS emerges already in this model, and to quantify the contribution of each of its components.

We closely follow the setup pioneered in Medvedev et al. (2025), where random feature networks also represent the student and teacher. Our two-step training procedure is the same as theirs: we first optimally train the teacher over the population error, so that its only source of error is its capacity, and then train the student, initialized at zero, on the teacher's labels using early-stopped gradient flow. The goal is to stop the training before the student imitates too many of the teacher's mistakes, which is an implicit regularization effect (Li et al., 2020; Yao et al., 2007; Raskutti et al., 2014; Ali et al., 2019; Heckel & Yilmaz, 2021). In our specific case, it can be understood as a spectral cutoff: the student imitates the teacher's components up to the target degree, while the remaining components barely move from their initialization.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We differ from Medvedev et al. (2025) both in method and in scope. In method, we use random matrix theory, specifically the theory of Dyson equations and deterministic equivalents (Erdős, 2019). A deterministic equivalent removes the randomness of the features from the population errors and gives explicit formulas for both student and teacher errors: a single-resolvent formula for the teacher and a two-resolvent formula for the student, the latter being a deterministic equivalent for a two-stage training pipeline. This simplified mechanism reproduces the errors accurately and makes their analysis straightforward. In scope, the formulas hold at every stopping time and for general targets, which lets us analyze the optimal improvement and the full effect of the stopping time.

We leave open the comparison between the deterministic equivalents and the random feature regression itself. Results of this kind exist in the single-stage setting (Defilippis et al., 2024; Cheng & Montanari, 2024; Misiakiewicz & Saeed, 2024); extending them to the two-stage pipeline is a matter of mathematical technicalities that we do not expect to reveal new information about the learning problem, and our numerical comparisons show that the deterministic equivalents match the gradient-flow simulations closely.

**Our contributions.**

- **Deterministic equivalents and a general lower bound.** We derive deterministic equivalents for the population errors of the optimally trained teacher and of the early-stopped student (Section 3, Appendix B). The student error splits into a squared bias and a variance term. The decomposition explains both the improvement over the teacher and its loss under prolonged training, and implies that the student error is at least the square of the teacher error, for any activation and any target (Theorem 1).
- **Quadratic improvement.** For ReLU activation on the sphere and a pure spherical-harmonic target, we show that an appropriate stopping time attains this quadratic improvement (Theorem 2, Section 4), attaining the lower bound of Theorem 1. The ReLU statements are conditional on a Gaussian universality assumption (Assumption 1).
- **Pure harmonic targets and general stopping times.** We characterize the phase diagram of the student error in terms of the teacher's capacity exponent $\alpha$ ($m = d^\alpha$ neurons) and the stopping time exponent $\gamma$ ($T = \Theta(d^{\gamma+1})$), identifying which of the two error terms controls the error in each region and the transitions between quadratic, nonquadratic, and no improvement (Theorem 4, Section 5).
- **Mixed targets and general stopping times.** For targets with components on several harmonic degrees, we give a condition for WTS in terms of unlearned target energy, and a coefficient-independent guarantee of polynomial improvement when the teacher can learn every mode. Two examples show that the improvement can nevertheless fail to be quadratic, because different modes require incompatible stopping times (Section 6).

**Related work.** Since the first empirical investigation of Burns et al. (2024), many works have considered a theoretical analysis using simple settings. Some focus on how data structure influences the phenomenon. Lang et al. (2024) study the student's correction of teacher mistakes and its performance on regions not covered by teacher labels. Shin et al. (2025) express improvement through the frequency of points containing both easy and hard patterns, while Wu & Sahai (2025) study a transition from random guessing to perfect generalization in classification.

Other works quantify the improvement through the *misfit error*, which measures the disagreement between the student and teacher (Charikar et al., 2024; Yao et al., 2025; Mulgund & Pabbaraju, 2025). From the point of view of representations, Xue et al. (2025) identify directions poorly captured by the teacher but well represented by the student. Dong et al. (2025) analyze how the student inherits teacher variance along shared directions and reduces it along others. In these works, differences between the two representation spaces prevent the student from fully inheriting the teacher's noise, which is the source of the improvement.

With respect to regularization, Moniri & Hassani (2026) show that a student in linear ridge regression can surpass the teacher through better regularization, especially when it is aligned with the target. This is related to Medvedev et al. (2025) and our work, where regularization is imposed implicitly by early-stopped gradient flow. Beyond linear settings, Oh et al. (2025) use convolutional networks and also find that early stopping improves student performance in a data-abundant regime.

Closest to our setting are analyses of weak-to-strong generalization in random feature and high-dimensional linear regression, where teacher and student differ in width or in regularization and the labels are produced by a fitted teacher. Ildiz et al. (2025) analyze ridge regression trained on the predictions of a surrogate model in the proportional regime, characterizing when the student improves on the teacher and deriving scaling laws, and Wu et al. (2026) show that weak-to-strong training improves the scaling law of random feature ridge regression. In both works, as in ours, deterministic equivalents play a central role. Appendix A gives an extended discussion.

**Notation.** Throughout, $d$ is the dimension of the data and all asymptotic statements are as $d \to \infty$. We write $a \asymp b$ for bounds in both directions by positive constants independent of $d$, $a \sim b$ for $a/b \to 1$, and $\widetilde{\Theta}$ for order up to polylogarithmic factors.

## 2 Problem setup

For both the weak teacher and the strong student, data vectors $x \in \mathbb{R}^d$ come from a distribution $\mathcal{D}$, and labels are in $\mathbb{R}$. The predictor is a random feature model, which can be seen as a two-layer neural network with a fixed bottom random layer:

$$f_{u,w}(x) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} w_i \sigma(\langle u_i, x \rangle), \qquad u_i \overset{\text{iid}}{\sim} \mathcal{U}.$$

Here $m$ is the number of neurons, $\sigma : \mathbb{R} \to \mathbb{R}$ is the activation function, the feature weights $u_i \in \mathbb{R}^d$ are fixed, and only the output weights $w_i \in \mathbb{R}$ are trained. The $1/\sqrt{m}$ normalization makes the kernel of the network, $\frac{1}{m}\sum_{i=1}^{m} \sigma(\langle u_i, x\rangle)\sigma(\langle u_i, x'\rangle)$, an average over the features, so that it converges to the population kernel $K$ below as $m \to \infty$. We will take both $\mathcal{D}$ and $\mathcal{U}$ uniform on the sphere $\mathbb{S}^{d-1}$, although the machinery of the theory would work for many other choices. Apart from their training procedures, the only difference between teacher and student is their width, which is our notion of model strength: the teacher has $M_{\mathrm{TE}}$ neurons, and the student is taken infinitely wide, $M_{\mathrm{ST}} \to \infty$. We expect that all results here would hold for a finite student as well, for student sizes which are sufficiently large to learn $f^\star$, but leave a quantitative analysis to future work.

Both student and teacher are evaluated against the ground-truth target $f^\star$ using population squared loss, which we write using the norm $\|f\|_{\mathcal{D}}^2 = \mathbb{E}_{x\sim\mathcal{D}} f(x)^2$, as

$$\mathcal{L}_{\mathrm{TE}} = \|f_{\mathrm{teacher}} - f^\star\|_{\mathcal{D}}^2, \qquad \mathcal{L}_{\mathrm{ST}}(t) = \|f_{\mathrm{student}}^{(t)} - f^\star\|_{\mathcal{D}}^2.$$

Weak-to-strong generalization means that the student error is smaller than the teacher error. However, in this paper we use WTS to refer to the stronger statement that $\mathcal{L}_{\mathrm{ST}}(t)/\mathcal{L}_{\mathrm{TE}}$ decays as a power of $d$.

**Teacher.** The teacher is optimally trained on the true labels: its output weights minimize the population loss $\mathcal{L}_{\mathrm{TE}}$ over $w \in \mathbb{R}^{M_{\mathrm{TE}}}$. Its only source of error is therefore its finite width. The optimal weights and the resulting error are computed in Appendix B, Lemma 1.

**Student.** The student receives the teacher's predictions as labels, starts from $w_{\mathrm{s}} = 0$, and is trained by the gradient flow,

$$\frac{dw_{s,j}}{dt} = -\frac{1}{2}\frac{\partial}{\partial w_{s,j}} \mathbb{E}_{x\sim\mathcal{D}} (f_{\mathrm{student}}^{(t)}(x) - f_{\mathrm{teacher}}(x))^2, \qquad j = 1, \dots, M_{\mathrm{ST}},$$

stopped at a time $t$ chosen below. As $M_{\mathrm{ST}} \to \infty$ this converges to the population spectral flow

$$f_{\mathrm{student}}^{(t)}(x) = \sum_i (1 - e^{-\psi_i t}) \langle f_{\mathrm{teacher}}, e_i \rangle_{\mathcal{D}} e_i(x), \tag{2.1}$$

where $(\psi_i, e_i)$ are the eigenpairs in $L^2(\mathcal{D})$ of the kernel $K(x, x') = \mathbb{E}_{u\sim\mathcal{U}} \sigma(\langle u, x\rangle)\sigma(\langle u, x'\rangle)$, the average of the student network kernel over the features (see Appendix A.1 of Medvedev et al., 2025).

We shall order the eigenvalues to be decreasing $\psi_1 \geq \psi_2 \geq \cdots$. Each feature $\sigma(\langle u_i, \cdot\rangle)$ has coordinates $\psi_j^{1/2} e_j(u_i)$ in this basis. We let $\Psi$ be the covariance of the feature coordinates $\psi_j^{1/2} e_j(u_i)$ over $u_i \sim \mathcal{U}$, so that $\Psi = \mathrm{diag}(\psi_1, \psi_2, \dots)$. The target is $f^\star = \sum_j \beta_j \psi_j^{1/2} \phi_j$, so that $\Psi^{1/2}\beta$ is its coefficient vector in the orthonormal basis and $\|f^\star\|_{\mathcal{D}}^2 = \langle \beta, \Psi\beta \rangle$, which we normalize to one.

## 3 DETERMINISTIC EQUIVALENTS

In the random features setting we have two sources of randomness: the data distribution and the random features distribution. A deterministic equivalent allows us to find a system of fixed-point equations, around whose solution the student and teacher test errors concentrate, over both sources of randomness.

**Teacher.** The deterministic equivalent of the teacher error is governed by a single scalar $\nu > 0$, defined as the unique solution of

$$\operatorname{tr}\big[\Psi(\Psi + \nu I)^{-1}\big] = m. \tag{3.1}$$

Since each summand $\psi_j/(\psi_j + \nu)$ is close to one for $\psi_j \gg \nu$ and close to zero for $\psi_j \ll \nu$, the equation says that the number of eigenvalues of $\Psi$ above $\nu$ is about $m$: $\nu$ is the eigenvalue scale at the edge of the teacher's capacity. It plays the role of an effective ridge parameter, since the teacher error has the form of a ridge regression error on the full feature space with ridge $\nu$ (Appendix B.3, Theorem 5):

$$\overline{\mathcal{L}}_{\mathrm{TE}} = \nu\langle \Psi^{1/2}\beta, (\Psi + \nu I)^{-1}\Psi^{1/2}\beta\rangle. \tag{3.2}$$

The factor $\nu/(\psi_j + \nu)$ multiplying the energy of the target along $\phi_j$ is small on the directions the teacher can represent ($\psi_j \gg \nu$) and close to one on those it cannot.

**Student.** Substituting the optimal teacher into the spectral flow (2.1) shows that the student error depends on the teacher's random features through two projectors instead of one (Appendix B.4). Its deterministic equivalent therefore requires a double-resolvent calculation, Theorem 7. Regrouping the resulting terms (Appendix B.5) gives the two-term expression that we take as the definition of the deterministic equivalent of the student error: a squared bias $\overline{B}(t)$ and a variance $\overline{V}(t)$,

$$\begin{aligned}\overline{\mathcal{L}}_{\mathrm{ST}}(t) = &\underbrace{\left\|(\nu I + \Psi e^{-\Psi t})(\Psi + \nu I)^{-1}\Psi^{1/2}\beta\right\|^2}_{\overline{B}(t)} \\ &\underbrace{+\,\nu\langle \Psi^{1/2}\beta, \Psi(\Psi + \nu I)^{-2}\Psi^{1/2}\beta\rangle \frac{\operatorname{tr}[\Psi(\Psi + \nu I)^{-2}(I - e^{-\Psi t})^2]}{\operatorname{tr}[\Psi(\Psi + \nu I)^{-2}]}}_{\overline{V}(t)}.\end{aligned} \tag{3.3}$$

At $t = 0$ this equals one, the norm of the target, and at $t = \infty$ it equals $\overline{\mathcal{L}}_{\mathrm{TE}}$: without early stopping the student simply reproduces the teacher. Appendix B.5 verifies (3.3) algebraically.

The deterministic equivalent is obtained from the matrix Dyson equation (Erdős, 2019). The derivation of the deterministic equivalent is for Gaussian feature coordinates, using Stein's lemma and removing lower order terms, which would vanish by concentration. Hence, the deterministic equivalent for the ReLU model simultaneously reflects two approximations: first a concentration property and second a universality property. See B.1 for a general introduction and (B.4) for a discussion of the universality.

We shall formulate all our theorems for the deterministic equivalent, and we expect that these transfer exactly to the random model as we have strong empirical evidence (see Figure 1) and theoretical support (see discussion at the end of Appendix B.1). Precisely, we anticipate the following holds:

**Assumption 1** (Gaussian universality and deterministic equivalence)**.** Let $\overline{\mathcal{L}}_{\mathrm{TE}}$ and $\overline{\mathcal{L}}_{\mathrm{ST}}(t)$ be the deterministic equivalents defined in (3.2) and (3.3) below, and let $(T_d)$ be any of the stopping-time sequences appearing in our statements. The random population losses of the ReLU model on the sphere satisfy, in probability as $d \to \infty$,

$$\mathcal{L}_{\mathrm{TE}}/\overline{\mathcal{L}}_{\mathrm{TE}} \to 1, \qquad \mathcal{L}_{\mathrm{ST}}(T_d)/\overline{\mathcal{L}}_{\mathrm{ST}}(T_d) \to 1.$$

For statements uniform over stopping times or target families, the corresponding relative approximation is assumed uniform.

### 3.1 THE BIAS–VARIANCE MECHANISM

Averaging over the teacher features $\mathcal{U}$ gives the exact identity

$$\mathbb{E}_{\mathcal{U}}\mathcal{L}_{\mathrm{ST}}(t) = \|\mathbb{E}_{\mathcal{U}} f^{(t)}_{\mathrm{student}} - f^\star\|^2_{\mathcal{D}} + \mathbb{E}_{\mathcal{U}}\|f^{(t)}_{\mathrm{student}} - \mathbb{E}_{\mathcal{U}} f^{(t)}_{\mathrm{student}}\|^2_{\mathcal{D}}.$$

The deterministic equivalent of the first term (the squared mean residual) is $\overline{B}$, and that of the second term (the fluctuation of the student around its mean) is $\overline{V}$ (see Appendix B.5 for details; see also Appendices C and D, where $\overline{V}$ is term 4)(2) and the rest sum to $\overline{B}$).

The bias can be seen to decrease in time, while the variance increases. Therefore, provided $\overline{B}(\infty) < \overline{V}(\infty)$ and $\overline{B}(0) > \overline{V}(0)$, there exists an optimal finite stopping time at which they are equal. As this can be substantially smaller than the sum of these losses at time $\infty$, which is equal to $\overline{\mathcal{L}}_{\mathrm{TE}}$, this shows how the student trained with early stopped gradient flow can outperform the teacher.

Further, we can give a general lower bound, using this representation, which will be sharp in some instances:

**Theorem 1** (General quadratic lower bound)**.** For any activation, any feature distribution, any data distribution, and any unit target for which (3.1)–(3.3) are defined, every stopping time $t$ satisfies

$$\overline{\mathcal{L}}_{\mathrm{ST}}(t) \geq \nu^2 \langle \Psi^{1/2}\beta, (\Psi + \nu I)^{-2}\Psi^{1/2}\beta\rangle \geq \overline{\mathcal{L}}^2_{\mathrm{TE}}.$$

The first inequality drops the nonnegative variance and the exponential in the bias. The second is Cauchy–Schwarz, using $\|\Psi^{1/2}\beta\| = 1$. Thus quadratic improvement is the best possible order for the deterministic equivalent, for arbitrary targets. This is the deterministic-equivalent counterpart of the lower bound of Medvedev et al. (2025), who prove the analogous result for the random student error.

## 4 OPTIMAL QUADRATIC IMPROVEMENT

**The ReLU model on the sphere.** We take $\sigma(r) = r_+$ and recall we use the uniform probability measure on $\mathbb{S}^{d-1}$ for both the data and the fixed weight layer, as in Medvedev et al. (2025). The population kernel then expands in an orthonormal spherical-harmonic basis $\phi_{k,l}$, $l = 1, \ldots, N_k$, where $N_k$ is the dimension of the space of degree-$k$ harmonics and $\sigma_k^2$ is the eigenvalue of $K$ on that space, so that $\sigma_k$ is the degree-$k$ coefficient of ReLU in the spherical-harmonic expansion; it vanishes for odd $k \geq 3$. The kernel is

$$K(x, x') = \sum_{k\in\mathcal{K}} \sigma_k^2 \sum_{l=1}^{N_k} \phi_{k,l}(x)\phi_{k,l}(x'), \qquad \mathcal{K} = \{0, 1\} \cup 2\mathbb{N}. \tag{4.1}$$

In the notation of Section 3, $\Psi$ has blocks $\sigma_k^2 I_{N_k}$; consequently traces become sums over degrees with multiplicities $N_k$, and, by the spectral flow (2.1), a target of degree $k$ is learned by the student on the time scale $1/\sigma_k^2$. For fixed $k$, $N_k \asymp d^k$ and $\sigma_k^2 \asymp d^{-k-1}$. Appendix C.1 gives their expressions and asymptotics.

Our first result is as follows.

**Theorem 2** (Quadratic improvement: pure harmonic target)**.** Let $f^\star$ be a unit spherical harmonic of fixed degree $k_0 \in \mathcal{K}$, $k_0 \geq 1$, and let $m = d^\alpha$ with fixed $\alpha > 0$. Then, with

$$T = \frac{\log m + \log \overline{\mathcal{L}}^{-1}_{\mathrm{TE}}}{\sigma^2_{k_0}}, \qquad \overline{\mathcal{L}}_{\mathrm{TE}} \asymp d^{-(\alpha-k_0)_+}, \qquad \overline{\mathcal{L}}_{\mathrm{ST}}(T) \asymp \overline{\mathcal{L}}^2_{\mathrm{TE}}. \tag{4.2}$$

The proof is in Appendix C. Together with Theorem 1, this establishes optimal quadratic improvement in the ReLU setting once $\alpha > k_0$.

### 4.1 THE ROLE OF THE ACTIVATION

We briefly investigate how the choice of activation affects Theorem 2. The short answer is that only the set $\mathcal{K}$ of degrees present in the kernel changes. Let $\sigma : \mathbb{R} \to \mathbb{R}$ be measurable with

$|\sigma(r)| \le Ce^{C|r|}$, and consider the features $\sigma(\sqrt{d}\,\langle u, x\rangle)$, so that the preactivation is of order one (for ReLU this is a constant multiple of the features above). With $z \sim \mathcal{N}(0,1)$ and $\mathrm{He}_k$ the probabilists' Hermite polynomials, set

$$c_k = \frac{\mathbb{E}[\sigma(z)\mathrm{He}_k(z)]}{\sqrt{k!}}, \qquad \mathcal{K}_\sigma = \{k \ge 0 : \ c_k \ne 0\}.$$

For ReLU, $\mathcal{K}_\sigma = \mathcal{K} = \{0,1\} \cup 2\mathbb{N}$; for $\tanh$, $\mathcal{K}_\sigma$ is the odd integers; for the logistic sigmoid it is $\{0\}$ together with the odd integers, and a generic bias makes it all of $\mathbb{N}$ (Appendix F, Examples 1–3).

**Theorem 3** (Quadratic improvement: general activations)**.** Let $\sigma$ be as above, let $f^\star$ be a unit spherical harmonic of fixed degree $k_0 \in \mathcal{K}_\sigma$, $k_0 \ge 1$, and let $m = d^\alpha$ with fixed $\alpha > 0$. Assume $c_{k_\star} \ne 0$ for some $k_\star > \alpha$ (which holds for every $\alpha$ if $\sigma$ is not a polynomial). Then, with the stopping time $T$ of (4.2),

$$\overline{\mathcal{L}}_{\mathrm{TE}} \asymp d^{-(\alpha - k_0)_+}, \qquad \overline{\mathcal{L}}_{\mathrm{ST}}(T) \asymp \overline{\mathcal{L}}_{\mathrm{TE}}^2.$$

The proof is in Appendix F. The kernel again has the form (4.1), with $\sigma_k$ the degree-$k$ spherical-harmonic coefficient of $\sigma(\sqrt{d}\,\cdot)$, and $N_k\sigma_k^2 \to c_k^2$ for each fixed $k$ (Lemma 7), so $\sigma_k^2 \asymp d^{-k}$ for $k \in \mathcal{K}_\sigma$, as for ReLU. The proof of Theorem 2 uses the $\sigma_k$ only through these orders at finitely many degrees and the bound $\mathrm{tr}\,\Psi \lesssim 1$, so it goes through unchanged.

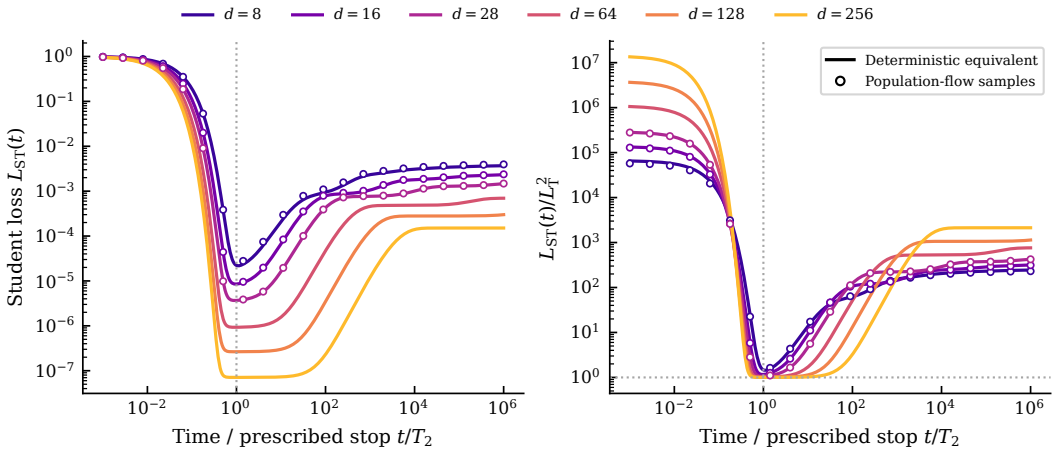



Figure 1: **Quadratic improvement and the stopping window.** Unit degree-two target, $m = d^3$, with time divided by $T_2 = 4\log(d)/\sigma_2^2$, the stopping time $T$ of Theorem 2 for $k_0 = 2$ and $m = d^3$. Left: student loss; right: loss divided by squared teacher loss. Lines are the deterministic equivalent; open markers (population-flow samples) are population spherical-teacher observations at $d = 8, 16, 28$ (32, 32, and 8 draws); see Appendix G for details. See also uncertainty and numerical approximation quantification in Appendix G.2. The dotted line marks $T_2$.

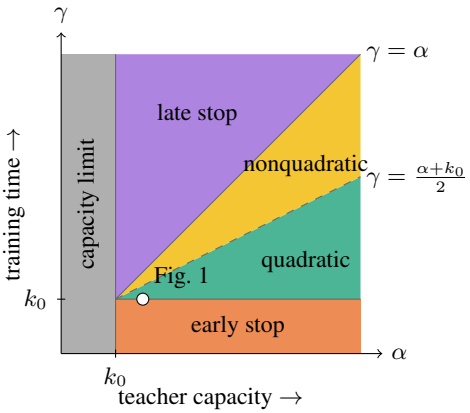
## 5 Stopping-time sensitivity and the phase diagram

To examine the effect of the stopping time in the WTS phenomenon, we vary the harmonic degree at which the student is stopped. For a $\gamma \in \mathcal{K}$ with $\gamma > 0$, set

$$T_\gamma = \frac{\log d}{\sigma_\gamma^2} \begin{cases} \alpha, & \alpha \le k_0, \\ 2\alpha - k_0, & \alpha > k_0. \end{cases}$$

This is the stopping time (4.2) of Theorem 2 with the target degree $k_0$ replaced by $\gamma$, the logarithm evaluated using $m = d^\alpha$, and the teacher error substituted by its leading asymptotic; $T_{k_0}$ recovers $T$. Thus $T_\gamma \asymp d^{\gamma+1}\log d$, the time at which the student has learned the teacher's components up to degree $\gamma$ and not beyond.

**Theorem 4** (Stopping-time phases)**.** In the pure-target setting of Theorem 2, for every stopping degree $\gamma \in \mathcal{K}$, the leading order of $\overline{\mathcal{L}}_{\mathrm{ST}}(T_\gamma)$ are as given in Figure 2.




| Phase | Region | $\overline{\mathcal{L}}_{\mathrm{ST}}(T_\gamma)$ | $\overline{B}/\overline{V}$ |
|---|---|---|---|
| capacity limit | $\alpha \le k_0$ | $\asymp 1$ | $\overline{B}$ |
| early stop | $\gamma < k_0$ | $\widetilde{\Theta}(1)$ | $\overline{B}$ |
| quadratic | $k_0 \le \gamma \le \frac{\alpha+k_0}{2}$ | $\widetilde{\Theta}(\overline{\mathcal{L}}_{\mathrm{TE}}^{\,2})$ | $\overline{B}$ |
| nonquadratic | $\frac{\alpha+k_0}{2} < \gamma < \alpha$ | $\widetilde{\Theta}(\overline{\mathcal{L}}_{\mathrm{TE}}^{\,\rho})$ | $\overline{V}$ |
| late stop | $\gamma \ge \alpha$ | $\widetilde{\Theta}(\overline{\mathcal{L}}_{\mathrm{TE}})$ | $\overline{V}$ |

Figure 2: **Stopping-time phases (Theorem 4).** Left: the five regions in the plane of teacher capacity exponent $\alpha$ and stopping degree $\gamma$ (drawn for $k_0 = 2$; admissible degrees form the discrete set $\mathcal{K}$). The marked point is the setting of Figure 1. Right: the phases, with the last four for $\alpha > k_0$, where $\overline{\mathcal{L}}_{\mathrm{TE}} \asymp d^{-(\alpha-k_0)}$, and $\rho = (3\alpha - k_0 - 2\gamma)/(\alpha - k_0) \in (1, 2)$. The $\overline{B}/\overline{V}$ indicates which error terms dominates the bias/variance decomposition. Details are in Appendix D.1.

The five phases of Figure 2 follow from comparing $\overline{B}(T_\gamma)$ and $\overline{V}(T_\gamma)$ (Appendix D.1, cf. the summary (D.11)). ■ If $\alpha \le k_0$ the teacher cannot represent the target, its residual dominates the bias, and both errors are of order one whatever the stop. ■ Before the target scale ($\gamma < k_0$), the bias is dominated by the unlearned target and leaves order-one student error even if the teacher error vanishes. ■ After the target scale and before the quadratic boundary, $\overline{B}$ reaches its squared-teacher floor $\overline{\mathcal{L}}_{\mathrm{TE}}^{\,2}$ while $\overline{V}$ is smaller; on the boundary itself the two have the same polynomial order and a factor $(\log d)^2$ can appear. ■ Further training leads the student to learn the variance of the teacher, leading to intermediate improvement exponents $\rho \in (1, 2)$. ■ At late harmonic stops ($\gamma \ge \alpha$) the variance has teacher-error order: the student has learned the degrees near the teacher's capacity edge, where the teacher's residual lives, and copies it.

## 6 Mixed targets: common guarantees and incompatible times

Since different spectral components of the target are learned on different timescales, for a target with several modes it may be impossible to choose a single stopping time at which all of them are learned optimally. However, if the teacher has sufficient capacity to learn all spectral components of the target, weak-to-strong generalization is possible for suitable stopping times.

Consider a general harmonic target with maximum degree $k_0$, and write $f_k^\star$ for its degree-$k$ component, so that $\|f^\star\|_{\mathcal{D}}^2 = \sum_{k \le k_0} \|f_k^\star\|_{\mathcal{D}}^2$. Figure 3 illustrates how $k_0$, the teacher capacity exponent $\alpha$ and the stopping-time exponent $\gamma$ give rise to different regimes of the WTS phenomenon. The regimes are established in full detail in Appendix D.2.

**Guaranteed weak-to-strong generalization.** ■ If the teacher has enough capacity to learn every mode of the target, $\alpha > k_0$, and the student is stopped after the target degree but before the teacher's capacity edge, $k_0 \le \gamma < \alpha$, then

$$\overline{\mathcal{L}}_{\mathrm{ST}}(T_\gamma) \lesssim d^{k_0-\alpha}\, \overline{\mathcal{L}}_{\mathrm{TE}} = o(\overline{\mathcal{L}}_{\mathrm{TE}}) \tag{6.1}$$

(Appendix D.2, Regime 4). This is weak-to-strong generalization for every mixed target, whatever its coefficients. The bound is sharp: a pure harmonic of degree $k_0$ attains it, since by Theorem 4 its student error at these stops is $\overline{\mathcal{L}}_{\mathrm{TE}}^{\,2} \asymp d^{k_0-\alpha}\overline{\mathcal{L}}_{\mathrm{TE}}$, so no better uniform guarantee is possible. The conceptual point is that although the modes of a mixed target are learned on different timescales, the conflict between them does not threaten polynomial improvement: as long as the teacher can learn the target, any stop in the window $[k_0, \alpha)$ delivers the full effect. The window ends at $\alpha$ for the same reason as in Theorem 4. The teacher's residual is concentrated on the degrees near its capacity edge,

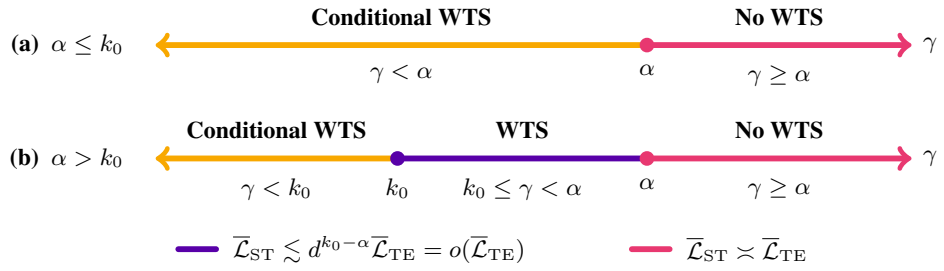



Figure 3: Weak-to-strong generalization regimes as a function of the stopping time exponent $\gamma$, for (a) $\alpha \le k_0$ and (b) $\alpha > k_0$. ■ guaranteed WTS, for every target; ■ conditional WTS, governed by (6.3); ■ no WTS. The ■ interval includes $\gamma = k_0$ and excludes $\gamma = \alpha$.

where $\sigma_k^2 \asymp \nu$, and once $\gamma \ge \alpha$ the student has learned those degrees and copies the residual, so $\overline{\mathcal{L}}_{\mathrm{ST}}(T_\gamma) \asymp \overline{\mathcal{L}}_{\mathrm{TE}}$ for every target (the ■ intervals of Figure 3).

**Conditional weak-to-strong generalization.** In the ■ intervals of Figure 3, the stop $\gamma < \alpha$ comes before some degree carrying target energy. Define the held-out energy, the target energy above the stopping degree,

$$\mathcal{E}_{>\gamma} := \sum_{k>\gamma}^{k_0} \|f_k^\star\|_{\mathcal{D}}^2 \asymp \sum_{k>\gamma}^{k_0} d^{-k-1}\beta_k^2, \tag{6.2}$$

where $\beta_k^2$ is the squared norm of the degree-$k$ coefficient vector of $\beta$ (Appendix D.2). This is the part of the target the student cannot have learned by time $T_\gamma$, and it is a floor for its error. Everything else is small: the bias on degrees up to $\gamma$ and the variance inherited from the teacher are both $o(\overline{\mathcal{L}}_{\mathrm{TE}})$ because $\gamma < \alpha$. Hence $\overline{\mathcal{L}}_{\mathrm{ST}}(T_\gamma) \asymp \mathcal{E}_{>\gamma} + o(\overline{\mathcal{L}}_{\mathrm{TE}})$, and the rule is a simple trade-off: WTS occurs if and only if the energy held out by the stop is negligible against the teacher's own loss

$$\mathcal{E}_{>\gamma} = o(\overline{\mathcal{L}}_{\mathrm{TE}}). \tag{6.3}$$

The condition is the same in both cases of Figure 3; what changes between them is the form of the teacher error,

$$\overline{\mathcal{L}}_{\mathrm{TE}} \asymp \sum_{k\le\alpha} d^{-\alpha-1}\beta_k^2 + \sum_{\alpha<k\le k_0} d^{-k-1}\beta_k^2 \ (\alpha \le k_0), \qquad \overline{\mathcal{L}}_{\mathrm{TE}} \asymp d^{-\alpha-1}\sum_{k\le k_0}\beta_k^2 \ (\alpha > k_0). \tag{6.4}$$

When $\alpha \le k_0$ the teacher learns the degrees up to $\alpha$ and leaves the rest essentially unlearned; when $\alpha > k_0$ it has the same error $d^{-\alpha-1}$ on every degree. In each case (6.3) compares the held-out energy with the teacher error (6.4) (Appendix D.2, Remark 2).

Whether the condition holds depends on the coefficients, and a two-mode target can illustrate the possible behaviors. Let

$$f_d^\star = \sqrt{1-w_d}\,Y_{2,d} + \sqrt{w_d}\,Y_{4,d}, \tag{6.5}$$

where $Y_{2,d}$ and $Y_{4,d}$ are orthonormal spherical harmonics of degrees two and four, so $k_0 = 4$, and $w_d \to 0$ is the energy of the small degree-four mode.

**A vanishing teacher error is not enough.** Take $m = d^3$, so $\alpha = 3 < k_0 = 4$ (case (a)), with $w_d = c/d$ for a constant $c > 0$. The teacher error tends to zero, but only because $w_d \to 0$: the teacher learns the dominant degree-two mode, while the degree-four direction lies above its capacity edge, so almost all of the small degree-four energy $w_d$ stays unlearned. The degree-four mode is then a floor for every student:

$$\overline{\mathcal{L}}_{\mathrm{TE}} \asymp d^{-1}, \qquad \inf_{t\ge 0}\overline{\mathcal{L}}_{\mathrm{ST}}(t) \asymp d^{-1}.$$

Early stopping removes the degree-two error, but the unlearned degree-four energy $w_d \asymp d^{-1}$ remains at every stopping time, so the best student improves on the teacher by at most a constant factor, and no polynomial improvement is possible. In contrast, the guarantee (6.1) needs the teacher to learn every mode, not just to have small error. This is the role of the condition $\alpha > k_0$ in (6.1). Proposition 2 in Appendix E proves the bounds, and Figure 5 there illustrates their convergence.

**Incompatible stopping times can rule out quadratic improvement.** Take $m = d^5$, so $\alpha = 5 > k_0 = 4$ (case (b)), with $w_d = Bd^{-5}$ for a constant $B > 0$. Now the teacher learns both modes, and each mode on its own would admit quadratic improvement. In the mixture, the time needed to learn the small degree-four mode is also the time at which the student is copying the teacher's degree-four error on the dominant mode. The infimum over $t$ is attained, up to constants, at any time of order $1/\sigma_4^2$, and

$$\overline{\mathcal{L}}_{\mathrm{TE}} \asymp d^{-3}, \qquad \inf_{t \geq 0} \overline{\mathcal{L}}_{\mathrm{ST}}(t) \asymp d^{-5} \asymp \overline{\mathcal{L}}_{\mathrm{TE}}^{5/3}. \tag{6.6}$$

The improvement is polynomial, as (6.1) guarantees, but the lower bound holds at every stopping time and rules out quadratic improvement (Proposition 3 in Appendix E).

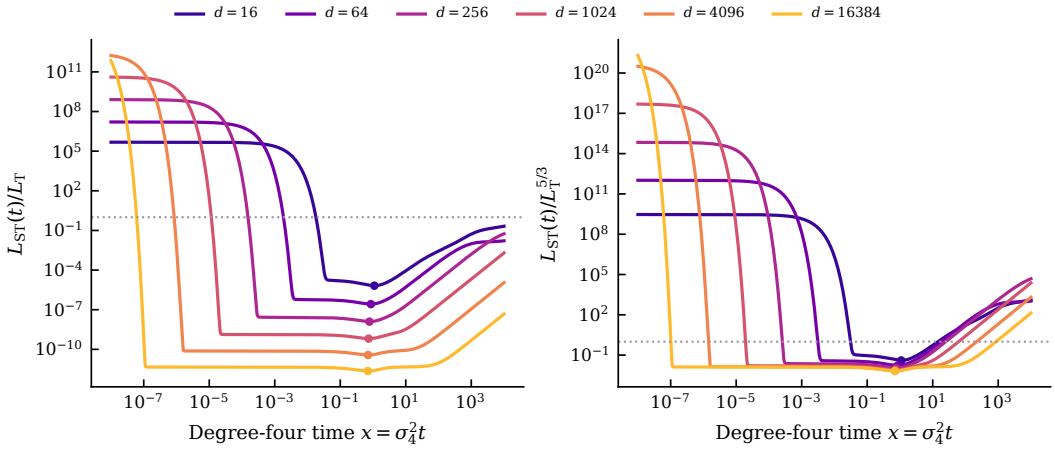



Figure 4: **A timing conflict that prevents quadratic improvement.** Deterministic-equivalent curves for the degree-two/degree-four mixture in (6.6), with $B = C^2/24$ and $C = (12\pi - 37)/24$ (Remark 3). Left: student-to-teacher loss ratio; right: student loss divided by $\overline{\mathcal{L}}_{\mathrm{TE}}^{5/3}$. Dots are numerically attained minima. The right-hand curves converge to $C^{1/3}/48$, the leading constant of the normalized optimum in (6.6) (Remark 3).

## 7 Conclusions and Further Discussion

Using deterministic equivalents, we have computed the leading-order population errors of a random feature teacher and an early-stopped student, and shown that a pure harmonic target admits optimal quadratic improvement. The mechanism is early stopping: the student learns each spectral component of the target on its own time scale and inherits the teacher's feature variance only on the slower time scale of the teacher's capacity edge, and weak-to-strong generalization is the window between the two.

One of the key takeaways is that this window is target dependent: different targets require different stopping times for optimal improvement, and for a target with several components, incompatible stopping times can rule out quadratic improvement even when every component would admit it on its own. Nonetheless, the effect is robust provided the teacher has the capacity to learn every component of the target: a fixed gap below capacity guarantees polynomial improvement for a range of stopping times. When the target has a component beyond the teacher's capacity, no stopping time gives more than a constant-factor gain. This illustrates a key limitation of weak-to-strong generalization: the student can suppress the teacher's variance, but it cannot recover what the teacher could not represent.

We mention some key limitations of our analysis. The conclusions remain conditional on Gaussian universality and deterministic-equivalent accuracy; controlling that approximation is a mathematical direction for further work. We have taken the student infinitely wide. A finite student of width $M_{\mathrm{ST}}$ adds a second source of feature variance, and one expects a threshold effect: once $M_{\mathrm{ST}}$ is large enough relative to the teacher. Finite student widths admit deterministic equivalent analaysis as well.

**AI Use Statement.** This paper was prepared with the assistance of LLMs. LLMs were used to generate experiment code and figures, restructure and revise the document, generate small proofs and other mathematical artifacts (specifically, the examples and the last three appendices), and verify sections of the paper. The majority of the text was hand-written and all main text was hand-revised.

## REFERENCES


Alnur Ali, J. Zico Kolter, and Ryan J. Tibshirani. A continuous-time view of early stopping for least squares regression. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1370–1378. PMLR, 2019. URL https://proceedings.mlr.press/v89/ali19a.html.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017. URL http://jmlr.org/papers/v18/14-546.html.

Simone Bombari and Marco Mondelli. Towards understanding the word sensitivity of attention layers: A study via random features. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4300–4328. PMLR, 2024. URL https://proceedings.mlr.press/v235/bombari24b.html.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeff Wu. Weak-to-strong generalization: eliciting strong capabilities with weak supervision. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Moses Charikar, Chirag Pabbaraju, and Kirankumar Shiragur. Quantifying the gain in weak-to-strong generalization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=MyVyH5Jo1l.

Chen Cheng and Andrea Montanari. Dimension free ridge regression. *The Annals of Statistics*, 52 (6):2879–2912, 2024. doi: 10.1214/24-AOS2449.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, volume 32, pp. 2937–2947. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/ae614c557843b1df326cb29c57225459-Abstract.html.

Romain Couillet and Zhenyu Liao. *Random Matrix Methods for Machine Learning*. Cambridge University Press, 2022. doi: 10.1017/9781009128490. https://zhenyu-liao.github.io/book/.

Leonardo Defilippis, Bruno Loureiro, and Theodor Misiakiewicz. Dimension-free deterministic equivalents and scaling laws for random feature regression. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=FBLJIfW64D.

Elvis Dohmatob and Alberto Bietti. On the (non-)robustness of two-layer neural networks in different learning regimes. *arXiv preprint arXiv:2203.11864*, 2022. doi: 10.48550/arXiv.2203.11864. URL https://arxiv.org/abs/2203.11864.

Yijun Dong, Yicheng Li, Yunai Li, Jason D. Lee, and Qi Lei. Discrepancies are virtue: Weak-to-strong generalization through lens of intrinsic dimension. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=q7PAIE0z5p.

Laszlo Erdős. The matrix dyson equation and its applications for random matrices. *arXiv preprint arXiv:1903.10060*, 2019.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029–1054, 2021. doi: 10.1214/20-AOS1990.

Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. The gaussian equivalence of generative models for learning with shallow neural networks. In *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, volume 145, pp. 426–471. PMLR, 2022. URL https://proceedings.mlr.press/v145/goldt22a.html.

Walid Hachem, Philippe Loubaton, Jamal Najim, and Pascal Vallet. On bilinear forms based on the resolvent of large random matrices. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 49(1):36–63, 2013. doi: 10.1214/11-AIHP450. URL https://numdam.org/articles/10.1214/11-AIHP450/.

Reinhard Heckel and Fatih Furkan Yilmaz. Early stopping in deep networks: Double descent and how to eliminate it. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=tlV90jvZbw.

Hong Hu and Yue M. Lu. Universality laws for high dimensional learning with random features. *IEEE Transactions on Information Theory*, 69(3):1932–1964, 2023. doi: 10.1109/TIT.2022.3217698. URL https://ieeexplore.ieee.org/document/9931146/.

Muhammed Emrullah Ildiz, Halil Alperen Gozeten, Ege Onur Taga, Marco Mondelli, and Samet Oymak. High-dimensional analysis of knowledge distillation: Weak-to-strong generalization and scaling laws. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=1xzqz73hvL.

Hunter Lang, David Sontag, and Aravindan Vijayaraghavan. Theoretical analysis of weak-to-strong generalization. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 46837–46880. Curran Associates, Inc., 2024. doi: 10.52202/079017-1486. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/5358d1e6138dff5718c5e9790f5fa593-Paper-Conference.pdf.

Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 4313–4324. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/li20j.html.

Marko Medvedev, Kaifeng Lyu, Dingli Yu, Sanjeev Arora, Zhiyuan Li, and Nathan Srebro. Weak-to-strong generalization even in random feature networks, provably. *arXiv preprint arXiv:2503.02877*, 2025.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022. doi: 10.1002/cpa.22008. URL https://doi.org/10.1002/cpa.22008.

James A. Mingo and Roland Speicher. *Free Probability and Random Matrices*, volume 35 of *Fields Institute Monographs*. Springer, New York, 2017. ISBN 978-1-4939-6942-5. doi: 10.1007/978-1-4939-6942-5.

Theodor Misiakiewicz and Basil Saeed. A non-asymptotic theory of kernel ridge regression: Deterministic equivalents, test error, and GCV estimator. *arXiv preprint arXiv:2403.08938*, 2024. doi: 10.48550/arXiv.2403.08938.

Behrad Moniri and Hamed Hassani. On the mechanisms of weak-to-strong generalization: A theoretical perspective. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=qR5B1Es9Ty.

Andrea Montanari and Basil N. Saeed. Universality of empirical risk minimization. In *Proceedings of the Thirty Fifth Conference on Learning Theory*, volume 178, pp. 4310–4312. PMLR, 2022. URL https://proceedings.mlr.press/v178/montanari22a.html.

Abhijeet Mulgund and Chirag Pabbaraju. Relating misfit to gain in weak-to-strong generalization beyond the squared loss. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=wITGhlMPXz.

Junsoo Oh, Jerry Song, and Chulhee Yun. From linear to nonlinear: Provable weak-to-strong generalization through feature learning. In *High-dimensional Learning Dynamics 2025*, 2025. URL https://openreview.net/forum?id=llHl4XN0yV.

Elliot Paquette. Random matrix theory of high-dimensional optimization, 2024. URL https://elliotpaquette.github.io/notes/RMTandOptimization-20240701.pdf. Lecture notes, version dated July 1, 2024.

Leonardo Petrini, Francesco Cagnetta, Eric Vanden-Eijnden, and Matthieu Wyart. Learning sparse features can lead to overfitting in neural networks. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In J. Platt, D. Koller, Y. Singer, and S. Roweis (eds.), *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/file/013a006f03dbc5392effeb8f18fda755-Paper.pdf.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Early stopping and non-parametric regression: An optimal data-dependent stopping rule. *Journal of Machine Learning Research*, 15(11):335–366, 2014. URL https://jmlr.org/papers/v15/raskutti14a.html.

Dominik Schröder, Hugo Cui, Daniil Dmitriev, and Bruno Loureiro. Deterministic equivalent and error universality of deep random features learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 30285–30320. PMLR, 2023. URL https://proceedings.mlr.press/v202/schroder23a.html.

Changho Shin, John Cooper, and Frederic Sala. Weak-to-strong generalization through the data-centric lens. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=uogG8BfLs2.

Jack W. Silverstein and Z. D. Bai. On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):175–192, 1995. URL https://jack.math.ncsu.edu/mandp.pdf.

James B. Simon, Madeline Dickens, Dhruva Karkada, and Michael R. DeWeese. The eigenlearning framework: A conservation law perspective on kernel regression and wide neural networks. *Transactions on Machine Learning Research*, 2023. URL https://arxiv.org/abs/2110.03922.

James B. Simon, Dhruva Karkada, Nikhil Ghosh, and Mikhail Belkin. More is better in modern machine learning: When infinite overparameterization is optimal and overfitting is obligatory. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2311.14646.

David Xing Wu and Anant Sahai. Provable weak-to-strong generalization via benign overfitting. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4vzGQcVUG8.

Diyuan Wu, Lehan Chen, Theodor Misiakiewicz, and Marco Mondelli. Improved scaling laws via weak-to-strong generalization in random feature ridge regression, 2026. URL https://arxiv.org/abs/2603.05691.

Yihao Xue, Jiping Li, and Baharan Mirzasoleiman. Representations shape weak-to-strong generalization: Theoretical insights and empirical predictions. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=ypEW077kle.

Wei Yao, Wenkai Yang, Ziqiao Wang, Yankai Lin, and Yong Liu. Understanding the capabilities and limitations of weak-to-strong generalization. In *Scaling Self-Improving Foundation Models without Human Supervision*, 2025.

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007. doi: 10.1007/s00365-006-0663-2.

## SUPPLEMENTARY MATERIAL

The appendices give the extended related work, the deterministic-equivalent calculations, and the proofs for the ReLU setting. The general bias–variance decomposition follows the resolvent calculations. ReLU spectral and fixed-point estimates precede the error asymptotics; the stopping-time analysis, the two mixed-target examples, and the extension to other activations follow. Numerical details for the main figures are collected at the end.


| | Contents | Page |
|---|---|---|
| A | Extended related work | 14 |
| | Comparison with Medvedev et al. (2025) | 15 |
| B | Deterministic equivalents for teacher and student errors | 17 |
| | Dyson equation and the Gaussian approximation | 17 |
| | Single-resolvent teacher calculation | 20 |
| | Double-resolvent student calculation | 22 |
| | Bias–variance decomposition | 27 |
| C | Proofs for ReLU activation | 28 |
| | Harmonic spectrum | 28 |
| | Teacher error scaling | 34 |
| | Student error scaling: quadratic improvement | 35 |
| D | General stopping times and targets | 43 |
| E | Two mixed-target examples | 60 |
| | A component above capacity: no polynomial improvement | 62 |
| | Incompatible stopping times: no quadratic improvement | 63 |
| | Leading constants used in the figures | 64 |
| F | Other activations | 64 |
| | Harmonic and Hermite coefficients | 66 |
| | Quadratic improvement for general activations | 66 |
| | Examples: ReLU, tanh, logistic sigmoid | 68 |
| G | Numerical details for the main figures | 69 |


## A EXTENDED RELATED WORK

Since Burns et al. (2024) first empirical investigation of the WTS phenomenon, many works attempted to do a theoretical analysis of it using simple settings. We discuss below a selection of works representing the range of different approaches used to understand the phenomenon.

**Data-dependent mechanisms.** First, some works also focus on understanding how data structure influences the WTS phenomenon. Lang et al. (2024) demonstrate that the gain in performance comes not only from the student's correction of the teacher's mistakes but also from the student's good performance on a portion not covered by the teacher labels. This occurs when one assumes that the strong model's predictions are robust in neighborhoods of pseudo-labeled points (teacher-labeled points) and that bad points, that is, points with incorrect pseudo-labels, have many good neighbors. They also provide empirical evidence for these assumptions in finite samples. Shin et al. (2025) define a quantity called *overlap density*, which represents how frequently a dataset contains points with both hard and easy patterns. The weak model learns only the easy patterns, but the strong model is able to learn the hard patterns from the overlap points and thereby generalize to points with only hard patterns. Building on this intuition, they theoretically prove that the WTS phenomenon can be expressed as a function of the overlap density. In the classification setting, Wu & Sahai (2025) define weak and strong models based on how well their features represent the data features. Considering a teacher whose labels are asymptotically equivalent to random guessing, they show that, under certain conditions, the student's performance undergoes an asymptotic transition from random guessing to perfect generalization.

**Misfit error.** Charikar et al. (2024) consider a more general notion of weak and strong models, in which the difference between them is determined by how well their representations capture the underlying data. This can include differences in expressivity, as in our setting, as well as other

factors such as the amount of pretraining. They study a regression setting with the squared loss and quantify the WTS gain in terms of the *misfit error*. This quantity is defined as the error of the strong student model on the labels generated by the weak teacher model; equivalently, it measures the squared disagreement between the student and the teacher. Under their assumptions, they show that the error of the strong model evaluated on the ground-truth labels is bounded above by the error of the weak model on the ground-truth labels minus the misfit error. In our notation, this corresponds to $\mathcal{L}_{\mathrm{ST}} \leq \mathcal{L}_{\mathrm{TE}} - \mathrm{misfit}$ Using representation and projection techniques, they show that, globally, the component of the teacher predictor that is not reproduced by the student contributes to the teacher's error relative to the ground-truth labels. Consequently, the larger the misfit, the larger the improvement. Yao et al. (2025) establish a corresponding upper bound while also studying lower bounds. Unlike Charikar et al. (2024), they use a KL-divergence-based loss instead of the squared loss. Other work that studies the *misfit error* includes Mulgund & Pabbaraju (2025), who extend the misfit-based inequality to loss functions given by Bregman divergences. This extension also includes the classification setting, since cross-entropy can be expressed in terms of the KL divergence, which is a Bregman divergence.

**Representation geometry.** Other works study WTS from a geometric and representation-based perspective. Xue et al. (2025) quantify the WTS improvement in terms of the overlap between the student principal-component subspace and the orthogonal complement of the teacher principal-component subspace. The principal-component subspace of each model contains the dominant directions of variance in its internal representation. Therefore, this overlap identifies directions that are not captured by the teacher but are well represented by the student, allowing the student to correct some of the teacher's errors. From a related geometric perspective, Dong et al. (2025) analyze ridgeless regression with intrinsically low-dimensional feature spaces. They show that the student inherits the teacher's variance along directions shared by their representations. In contrast, along directions represented by the teacher but not by the student, this variance is reduced by a factor given by the ratio between the dimension of the student feature space and the number of weakly labeled samples. Thus, differences between the two representation spaces prevent the student from fully inheriting the teacher noise and then improving generalization.

**Regularization and early-stopping.** With respect to how regularization procedures influence the phenomenon, Moniri & Hassani (2026) show that, in the setting of linear ridge regression, the student can surpass the teacher by having better regularization. This is similar to the findings of Medvedev et al. (2025) and our work, but we use implicit regularization imposed by early-stopped gradient flow instead of explicit regularization. They also show that student regularization that is better aligned with the target function improves performance. They also go beyond linear models and frozen representations by considering a pretrained student in a nonlinear setting. The student learns easy, task-specific features from the teacher while retaining hard features acquired during pretraining that the teacher cannot capture, allowing it to outperform the teacher.

**Beyond linear.** Going beyond linear settings, Oh et al. (2025) use CNNs for both the teacher and the student and adopt yet another characterization of model strength. They characterize it through the activation function: a linear activation for the weak teacher and ReLU for the strong student. Even in this setting, which differs from ours, they find that, in a data-abundant regime, early stopping the student training leads to better performance.

**Scaling laws.** Finally, regarding the important problem of scaling laws, Ildiz et al. (2025) study ridge regression trained on the predictions of a surrogate model in the proportional high-dimensional regime. Using deterministic equivalents for the two-stage procedure, they characterize when the student improves on the surrogate and derive scaling laws in the sample size and model size. Wu et al. (2026) show that WTS can improve the scaling law of the test error in a random feature ridge regression setting. Closely related to our work, they also derive a deterministic equivalent for the student test error.

### A.1 Comparison with the ReLU bounds of Medvedev et al. (2025)

Our model and two-step procedure are those of Medvedev et al. (2025), so their ReLU bounds can be evaluated in the setting of Theorems 2 and 4 and compared term by term with the deterministic equivalent. Medvedev et al. (2025, Theorem 6.7) bounds the student error for a two-layer random-feature teacher with any activation, and Medvedev et al. (2025, Corollary 6.8) specializes it to ReLU

on the sphere, where the eigenvalues come with the multiplicities $N_k$. In our notation (teacher width $m$, harmonic coefficients $\sigma_k$, admissible degrees $\mathcal{K}$, and $s^+$ the smallest degree in $\mathcal{K}$ larger than $s$), the corollary reads as follows. Let $f^\star$ be a unit target of degree $k_0$, let the teacher be the population-optimal one, and let $T > 0$. Then, for every $s \in \mathcal{K}$ with $s \geq k_0$, with probability at least $1 - 2/m - N_{k_0} e^{-cm/N_{k_0}}$ over the teacher features,

$$\begin{aligned} \mathcal{L}_{\mathrm{ST}}(T) \leq & \frac{4\big(\sum_{j\leq s}\sigma_j^2 N_j\big)\big(\sum_{j>s}\sigma_j^2 N_j\big)\log^2 m}{\sigma_s^4\, m}\,\mathcal{L}_{\mathrm{TE}} && \Big\} \text{(A) alignment} \\ & + \sigma_{s^+}^4 T^2\,\mathcal{L}_{\mathrm{TE}} && \Big\} \text{(B) learned high degrees} \\ & + \frac{e^{-\sigma_{k_0}^2 T}}{2 - e^{-\sigma_{k_0}^2 T}}\,\|f^\star\|_{\mathcal{D}}^2 && \Big\} \text{(C) unlearned target.} \end{aligned} \tag{A.1}$$

Since the bound holds for every admissible $s$ and the probability does not depend on $s$, it also holds with the infimum over $s \in \mathcal{K}$, $s \geq k_0$, of the right-hand side, which is the form stated in their corollary.

**Evaluation.** Take $m = d^\alpha$ with $\alpha > k_0$, so that $\overline{\mathcal{L}}_{\mathrm{TE}} \asymp d^{-(\alpha-k_0)}$, and stop at the harmonic time $T_\gamma \asymp d^{\gamma+1}\log d$ of Section 5 with $\gamma \geq k_0$; Theorem 2 is the case $\gamma = k_0$. Using $N_j \asymp d^j$, $\sigma_j^2 \asymp d^{-j-1}$ and $\sum_j \sigma_j^2 N_j = K(x,x) = 1/(2d)$, where every fixed degree carries a $d$-independent share (Appendix C.1), the three terms of (A.1) become

| term | at a free $s$ | at the best $s$, namely $s = \gamma$ |
|---|---|---|
| (A) alignment | $d^{\,2s-\alpha}\log^2 d \cdot \mathcal{L}_{\mathrm{TE}}$ | $d^{\,2\gamma-\alpha}\log^2 d \cdot \mathcal{L}_{\mathrm{TE}}$ |
| (B) learned high degrees | $d^{-2(s^+-\gamma)}\log^2 d \cdot \mathcal{L}_{\mathrm{TE}}$ | $d^{-2(\gamma^+-\gamma)}\log^2 d \cdot \mathcal{L}_{\mathrm{TE}}$ |
| (C) unlearned target | $\leq \mathcal{L}_{\mathrm{TE}}/m = d^{-\alpha}\,\mathcal{L}_{\mathrm{TE}}$ | $\leq \mathcal{L}_{\mathrm{TE}}/m = d^{-\alpha}\,\mathcal{L}_{\mathrm{TE}}$ |

Table 1: The three terms of (A.1) for a unit target of degree $k_0 < \alpha$, stopped at $T_\gamma$ with $\gamma \geq k_0$, at a free $s \in \mathcal{K}$ with $s \geq \gamma$ and at the $s$ that minimizes their sum. Term (A) does not depend on the stopping time, and term (C) does not depend on $s$.

**Choice of $s$.** Term (A) increases with $s$ and term (B) decreases with $s$, so the sum is smallest where the two cross. Because $\mathcal{K}$ is discrete the crossing cannot be attained exactly, and the smallest admissible choice $s = \gamma$ is optimal for every $\alpha \leq 2\gamma + 8$; beyond that a larger $s$ is better. Writing $\mathcal{L}_{\mathrm{ST}}(T_\gamma) \lesssim d^{-r}\log^2 d \cdot \mathcal{L}_{\mathrm{TE}}$, the third column gives $r = \min\big(\alpha - 2\gamma,\, 2(\gamma^+ - \gamma)\big)$, so term (A) is the binding one exactly when $\alpha \leq 2\gamma + 2(\gamma^+ - \gamma)$, which is $\alpha \leq 2\gamma + 4$ for even $\gamma$.

**Key differences between Medvedev et al. (2025) and this work:**

- **Improvement threshold.** At $\gamma = k_0$ the bound (A.1) improves on the teacher only when $\alpha > 2k_0$, whereas the deterministic equivalent improves for every $\alpha > k_0$. The same threshold appears as the hypothesis $m \geq \Theta(d^{2k})$ of Medvedev et al. (2025, Theorem 3.2).
- **Improvement at the quadratic stop.** At $\gamma = k_0$, in the range where term (A) is rate-limiting, the bound (A.1) gives $\mathcal{L}_{\mathrm{ST}}/\mathcal{L}_{\mathrm{TE}} \lesssim d^{-(\alpha-2k_0)}\log^2 d$, whereas Theorem 2 gives $\overline{\mathcal{L}}_{\mathrm{ST}}/\overline{\mathcal{L}}_{\mathrm{TE}} \asymp d^{-(\alpha-k_0)}$.
- **Best exponent.** Writing $\mathcal{L}_{\mathrm{ST}} \lesssim \mathcal{L}_{\mathrm{TE}}^e$ with $e = 1 + r/(\alpha - k_0)$, at $\gamma = k_0$ this is $e = 2 - k_0/(\alpha - k_0)$ while term (A) is rate-limiting, at most $5/3$ over all $k_0$ and $\alpha$, and tending to $3/2$ as $\alpha \to \infty$. That limit is the exponent $1.49$ of Medvedev et al. (2025, Theorem 3.2), obtained there for fixed $d$ and $m \to \infty$. Theorem 2 gives $e = 2$.
- **Later stops.** Here $r \leq \alpha - 2\gamma$, so the bound (A.1) decays monotonically in $\gamma$ and is vacuous once $\gamma \geq \alpha/2$, with no phase structure. Theorem 4 gives improvement for every $\gamma < \alpha$, with exponent $\rho \in (1, 2]$ and the five phases of Figure 2.

**Scope.** The comparison is one-sided in an important respect. Medvedev et al. (2025, Theorem 6.7) is nonasymptotic, holds with high probability for the actual random-feature teacher without a Gaussian universality assumption, covers any bounded activation, and applies at fixed $d$ with $m \to \infty$,

a regime about which our $d \to \infty$ scaling with $m = d^\alpha$ says nothing. Our results describe the deterministic equivalent and transfer to the spherical model under Assumption 1. Within the common regime of a pure harmonic target and polynomial teacher width, the deterministic equivalent gives the sharp exponent, the sharp threshold, and the phase diagram; the general bound gives an exponent strictly below the quadratic ceiling that their own lower bound establishes.

# B Deterministic equivalents for the teacher and student errors

In this appendix we calculate the deterministic equivalents of the teacher and student population errors. The calculations are general and do not depend on the activation. The specific values of $\mathcal{U}, \mathcal{D}, \beta$ and $\Psi$ will be substituted in Appendix C. Also, to ease notation, we will use $m$ here to denote the number of neurons of the teacher, $m = M_{\mathrm{TE}}$; the student is the infinite-width spectral flow (2.1).

## B.1 Background on the Dyson equation and Gaussian approximation

In the random features setting we have two sources of randomness: the data distribution and the random features distribution. It turns out that when calculating the population error, the latter remains as a resolvent of a random covariance matrix. The concept of deterministic equivalents of random matrices can be useful in this case. The idea is to find a deterministic matrix around which this random resolvent concentrates. This deterministic matrix will have the same spectral information as its random counterpart when the dimension is very large and the reduced randomness makes the computations more tractable.

One might consider the expectation matrix as a natural choice of deterministic equivalent but, as argued in Couillet & Liao (2022), this choice is not desirable since it involves an integration over a probability space which can be very challenging numerically. Instead, a more tractable deterministic equivalent can be derived through the associated matrix Dyson equation (Erdős, 2019). For a Gaussian random matrix $A \in \mathbb{R}^{d\times d}$ the matrix Dyson equation for its resolvent $R(z, A)$ can be derived by applying the Stein's lemma together with the tautological equation

$$R(z; A)(A - zI) = I \tag{B.1}$$

where $I \in \mathbb{R}^{d\times d}$ is the identity matrix, and some concentration properties of the Gaussian distribution. This leads to a self-consistent equation for a scalar quantity, and by some technical stability results, one can show that the Dyson equation has a unique solution which is given in terms of this scalar quantity.

We illustrate this theory with a related example. Put $n = d$ and let $A = P + SGS$, where $P = \mathbb{E}[A]$, and $G$ is a standard GOE matrix, i.e, a matrix with independent upper-triangular entries given by

$$G_{ij} \sim \begin{cases} \mathcal{N}\left(0, \frac{1}{n}\right), & i < j, \\ \mathcal{N}\left(0, \frac{2}{n}\right), & i = j, \end{cases}$$

and $S$ is a symmetric matrix that encodes the variance profile.

Stein's lemma in matrix format gives

$$\mathbb{E}[Zf(Z)] = \Sigma\mathbb{E}[\nabla f(Z)]$$

where $f$ is a nice test function and $Z$ is a gaussian matrix. Define $Z = A - \mathbb{E}[A]$, we have $Z \sim \mathcal{N}(0, \Sigma)$ where $\Sigma \in \mathbb{R}^{n^2\times n^2}$ has entries

$$\Sigma_{ij,k\ell} = \frac{1}{n}\left[\left(S^2\right)_{ik}\left(S^2\right)_{j\ell} + \left(S^2\right)_{i\ell}\left(S^2\right)_{jk}\right].$$

Applying the lemma for $f(Z) = R(z, A)$ gives, to leading order,

$$\mathbb{E}[R(z; A)(A - P)] \approx -\mathbb{E}[R(z; A)\mathcal{S}(R(z; A))]$$

where

$$\mathcal{S}(R) \coloneqq \frac{1}{n}S^2 \operatorname{tr}(RS^2).$$

The other GOE contraction contributes $-n^{-1}\mathbb{E}[RS^2RS^2]$, which is of smaller order for bounded $S$ and a spectral parameter bounded away from the real axis. Concentration then motivates factorizing the leading expected product. Writing $\overline{R} = \mathbb{E}[R(z;A)]$, we obtain

$$\mathbb{E}[R(z;A)(A-P)] \approx -\overline{R}\mathcal{S}(\overline{R}).$$

Combining this with the expectation of (B.1) motivates the **Dyson equation**:

$$M(P - \mathcal{S}(M) - zI) = I. \tag{B.2}$$

where $M \in \mathbb{C}^{d\times d}$ is the deterministic equivalent of $R(z;A)$ for a nonreal spectral parameter $z$. This is a background derivation; extending its error control toward the spectrum requires the relevant stability estimates.

We then show that the resolvent $R(z;A)$ is a solution of a perturbed Dyson equation and by stability and uniqueness of the solution, the resolvent is close to the solution of the original one, that is, we ended up showing that $R(z;A)$ is close to $M$ in a suitable sense (see Paquette (2024) for more details).

Defining the scalar quantity $\kappa(z) \coloneqq n^{-1}\operatorname{tr}(MS^2)$, (B.2) becomes

$$M = (\mathbb{E}[A] - S^2\kappa(z) - zI)^{-1} \tag{B.3}$$

and we obtain the scalar self consistent equation

$$\kappa(z) = \frac{1}{n}\operatorname{tr}(S^2(\mathbb{E}[A] - S^2\kappa(z) - zI)^{-1}).$$

Thus, to find $M$ we only need to solve this self-consistent equation for $\kappa(z)$ and substitute in (B.3).

In our case, we will be looking for the deterministic equivalent of a feature-feature covariance matrix. Let $X_1, X_2, ..., X_n \in \mathbb{R}^d$ be the rows of $X \in \mathbb{R}^{n\times d}$. The corresponding feature-feature covariance matrix is given by

$$A = \frac{1}{n}X^TX \in \mathbb{R}^{d\times d}.$$

In this situation, if $X$ has Gaussian entries, the matrix $A$ will depend quadratically on a Gaussian distribution so applying Stein's lemma does not directly yield the linear Dyson form above. It turns out that it is possible to use a linearization trick from free probability (see subsection 10.3 of Mingo & Speicher (2017)) to solve this issue and also obtain a Dyson equation for the resolvent of these type of matrices. Again, see Paquette (2024) for details on how to derive the Dyson equation and deterministic equivalent in this case. We use these calculations in the single- and double-resolvent analyses below.

An issue arises for our ReLU random features setting. In this case, both the data and the random features are sampled uniformly from the sphere. The Dyson equation framework, however, is derived using Gaussian integration by parts, or equivalently Stein's lemma. We therefore invoke the **Gaussian Universality Ansatz** to connect the spherical ReLU model to a Gaussian model with the same relevant covariance structure. We explain heuristically why one should expect it, and then we discuss related literature where versions of this have been proven rigorously.

**Cumulant expansion heuristic.** Stein's lemma can be generalized to general random variables, in the form of a *cumulant expansion*. For a random variable $X$ and a smooth enough function $f$

$$\mathbb{E}[Xf(X)] = \kappa_1(X)\mathbb{E}[f(X)] + \kappa_2(X)\mathbb{E}[f'(X)] + \frac{\kappa_3(X)}{2}\mathbb{E}[f''(X)] + \cdots, \tag{B.4}$$

where $\kappa_i(X)$ is the $i$-th cumulant of $X$. Note that since all cumulants for $k > 2$ of the Gaussian distribution are zero, (B.4) gives Stein lemma when $X$ is Gaussian. With that, what we are in fact supposing is that the higher cumulants of the random vector $\sigma(U^\top x)$ (for $U$ a matrix of uniform spherical vectors and data vectors $x$) do not influence resolvent statistics; indeed, entries of this vector are affected by large collections of almost independent random variables which lead to central limit theorem effects.

**Related literature on the Gaussian Universality Ansatz.** A central work in this area is Hu & Lu (2023), who prove that, under suitable assumptions in the proportional regime, nonlinear random feature models and Gaussian models with matching covariance have asymptotically equivalent

training and generalization errors. Related results were obtained by Montanari & Saeed (2022), who establish covariance universality for the training error and the minimum test error among near empirical risk minimizers. The Gaussian equivalence principle was also studied by Goldt et al. (2022), who established rigorous conditions for single layer generative models and applied the Gaussian description to pre learned features and kernel methods. The work of Schröder et al. (2023) is particularly close to ours, since it combines Gaussian universality with deterministic equivalents for deep random feature models. They prove universality of the ridge regression test error when the learner and target share the same intermediate layers, and conjecture extensions to more general architectures and convex losses.

**Related literature on deterministic equivalent approximations.** Deterministic equivalents can provide quantitatively accurate approximations to population errors. In particular, Defilippis et al. (2024) establish a dimension free nonasymptotic multiplicative approximation for random feature ridge regression under concentration and spectral assumptions. Their bounds give vanishing relative errors under suitable scaling of the sample size, number of features and regularization, even when the population error itself vanishes. Moreover, their infinite data limit agrees with our teacher deterministic equivalent. Related nonasymptotic bounds for ridge regression and kernel ridge regression are given by Cheng & Montanari (2024) and Misiakiewicz & Saeed (2024), respectively. These results support expecting accurate deterministic approximations in our setting, although they do not directly control our approximation errors. Our analysis considers a two stage learning procedure, where teacher generated labels introduce additional dependencies and cross terms. A closely related result is given by Wu et al. (2026), who derive a deterministic equivalent for the student excess error in a teacher student random feature ridge regression model, together with an additive nonasymptotic bound. Their result establishes relative accuracy when the approximation remainder is negligible compared with the student error.

### B.2 Optimally trained teacher

Since we assume that the teacher is optimally trained, let us first calculate the optimal weight vector. This is a standard calculation, but we repeat it here for completeness.

**Lemma 1.** Given the setting of Section 2, for the squared population loss of the teacher

$$\mathcal{L}_{\mathrm{TE}} = \mathbb{E}_{x\sim\mathcal{D}}\left[(f_{\mathrm{teacher}}(x) - f^*(x))^2\right],$$

we have that the optimal weight is

$$w_* = (U_m \Psi U_m^T)^{-1} U_m \Psi \beta$$

where $U_m \coloneqq U/\sqrt{m}$, and $U \in \mathbb{R}^{m\times d}$ has the fixed weight vectors $u_i \in \mathbb{R}^d$ as its rows. With that, we obtain

$$\mathcal{L}_{\mathrm{TE}} = \langle U_m^T (U_m \Psi U_m^T)^{-1} U_m \Psi \beta - \beta, \Psi[U_m^T (U_m \Psi U_m^T)^{-1} U_m \Psi \beta - \beta]\rangle.$$

*Proof.* Just plugging in the definitions, we get

$$\begin{aligned}
\mathcal{L}_{\mathrm{TE}} &= \mathbb{E}_{x\sim\mathcal{D}}\left[(f_{\mathrm{teacher}}(x) - f^*(x))^2\right] \\
&= \mathbb{E}_{x\sim\mathcal{D}}\left[\left(\frac{1}{\sqrt{m}} w^T U x - \beta^T x\right)^2\right] \\
&= \mathbb{E}_{x\sim\mathcal{D}}\left[\left(\sum_{l=1}^{d}\left(\frac{1}{\sqrt{m}}\sum_{k=1}^{m} w_k U_{kl} x_l - \beta_l x_l\right)\right)^2\right] \\
&= \sum_{l=1}^{d} \psi_l \left(\frac{1}{\sqrt{m}}\sum_{k=1}^{m} w_k U_{kl} - \beta_l\right)^2
\end{aligned} \tag{B.5}$$

where $w \in \mathbb{R}^m$ is the vector of trainable weights.

Thus, the optimal weight $w_* \in \mathbb{R}^m$ is given by

$$w_* \coloneqq \operatorname*{arg\,min}_{w\in\mathbb{R}^m} \left\langle \frac{1}{\sqrt{m}} U^T w - \beta, \Psi\left(\frac{1}{\sqrt{m}} U^T w - \beta\right)\right\rangle.$$

and it solves

$$\nabla_w \left\langle \frac{1}{\sqrt{m}} U^T w - \beta, \Psi \left( \frac{1}{\sqrt{m}} U^T w - \beta \right) \right\rangle = 0. \tag{B.6}$$

Using that

$$\begin{aligned} \nabla_w \left\langle \frac{1}{\sqrt{m}} U^T w - \beta, \Psi \left( \frac{1}{\sqrt{m}} U^T w - \beta \right) \right\rangle &= \left( \frac{1}{\sqrt{m}} U^T w - \beta \right)^T (\Psi + \Psi^T) \nabla_w \left( \frac{1}{\sqrt{m}} U^T w - \beta \right) \\ &= 2 \frac{1}{\sqrt{m}} U \Psi \left( \frac{1}{\sqrt{m}} U^T w - \beta \right), \end{aligned}$$

we obtain from (B.6) that

$$2 \frac{1}{\sqrt{m}} U \Psi \left( \frac{1}{\sqrt{m}} U^T w_* - \beta \right) = 0,$$

such that

$$w_* = (U_m \Psi U_m^T)^{-1} U_m \Psi \beta.$$

Plugging this expression into the teacher population error (B.5) we get

$$\mathcal{L}_{\mathrm{TE}} = \langle U_m^T (U_m \Psi U_m^T)^{-1} U_m \Psi \beta - \beta, \Psi [U_m^T (U_m \Psi U_m^T)^{-1} U_m \Psi \beta - \beta] \rangle \tag{B.7}$$

as claimed. □

### B.3 Single-resolvent teacher error deterministic equivalent

The self-consistent equation used below belongs to the classical sample covariance theory of Silverstein & Bai (1995); deterministic equivalents for bilinear forms of the resolvent are developed by Hachem et al. (2013). Following the exposition in the lecture notes of Paquette (2024), we introduce the deterministic resolvent and calculate its associated teacher error. The connection to the ReLU model uses the Gaussian Universality Ansatz. Theorem 5 below evaluates the deterministic quantity; approximation of the random teacher error is a separate question.

Recall that, in the coordinates used in this appendix, $\Psi = \mathrm{diag}(\psi_1, \ldots, \psi_d)$ is the population covariance matrix and $\beta$ is the target coefficient vector, so that $f^*(x) = \beta^T x$ and $\|f^*\|^2_{L^2(\mathcal{D})} = \langle \beta, \Psi \beta \rangle$. In particular, $\Psi^{1/2}\beta$ is the target coefficient vector in the corresponding orthonormal basis. Here $m = M_{\mathrm{TE}}$ denotes the teacher width. We use $\Psi + \nu$ as shorthand for $\Psi + \nu I$.

**Definition 1** (Self-consistent parameter)**.** For $0 < m < \mathrm{rank}(\Psi)$, define $\nu > 0$ to be the unique solution of

$$m = \mathrm{tr}\big(\Psi(\Psi + \nu I)^{-1}\big). \tag{B.8}$$

The solution exists and is unique because the right-hand side is continuous and strictly decreasing in $\nu > 0$, from $\mathrm{rank}(\Psi)$ to 0.

**Definition 2** (Deterministic resolvent)**.** For the sample covariance matrix

$$H = \Psi^{1/2} U_m^T U_m \Psi^{1/2}, \qquad R(z) = (H - zI)^{-1},$$

define its deterministic resolvent by

$$\overline{R}(z) = (-\Psi \kappa(z) - zI)^{-1},$$

where $\kappa$ is the analytic solution on the upper half-plane with $\mathrm{Im}\, \kappa(z) > 0$ of

$$-\kappa(z) = \left( 1 + \frac{1}{m} \mathrm{tr}\big(\Psi \overline{R}(z)\big) \right)^{-1}. \tag{B.9}$$

We extend this solution to the lower half-plane by complex conjugation. This is the deterministic resolvent associated with the deformed Marchenko–Pastur law (cf. Theorem 17 in the notes).

Let us first rewrite the teacher population error in terms of the resolvent. In fact, by (B.7) we have

$$\begin{aligned}\mathcal{L}_{\mathrm{TE}} &= -\langle (U_m \Psi U_m^T)^{-1} U_m \Psi\beta, U_m\Psi\beta\rangle + \langle \beta, \Psi\beta\rangle \\ &= -\left\langle \Psi^{1/2}\beta, \left(\sum_{\lambda_j\neq 0} v_j v_j^T\right)\Psi^{1/2}\beta\right\rangle + \langle\beta,\Psi\beta\rangle \\ &= \left\langle \Psi^{1/2}\beta, \left(\sum_{\lambda_j = 0} v_j v_j^T\right)\Psi^{1/2}\beta\right\rangle,\end{aligned}$$

where $\lambda_j$ and $v_j$ are the eigenvalues and an orthonormal basis of eigenvectors of $H$. Here we used that $\Psi^{1/2}U_m^T(U_m\Psi U_m^T)^{-1}U_m\Psi^{1/2}$ is the orthogonal projector onto the range of $H$. Thus, writing $P_0 \coloneqq \sum_{\lambda_j=0} v_j v_j^T$, we obtain $\mathcal{L}_{\mathrm{TE}} = \langle \Psi^{1/2}\beta, P_0\Psi^{1/2}\beta\rangle$.

Then, if $\Gamma \subset \mathbb{C}$ is a small positively oriented contour around zero enclosing no nonzero eigenvalues of $H$, we can write

$$\mathcal{L}_{\mathrm{TE}} = -\frac{1}{2\pi i}\oint_\Gamma \langle \Psi^{1/2}\beta, R(z)\Psi^{1/2}\beta\rangle\, dz.$$

Replacing the resolvent in this expression by its deterministic counterpart, we define the deterministic-equivalent teacher error as

$$\overline{\mathcal{L}}_{\mathrm{TE}} \coloneqq -\frac{1}{2\pi i}\oint_\Gamma \langle \Psi^{1/2}\beta, \overline{R}(z)\Psi^{1/2}\beta\rangle\, dz, \tag{B.10}$$

where we now take $\Gamma$ sufficiently small to enclose only the isolated pole at zero of $\overline{R}$. The following theorem gives an explicit expression for this deterministic quantity.

**Theorem 5** (Deterministic-equivalent teacher error)**.** With $\nu$ as in Definition 1, the deterministic-equivalent teacher error is

$$\overline{\mathcal{L}}_{\mathrm{TE}} = \nu\langle \Psi^{1/2}\beta, (\Psi+\nu I)^{-1}\Psi^{1/2}\beta\rangle.$$

*Proof.* In fact, by (B.10) and Definition 2, we have

$$\overline{\mathcal{L}}_{\mathrm{TE}} = -\frac{1}{2\pi i}\oint_\Gamma \langle \Psi^{1/2}\beta, (-\Psi\kappa(z) - zI)^{-1}\Psi^{1/2}\beta\rangle\, dz.$$

We claim that $\kappa$ is holomorphic near 0, with $\kappa(0)=0$, and that $\kappa(z) = cz + O(z^2)$ with $c = \nu^{-1}$. To see this, let us write $\kappa(z) = zc(z)$. Substituting this into (B.9) gives, for $z \neq 0$,

$$-zc(z) = \left(1 - \frac{1}{mz}\operatorname{tr}\big(\Psi(I + c(z)\Psi)^{-1}\big)\right)^{-1},$$

$$-mzc(z) + c(z)\operatorname{tr}\big(\Psi(I+c(z)\Psi)^{-1}\big) = m.$$

Thus, the equation becomes

$$m = \operatorname{tr}\big(c(z)\Psi(I+c(z)\Psi)^{-1}\big) - mzc(z).$$

At $z=0$, this gives

$$m = c(0)\operatorname{tr}\big(\Psi(I+c(0)\Psi)^{-1}\big),$$

such that identifying $c(0) = \nu^{-1}$ gives exactly (B.8), as we wanted. To justify the expansion near zero, observe that

$$\frac{d}{dc}\operatorname{tr}\big(c\Psi(I+c\Psi)^{-1}\big)\bigg|_{c=\nu^{-1}} = \operatorname{tr}\big(\Psi(I+\nu^{-1}\Psi)^{-2}\big) > 0.$$

The analytic implicit function theorem therefore gives a local analytic solution $c(z) = \nu^{-1} + O(z)$. Since $\kappa(z) = z/\nu + O(z^2)$ has positive imaginary part just above zero, it agrees with the branch in Definition 2. We conclude that $\kappa(z) = cz + O(z^2)$ with $c = \nu^{-1}$, such that

$$\overline{R}(z) = -\frac{1}{z}(I+\nu^{-1}\Psi)^{-1} + O(1).$$

Finally, substituting this expansion into (B.10) and integrating in $z$ gives

$$\overline{\mathcal{L}}_{\mathrm{TE}} = \langle \Psi^{1/2}\beta, (I+\nu^{-1}\Psi)^{-1}\Psi^{1/2}\beta\rangle = \nu\langle \Psi^{1/2}\beta, (\Psi+\nu I)^{-1}\Psi^{1/2}\beta\rangle.$$

as claimed. □

The next lemma gives the deterministic equivalent of the projection operator $P_0$ which also appears during the calculation of the student deterministic equivalent.

**Lemma 2** (Single-resolvent deterministic equivalent)**.** Define $\overline{P}_0 \coloneqq -\operatorname{Res}_{z=0} \overline{R}(z)$. Then

$$\overline{P}_0 = \nu(\Psi + \nu I)^{-1} = (I + \nu^{-1}\Psi)^{-1}.$$

*Proof.* This can be seen directly from the expansion obtained in the proof of Theorem 5. In fact,

$$\overline{R}(z) = -\frac{1}{z}(I + \nu^{-1}\Psi)^{-1} + O(1),$$

such that

$$\overline{P}_0 = (I + \nu^{-1}\Psi)^{-1} = \nu(\Psi + \nu I)^{-1},$$

as we wanted. □

### B.4 Double-resolvent student error deterministic equivalent

Since student training follows teacher training in a two-step procedure, the student population error has a more complicated expression that reflects this dependence. For this reason, deriving its deterministic equivalent is more intricate and requires a double-resolvent trick, as we will see below.

**Theorem 6.** Under our setting, if we define $P_0 \coloneqq \sum_{\lambda_j=0} v_j v_j^T \in \mathbb{R}^{d\times d}$ to be the projection operator onto the zero eigenspace of $\Psi^{\frac{1}{2}} U_m^T (U_m \Psi U_m^T)^{-1} U_m \Psi^{\frac{1}{2}}$ we can write

$$\mathcal{L}_{\mathrm{ST}}(t) = \left\langle \sqrt{\Psi}\beta, [e^{-2\Psi t} - (I - e^{-\Psi t})^2 P_0 + P_0(I - e^{-2\Psi t}) + P_0(I - e^{-\Psi t})^2 P_0]\sqrt{\Psi}\beta \right\rangle.$$

*Proof.* From (2.1) we have that the student neural network for a time $t > 0$ of gradient flow is given by

$$f_{\mathrm{student}}(t)(\cdot) = \sum_{k=1}^{d} (1 - e^{-\lambda_k t}) \langle f_{\mathrm{teacher}}(\cdot), e_k \rangle_{\mathcal{D}} e_k,$$

where $\lambda_k$ and $e_k$ are the eigenvalues and eigenvectors of the population random-feature kernel induced by the activation function. Thus, we can write the student population error at time $t$ as

$$\begin{aligned}
\mathcal{L}_{\mathrm{ST}}(t) &= \mathbb{E}_{x\sim\mathcal{D}}[(f_{\mathrm{student}}(t) - f^*)^2] \\
&= \mathbb{E}_{x\sim\mathcal{D}} \left[ \left( \sum_{k=1}^{d} (1 - e^{-\lambda_k t}) \langle f_{\mathrm{teacher}}, e_k \rangle e_k - f^* \right)^2 \right].
\end{aligned}$$

Now, writing $f^* = \sum_{k=1}^{d} \langle f^*, e_k \rangle_{\mathcal{D}} e_k$ and using the orthonormality of $\{e_k\}_{k=1}^{d}$, we obtain

$$\begin{aligned}
\mathcal{L}_{\mathrm{ST}}(t) &= \mathbb{E}_{x\sim\mathcal{D}} \left[ \left( \sum_{k=1}^{d} \langle (1 - e^{-\lambda_k t}) f_{\mathrm{teacher}} - f^*, e_k \rangle_{\mathcal{D}} e_k \right)^2 \right] \\
&= \mathbb{E}_{x\sim\mathcal{D}} \left[ \sum_{k=1}^{d} \langle (1 - e^{-\lambda_k t}) f_{\mathrm{teacher}} - f^*, e_k \rangle_D^2 \right] \\
&= \sum_{k\geq 1} (\mathbb{E}_{x\sim\mathcal{D}}[(1 - e^{-\lambda_k t}) f_{\mathrm{teacher}} - f^*] e_k)^2.
\end{aligned}$$

Replacing

$$f_{\mathrm{teacher}}(x) = w_* U_m x \quad \text{and} \quad f^*(x) = \beta^T x$$

we get

$$\begin{aligned}\mathcal{L}_{\mathrm{ST}}(t) &= \sum_{k=1}^{d}\left(\mathbb{E}_{x\sim\mathcal{D}}\left[\left((1-e^{-\psi_k t})w_* U_m x - \beta^T x\right)\frac{x_k}{\sqrt{\psi_k}}\right]\right)^2 \\ &= \sum_{k=1}^{d}\left(\mathbb{E}_{x\sim\mathcal{D}}\left[\left((1-e^{-\psi_k t})\sum_{l=1}^{d}(w_* U_m)_l x_l - \sum_{l=1}^{d}\beta_l^T x_l\right)\frac{x_k}{\sqrt{\psi_k}}\right]\right)^2 \\ &= \sum_{k=1}^{d}\left((1-e^{-\psi_k t})(w_* U_m)_k\sqrt{\psi_k} - \beta_k\sqrt{\psi_k}\right)^2 \\ &= \sum_{k=1}^{d}\psi_k((1-e^{-\psi_k t})(w_* U_m)_k - \beta_k)^2.\end{aligned} \tag{B.11}$$

Substituting the optimal weight

$$w_* = (U_m \Psi U_m^T)^{-1} U_m \Psi \beta$$

in (B.11) gives

$$\mathcal{L}_{\mathrm{ST}}(t) = \langle (I-e^{-\Psi t})[U_m^T(U_m\Psi U_m^T)^{-1}U_m\Psi\beta]-\beta, \Psi\{(I-e^{-\Psi t})[U_m^T(U_m\Psi U_m^T)^{-1}U_m\Psi\beta]-\beta\}\rangle.$$

Now, we can rewrite this error as

$$\begin{aligned}\mathcal{L}_{\mathrm{ST}}(t) =& \langle (I-e^{-\Psi t})\sqrt{\Psi}U_m^T(U_m\Psi U_m^T)^{-1}U_m\Psi\beta \\ & (I-e^{-\Psi t})\sqrt{\Psi}U_m^T(U_m\Psi U_m^T)^{-1}U_m\Psi\beta)\rangle \\ & -2\langle (I-e^{-\Psi t})\sqrt{\Psi}U_m^T(U_m\Psi U_m^T)^{-1}U_m\Psi\beta, \sqrt{\Psi}\beta\rangle \\ & +\langle\sqrt{\Psi}\beta, \sqrt{\Psi}\beta\rangle,\end{aligned}$$

such that

$$\begin{aligned}\mathcal{L}_{\mathrm{ST}}(t) &= \left\langle (I-e^{-\Psi t})\left(\sum_{\tilde{\lambda}_j\neq 0}\tilde{v}_j\tilde{v}_j^T\right)\sqrt{\Psi}\beta, (I-e^{-\Psi t})\left(\sum_{\tilde{\lambda}_j\neq 0}\tilde{v}_j\tilde{v}_j^T\right)\sqrt{\Psi}\beta\right\rangle \\ &\quad - 2\left\langle (I-e^{-\Psi t})\left(\sum_{\tilde{\lambda}_j\neq 0}\tilde{v}_j\tilde{v}_j^T\right)\sqrt{\Psi}\beta, \sqrt{\Psi}\beta\right\rangle + \langle\sqrt{\Psi}\beta, \sqrt{\Psi}\beta\rangle \\ &= \left\langle (I-e^{-\Psi t})\left(\sum_{\tilde{\lambda}_j\neq 0}\tilde{v}_j\tilde{v}_j^T\right)\sqrt{\Psi}\beta, \left[(I-e^{-\Psi t})\left(\sum_{\tilde{\lambda}_j\neq 0}\tilde{v}_j\tilde{v}_j^T\right) - 2I\right]\sqrt{\Psi}\beta\right\rangle + \langle\sqrt{\Psi}\beta, \sqrt{\Psi}\beta\rangle \\ &= \left\langle \sqrt{\Psi}\beta, \left[\left(\sum_{\tilde{\lambda}_j\neq 0}\tilde{v}_j\tilde{v}_j^T\right)(I-e^{-\Psi t})^2\left(\sum_{\tilde{\lambda}_j\neq 0}\tilde{v}_j\tilde{v}_j^T\right) - 2\left(\sum_{\tilde{\lambda}_j\neq 0}\tilde{v}_j\tilde{v}_j^T\right)(I-e^{-\Psi t}) + I\right]\sqrt{\Psi}\beta\right\rangle.\end{aligned}$$

Then, using $P_0 \coloneqq \sum_{\lambda_j=0} v_j v_j^T$, the term in square brackets can be rewritten as follows

$$\begin{aligned}&(I-P_0)(I-e^{-\Psi t})^2(I-P_0) - 2(I-P_0)(I-e^{-\Psi t}) + I \\ &= (I-e^{-\Psi t})^2(I-P_0) - P_0(I-e^{-\Psi t})^2(I-P_0) - 2(I-e^{-\Psi t}) + 2P_0(I-e^{-\Psi t}) + I \\ &= (I-e^{-\Psi t})^2 - (I-e^{-\Psi t})^2P_0 - P_0(I-e^{-\Psi t})^2 + P_0(I-e^{-\Psi t})^2P_0 - 2P_0(I-e^{-\Psi t}) + I \\ &= I - 2e^{-\Psi t} + e^{-2\Psi t} - (I-e^{-\Psi t})^2P_0 - P_0(I-2e^{-\Psi t}+e^{-2\Psi t}) + P_0(I-e^{-\Psi t})^2P_0 \\ &\quad - 2(I-e^{-\Psi t}) + 2P_0(I-e^{-\Psi t}) + I.\end{aligned}$$

Thus, we finally obtain

$$\mathcal{L}_{\mathrm{ST}}(t) = \left\langle \sqrt{\Psi}\beta, [e^{-2\Psi t} - (I-e^{-\Psi t})^2P_0 + P_0(I-e^{-2\Psi t}) + P_0(I-e^{-\Psi t})^2P_0]\sqrt{\Psi}\beta\right\rangle \tag{B.12}$$

as we wanted. □

For the terms in (B.12) that depend on a single $P_0$, we use the deterministic representative $\overline{P}_0$ provided by Lemma 2. For the last term, we will need a double-resolvent calculation. The idea is to insert the matrix $S \coloneqq (I - e^{-\Psi t})^2$ by differentiating a perturbed resolvent. We first define the corresponding deterministic quantities, and then calculate them.

An alternative point of view is provided by the eigenlearning framework, where second moments of a learning transfer operator are computed by spectral differentiation (Simon et al., 2023, Appendix I.9); related resolvent differentiation identities are used for random feature regression in Simon et al. (2024, Appendix E.3).

**Definition 3** (Perturbed deterministic resolvent)**.** For $S = (I - e^{-\Psi t})^2$, let

$$\overline{R}_\epsilon(z) = (\epsilon S - zI - \Psi\kappa_\epsilon(z))^{-1}, \qquad -\kappa_\epsilon(z) = \left(1 + \frac{1}{m}\operatorname{tr}\big(\Psi\overline{R}_\epsilon(z)\big)\right)^{-1}.$$

For $z$ off the deterministic spectrum, we take the branch analytic in $\epsilon$ near zero that continues Definition 2, so that $\overline{R}_0 = \overline{R}$ and $\kappa_0 = \kappa$. Define

$$\overline{Q}_S(z) \coloneqq -\left.\frac{\partial\overline{R}_\epsilon(z)}{\partial\epsilon}\right|_{\epsilon=0}, \qquad \overline{\mathcal{P}}(S) \coloneqq \frac{1}{2\pi i}\oint_\Gamma z\,\overline{Q}_S(z)\,dz,$$

where $\Gamma$ is a sufficiently small positively oriented contour around zero. These are the deterministic counterparts of $R(z)SR(z)$ and $P_0SP_0$, respectively.

The following theorem evaluates $\overline{\mathcal{P}}(S)$.

**Theorem 7** (Double-resolvent deterministic equivalent)**.** For $S = (I - e^{-\Psi t})^2$, we have

$$\overline{\mathcal{P}}(S) = \frac{(I - e^{-\Psi t})^2}{(I + \Psi\nu^{-1})^2} + \frac{\frac{\Psi}{(I+\Psi\nu^{-1})^2}\operatorname{tr}\left(\frac{\nu^{-2}\Psi(I-e^{-\Psi t})^2}{(I+\Psi\nu^{-1})^2}\right)}{m - \operatorname{tr}\left(\frac{\Psi^2\nu^{-2}}{(I+\Psi\nu^{-1})^2}\right)}$$

where $\nu$ is the solution of (B.8).

*Proof.* For $\epsilon$ near zero, let us define the following resolvent

$$R_\epsilon(z) \coloneqq \left(\frac{1}{m}\Psi^{\frac{1}{2}}U^TU\Psi^{\frac{1}{2}} + \epsilon S - zI\right)^{-1}$$

where $S \coloneqq (I - e^{-\Psi t})^2$.

Observe that

$$\partial_\epsilon R_\epsilon(z)|_{\epsilon=0} = -R_0(z)SR_0(z), \tag{B.13}$$

on the other hand, we also have that

$$R_0(z) = \sum_j \frac{v_jv_j^T}{\lambda_j - z}$$

where $\lambda_j$ are the eigenvalues of $H = \frac{1}{m}\Psi^{\frac{1}{2}}U^TU\Psi^{\frac{1}{2}}$ and $v_j$ are the associated eigenvectors. Thus

$$\begin{aligned} R_0(z)SR_0(z) &= \left(\sum_j \frac{v_jv_j^T}{\lambda_j - z}\right) S \left(\sum_k \frac{v_kv_k^T}{\lambda_k - z}\right) \\ &= \sum_{j,k}(v_jv_j^TSv_kv_k^T)\frac{1}{z^2}\left(\frac{1}{1 - (\lambda_j + \lambda_k)\frac{1}{z} + \frac{\lambda_j\lambda_k}{z^2}}\right). \end{aligned}$$

Defining $f(z) = zR_0(z)SR_0(z)$ and $\Gamma$ a small positively oriented contour around 0 enclosing no nonzero eigenvalues of $H$, consider the integral

$$\frac{1}{2\pi i}\oint_\Gamma f(z)dz = \frac{1}{2\pi i}\oint_\Gamma \sum_{j,k}\frac{v_jv_j^TSv_kv_k^T}{z}\left(\frac{1}{1 - (\lambda_j + \lambda_k)\frac{1}{z} + \frac{\lambda_j\lambda_k}{z^2}}\right)dz.$$

If at least one of $\lambda_j$ and $\lambda_k$ is nonzero, the corresponding summand of $f(z)$ is holomorphic inside $\Gamma$ and by Cauchy's theorem its integral is 0. Thus, only the terms with $\lambda_j = \lambda_k = 0$ remain, and we have

$$\frac{1}{2\pi i}\oint \sum_{\substack{\lambda_j=0\\ \lambda_k=0}} \frac{v_j v_j^T S v_k v_k^T}{z} dz = P_0 S P_0$$

since

$$\frac{1}{2\pi i}\oint_\Gamma \frac{1}{z} dz = 1.$$

Then, we get

$$\frac{1}{2\pi i}\oint_\Gamma z R_0(z) S R_0(z) dz = P_0 S P_0. \tag{B.14}$$

We now use a linearization procedure to recover the Dyson equation defining $\overline{R}_\epsilon(z)$. We define the larger matrix

$$L \coloneqq \begin{pmatrix} \epsilon S - zI_d & \frac{1}{\sqrt{m}}\Psi^{\frac{1}{2}}U^T \\ \frac{1}{\sqrt{m}}U\Psi^{\frac{1}{2}} & -I_m \end{pmatrix} \in \mathbb{C}^{(d+m)\times(d+m)}.$$

Using the Schur complement formula, we obtain

$$(L^{-1})_{11} = \left(\frac{1}{m}\Psi^{\frac{1}{2}}U^T U\Psi^{\frac{1}{2}} + \epsilon S - zI\right)^{-1}.$$

Now, by the Dyson equation

$$\begin{pmatrix} \mathcal{M}_{11} & 0 \\ 0 & \mathcal{M}_{22} \end{pmatrix}\left(\begin{pmatrix} \epsilon S - zI & 0 \\ 0 & -I_m \end{pmatrix} - \begin{pmatrix} \frac{1}{m}\Psi \operatorname{tr}(\mathcal{M}_{22}) & 0 \\ 0 & \frac{1}{m}\operatorname{tr}(\mathcal{M}_{11}\Psi) I_m \end{pmatrix}\right) = \begin{pmatrix} I & 0 \\ 0 & I_m \end{pmatrix},$$

we get

$$\begin{aligned} \mathcal{M}_{11} &= \left(\epsilon S - zI - \frac{1}{m}\Psi \operatorname{tr}(\mathcal{M}_{22})\right)^{-1} \\ \mathcal{M}_{22} &= \left(-1 - \frac{1}{m}\operatorname{tr}(\mathcal{M}_{11}\Psi)\right)^{-1} I_m \end{aligned} \tag{B.15}$$

and $\mathcal{M}_{11} = \overline{R}_\epsilon(z)$ is exactly the matrix in Definition 3. In our case, $S = (I - e^{-\Psi t})^2$ and $\Psi$ share the same orthonormal basis of eigenvectors $\{w_j\}$. Let $\lambda_j^S$ and $\lambda_j^\Psi$ be the associated eigenvalues, then

$$\begin{aligned} \langle w_j, \mathcal{M}_{11} w_j\rangle &= \left(\epsilon\lambda_j^S - z - \frac{\lambda_j^\Psi}{m}\operatorname{tr}(\mathcal{M}_{22})\right)^{-1} \\ \mathcal{M}_{22} &= \left(-1 - \frac{1}{m}\operatorname{tr}(\mathcal{M}_{11}\Psi)\right)^{-1} I_m. \end{aligned}$$

To ease notation in the calculation below, write $\kappa(z) = \kappa_\epsilon(z)$, suppressing its dependence on $\epsilon$. Then

$$\kappa(z) = \frac{1}{m}\operatorname{tr}(\mathcal{M}_{22}),$$

then

$$\langle w_j, \mathcal{M}_{11} w_j\rangle = \frac{1}{\epsilon\lambda_j^S - z - \lambda_j^\Psi \kappa(z)}. \tag{B.16}$$

Note that

$$\operatorname{tr}(\mathcal{M}_{22}) = m\left(-1 - \frac{1}{m}\operatorname{tr}(\mathcal{M}_{11}\Psi)\right)^{-1}$$

such that by (B.15) we have

$$\kappa(z) = \frac{1}{-1 - \frac{1}{m}\sum_{j=1}^d \langle w_j, \mathcal{M}_{11} w_j\rangle \lambda_j^\Psi}.$$

Also, from (B.16) we get

$$\mathcal{M}_{11} = \sum_{j=1}^{d} w_j w_j^T \left( \frac{1}{\epsilon\lambda_j^S - z - \lambda_j^\Psi \kappa(z)} \right),$$

such that we have the following scalar self-consistent equation

$$\kappa(z) = \frac{1}{-1 - \frac{1}{m}\sum_{j=1}^{d} \frac{\lambda_j^\Psi}{\epsilon\lambda_j^S - z - \lambda_j^\Psi \kappa(z)}}.$$

The exact identities (B.13) and (B.14) motivate the definitions of $\overline{Q}_S$ and $\overline{\mathcal{P}}(S)$. Since $\mathcal{M}_{11} = \overline{R}_\epsilon$, Definition 3 gives

$$\overline{\mathcal{P}}(S) = -\frac{1}{2\pi i} \oint_\Gamma z \left. \frac{\partial \mathcal{M}_{11}}{\partial \epsilon} \right|_{\epsilon=0} dz.$$

Calculating, we get

$$\frac{\partial \mathcal{M}_{11}}{\partial \epsilon} = -\sum_{j=1}^{d} \frac{w_j w_j^T}{(\epsilon\lambda_j^S - z - \lambda_j^\Psi \kappa(z))^2} \left( \lambda_j^S - \lambda_j^\Psi \frac{\partial \kappa(z)}{\partial \epsilon} \right).$$

Also,

$$\frac{\partial \kappa(z)}{\partial \epsilon} = \frac{-1}{\left(-1 - \frac{1}{m}\sum_{j=1}^{d} \frac{\lambda_j^\Psi}{\epsilon\lambda_j^S - z - \lambda_j^\Psi \kappa(z)}\right)^2} \left( \frac{1}{m} \sum_{j=1}^{d} \frac{\lambda_j^\Psi (\lambda_j^S - \lambda_j^\Psi \frac{\partial \kappa(z)}{\partial \epsilon})}{(\epsilon\lambda_j^S - z - \lambda_j^\Psi \kappa(z))^2} \right),$$

such that

$$\left. \frac{\partial \kappa(z)}{\partial \epsilon} \right|_{\epsilon=0} = \left. \frac{-\frac{1}{m} \operatorname{tr}\left( \frac{\kappa(z)^2 \Psi S}{(z+\Psi\kappa(z))^2} \right)}{1 - \frac{1}{m} \operatorname{tr}\left( \frac{\Psi^2 \kappa(z)^2}{(z+\Psi\kappa(z))^2} \right)} \right|_{\epsilon=0}.$$

Then

$$\overline{\mathcal{P}}(S) = \frac{1}{2\pi i} \oint_\Gamma \sum_{j=1}^{d} \frac{z w_j w_j^T}{(z + \lambda_j^\Psi \kappa(z))^2} \left[ \lambda_j^S + \frac{\frac{\lambda_j^\Psi}{m} \operatorname{tr}\left( \frac{\kappa(z)^2 \Psi (I - e^{-\Psi t})^2}{(z+\Psi\kappa(z))^2} \right)}{1 - \frac{1}{m} \operatorname{tr}\left( \frac{\Psi^2 \kappa(z)^2}{(z+\Psi\kappa(z))^2} \right)} \right]_{\epsilon=0} dz. \tag{B.17}$$

At $\epsilon = 0$, the proof of Theorem 5 gives $\kappa(z) = cz + O(z^2)$ with $c = \nu^{-1}$. Observe also that the denominator at zero is strictly positive:

$$m - \operatorname{tr}\left( \frac{c^2 \Psi^2}{(I + c\Psi)^2} \right) = \operatorname{tr}\left( \frac{c\Psi}{(I + c\Psi)^2} \right) > 0,$$

where we used (B.8). Thus, keeping the coefficient of $z^{-1}$ in the integrand, we can rewrite (B.17) as follows

$$\begin{aligned} \overline{\mathcal{P}}(S) &= \frac{1}{2\pi i} \oint_\Gamma \sum_{j=1}^{d} \frac{w_j w_j^T}{z(1 + \Psi_j c)^2} \left[ (1 - e^{-\Psi_j t})^2 + \frac{\frac{\Psi_j}{m} \operatorname{tr}\left( \frac{c^2 \Psi (I - e^{-\Psi t})^2}{(I+\Psi c)^2} \right)}{1 - \frac{1}{m} \operatorname{tr}\left( \frac{\Psi^2 c^2}{(I+\Psi c)^2} \right)} \right] dz \\ &= \frac{(I - e^{-\Psi t})^2}{(I + \Psi c)^2} + \frac{\frac{\Psi}{(I+\Psi c)^2} \operatorname{tr}\left( \frac{c^2 \Psi (I - e^{-\Psi t})^2}{(I+\Psi c)^2} \right)}{m - \operatorname{tr}\left( \frac{\Psi^2 c^2}{(I+\Psi c)^2} \right)} \end{aligned}$$

as claimed. □

The two terms in Theorem 7 come from differentiating the explicit perturbation $\epsilon S$ and the self-consistent parameter $\kappa_\epsilon$, respectively. We can now combine this calculation with the single-resolvent formula to define the deterministic-equivalent student error.

**Definition 4** (Deterministic-equivalent student error)**.** Let $S_t = (I - e^{-\Psi t})^2$. Replacing the single-projector terms and the double-projector term in (B.12) by their deterministic counterparts, define

$$\overline{\mathcal{L}}_{\mathrm{ST}}(t) \coloneqq \left\langle \Psi^{1/2}\beta, \left[ e^{-2\Psi t} - S_t \overline{P}_0 + \overline{P}_0 (I - e^{-2\Psi t}) + \overline{\mathcal{P}}(S_t) \right] \Psi^{1/2}\beta \right\rangle, \tag{B.18}$$

where $\overline{P}_0 = \nu(\Psi + \nu I)^{-1}$ and $\overline{\mathcal{P}}(S_t)$ is evaluated in Theorem 7.

### B.5 BIAS–VARIANCE DECOMPOSITION

We first verify (3.3) directly from Definition 4, then give its moment interpretation.

**Direct algebraic verification.** The fixed-point equation gives

$$\begin{aligned} m - \operatorname{tr}[\Psi^2(\Psi+\nu I)^{-2}] &= \operatorname{tr}[\Psi(\Psi+\nu I)^{-1} - \Psi^2(\Psi+\nu I)^{-2}] \\ &= \nu \operatorname{tr}[\Psi(\Psi+\nu I)^{-2}] > 0. \end{aligned}$$

Using $(I+\Psi\nu^{-1})^{-2} = \nu^2(\Psi+\nu I)^{-2}$, Theorem 7 evaluated at $S = S_t = (I-e^{-\Psi t})^2$ therefore yields

$$\begin{aligned} \overline{\mathcal{P}}(S_t) = \nu^2(\Psi+\nu I)^{-2}(I-e^{-\Psi t})^2 \\ + \nu\Psi(\Psi+\nu I)^{-2}\frac{\operatorname{tr}[\Psi(\Psi+\nu I)^{-2}(I-e^{-\Psi t})^2]}{\operatorname{tr}[\Psi(\Psi+\nu I)^{-2}]}. \end{aligned}$$

The second term gives exactly $\overline{V}(t)$ after taking the quadratic form against $\Psi^{1/2}\beta$. The remaining terms in (B.18) combine using

$$I - e^{-2\Psi t} - (I-e^{-\Psi t})^2 = 2e^{-\Psi t}(I-e^{-\Psi t}).$$

Since all the matrices below are functions of $\Psi$ and hence commute, their sum is

$$\begin{aligned} & e^{-2\Psi t} + 2\nu e^{-\Psi t}(I-e^{-\Psi t})(\Psi+\nu I)^{-1} + \nu^2(I-e^{-\Psi t})^2(\Psi+\nu I)^{-2} \\ &= \left[e^{-\Psi t} + \nu(I-e^{-\Psi t})(\Psi+\nu I)^{-1}\right]^2 \\ &= \left[(e^{-\Psi t}(\Psi+\nu I) + \nu(I-e^{-\Psi t}))(\Psi+\nu I)^{-1}\right]^2 \\ &= (\nu I + \Psi e^{-\Psi t})^2(\Psi+\nu I)^{-2}. \end{aligned}$$

Its quadratic form against $\Psi^{1/2}\beta$ is

$$\left\|(\nu I + \Psi e^{-\Psi t})(\Psi+\nu I)^{-1}\Psi^{1/2}\beta\right\|^2 = \overline{B}(t),$$

which proves (3.3).

**Moment interpretation.** Here the bias and variance refer to the student prediction as an estimator of $f^*$, with respect to the randomness of the teacher features $\mathcal{U}$. The target and the time $t$ are fixed, while different draws of $\mathcal{U}$ produce different teacher labels and hence different student predictions. Recall that $P_0$ is the orthogonal projector onto the nullspace of $H$, and that the student residual is $[e^{-\Psi t} + (I-e^{-\Psi t})P_0]\Psi^{1/2}\beta$. Its squared Euclidean norm equals the squared $L^2(\mathcal{D})$ norm of $f^* - f_{\text{student}}(t)$. Averaging over the teacher features gives the exact identity

$$\begin{aligned} & \mathbb{E}_{\mathcal{U}}\|[e^{-\Psi t} + (I-e^{-\Psi t})P_0]\Psi^{1/2}\beta\|^2 \\ &\quad = \|[e^{-\Psi t} + (I-e^{-\Psi t})\mathbb{E}_{\mathcal{U}}[P_0]]\Psi^{1/2}\beta\|^2 \\ &\qquad + \mathbb{E}_{\mathcal{U}}\|(I-e^{-\Psi t})(P_0 - \mathbb{E}_{\mathcal{U}}[P_0])\Psi^{1/2}\beta\|^2. \end{aligned}$$

The first term is the squared norm of the average residual. Since $f^*$ is deterministic,

$$\begin{aligned} & \|[e^{-\Psi t} + (I-e^{-\Psi t})\mathbb{E}_{\mathcal{U}}[P_0]]\Psi^{1/2}\beta\|^2 \\ &\quad = \|f^* - \mathbb{E}_{\mathcal{U}}[f_{\text{student}}(t)]\|_{\mathcal{D}}^2 \\ &\quad = \mathbb{E}_{x\sim\mathcal{D}}\left[\left(\mathbb{E}_{\mathcal{U}}[f_{\text{student}}(t)(x)] - f^*(x)\right)^2\right]. \end{aligned}$$

For each fixed $x$, the difference $\mathbb{E}_{\mathcal{U}}[f_{\text{student}}(t)(x)] - f^*(x)$ is the bias of the student prediction as an estimator of $f^*(x)$. Thus, the first term is the squared bias averaged over the input distribution.

The second term is the expected squared norm of the residual minus its mean. Centering the residual cancels the target:

$$\begin{aligned} & (f^* - f_{\text{student}}(t)) - \mathbb{E}_{\mathcal{U}}[f^* - f_{\text{student}}(t)] \\ &\quad = \mathbb{E}_{\mathcal{U}}[f_{\text{student}}(t)] - f_{\text{student}}(t). \end{aligned}$$

Consequently,

$$\begin{aligned}&\mathbb{E}_{\mathcal{U}}\|(I-e^{-\Psi t})(P_0-\mathbb{E}_{\mathcal{U}}[P_0])\Psi^{1/2}\beta\|^2\\&\quad=\mathbb{E}_{\mathcal{U}}\|f_{\text{student}}(t)-\mathbb{E}_{\mathcal{U}}[f_{\text{student}}(t)]\|_{\mathcal{D}}^2\\&\quad=\mathbb{E}_{x\sim\mathcal{D}}\left[\mathrm{Var}_{\mathcal{U}}(f_{\text{student}}(t)(x))\right].\end{aligned}$$

Thus, the second term is the variance of the student prediction across draws of the teacher features, averaged over the input distribution.

To identify these moments with $\overline{B}(t)$ and $\overline{V}(t)$, the deterministic equivalents must approximate the relevant first and second moments at the claimed asymptotic scale. Under this additional control, replacing $\mathbb{E}_{\mathcal{U}}[P_0]$ by $\nu(\Psi+\nu I)^{-1}$ in the squared mean residual gives $\overline{B}(t)$ by the algebra above.

For the variance, we first write the exact identity

$$\begin{aligned}&\mathbb{E}_{\mathcal{U}}\|(I-e^{-\Psi t})(P_0-\mathbb{E}_{\mathcal{U}}[P_0])\Psi^{1/2}\beta\|^2\\&\quad=\mathbb{E}_{\mathcal{U}}\langle\Psi^{1/2}\beta,P_0(I-e^{-\Psi t})^2P_0\Psi^{1/2}\beta\rangle\\&\qquad-\|(I-e^{-\Psi t})\mathbb{E}_{\mathcal{U}}[P_0]\Psi^{1/2}\beta\|^2.\end{aligned}$$

The first term on the right-hand side is approximated by the quadratic form of $\overline{\mathcal{P}}(S_t)$, against $\Psi^{1/2}\beta$. Replacing $\mathbb{E}_{\mathcal{U}}[P_0]$ by $\nu(\Psi+\nu I)^{-1}$ in the subtracted squared mean gives precisely the quadratic form of the first term of $\overline{\mathcal{P}}(S_t)$. These contributions therefore cancel, leaving the quadratic form of its second term, which is $\overline{V}(t)$. As above, this identification requires the relevant moment approximations; convergence in probability of the loss alone does not establish them.

# C Proofs for ReLU activation

**Scope.** The asymptotic loss calculations in this appendix concern the deterministic equivalents. Transfer to the random spherical losses uses Assumption 1. Lemma 4 gives the order of $\nu$, which Appendix E reuses for the two mixed-target examples.

The kernel (4.1) contains the degree-zero mode, a single direction with $N_0 = 1$ and coefficient $\sigma_0 = \mu_d = \mathbb{E}\left(\langle u, x\rangle\right)_+ = \Gamma(d/2)/(2\sqrt{\pi}\,\Gamma((d+1)/2)) \asymp d^{-1/2}$, the population mean of a feature. It is included in the sums and traces below, which run over $k \in \mathcal{K} = \{0,1\} \cup 2\mathbb{N}$.

In this appendix we calculate the deterministic equivalents of the teacher and student model in the setting of Theorem 2. From (4.1), we see that the important quantities to calculate are the ReLU coefficients in spherical harmonics $\sigma_k$ and the dimension of the space of spherical harmonics of degree $k$, namely $N_k$. Thus, we start calculating their the asymptotics.

## C.1 Asymptotics of $N_k$ and $\sigma_k$

The explicit value of $N_k$ can be found, for example, in Appendix $D.1$ of Bach (2017). For a $d$-dimensional sphere $\mathbb{S}^{d-1}$, we have

$$N_k = \frac{2k+d-2}{k}\binom{k+d-3}{d-2}. \tag{C.1}$$

for $k \geq 1$ integer. For $k = 0$, the basis element is the constant function such that $N_0 = 1$.

Expanding (C.1), we obtain

$$\begin{aligned}N_k &= \frac{2k+d-2}{k}\frac{(k+d-3)(k+d-4)...(d-2)!}{(k-1)!(d-2)!}\\&= \frac{2k+d-2}{k!}(d-2+(k-1))...(d-1)\end{aligned}$$

that is

$$N_k = \frac{(2k+d-2)(d-2+(k-1))(d-2+(k-2))...(d-1)}{k!}. \tag{C.2}$$

The expression for $\sigma_k$ can be found in Appendix $A.1$ of Petrini et al. (2022). For $k > 1$ we get

$$\sigma_k = \frac{-\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{d}{2}\right)} \frac{\Gamma\left(\frac{d-1}{2}\right)}{\Gamma\left(k+\frac{d-1}{2}\right)} \left(-\frac{1}{2}\right)^k \frac{d^{k-2}}{dt^{k-2}}(1-t^2)^{k+\frac{d-3}{2}}\Big|_{t=0}^{t=1}.$$

For $k = 0$ we have another expression which is

$$\begin{aligned}
\sigma_0 &= \frac{\Gamma(d/2)}{\sqrt{\pi}\,\Gamma((d-1)/2)} \int_0^1 t(1-t^2)^{\frac{d-3}{2}}\, dt \\
&= \frac{\Gamma(d/2)}{2\sqrt{\pi}\,\Gamma((d-1)/2)} \int_0^1 u^{\frac{d-3}{2}}\, du \\
&= \frac{\Gamma(d/2)}{2\sqrt{\pi}\,\Gamma((d-1)/2)} \left[\frac{2}{d-1} u^{\frac{d-1}{2}}\right]_0^1 \\
&= \frac{\Gamma(d/2)}{(d-1)\sqrt{\pi}\,\Gamma((d-1)/2)} \\
&= \frac{\Gamma(d/2)}{2\sqrt{\pi}\,\Gamma((d+1)/2)} \\
&\asymp \frac{1}{\sqrt{d}}
\end{aligned} \tag{C.3}$$

For $k = 1$ we have a similar expression

$$\sigma_1 = \frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\Gamma(\frac{d-1}{2})} \int_0^1 t^2(1-t^2)^{\frac{d-3}{2}} dt.$$

Using the change of variables $u = t^2$ such that $du = 2tdt$ and we get for the integral

$$\begin{aligned}
\int_0^1 \frac{u(1-u)^{\frac{d-3}{2}}}{2\sqrt{u}} du &= \frac{1}{2}\int_0^1 u^{\frac{1}{2}}(1-u)^{\frac{d-3}{2}} du \\
&= \frac{1}{2} B\left(\frac{3}{2}, \frac{d-1}{2}\right) = \frac{1}{2}\frac{\Gamma\left(\frac{3}{2}\right)\Gamma\left(\frac{d-1}{2}\right)}{\Gamma\left(\frac{d}{2}+1\right)} \\
&= \frac{\sqrt{\pi}}{4}\frac{\Gamma\left(\frac{d-1}{2}\right)}{\frac{d}{2}\Gamma\left(\frac{d}{2}\right)} = \frac{\sqrt{\pi}}{2d}\frac{\Gamma\left(\frac{d-1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}
\end{aligned}$$

where we have written the integral in terms of a Beta function. Then, we obtain a simple expression

$$\sigma_1 = \frac{1}{2d}. \tag{C.4}$$

Substituting $k = 1$ in (C.1) we conclude that

$$N_1\sigma_1^2 = \frac{d}{4d^2} = \frac{1}{4d}. \tag{C.5}$$

For the derivative appearing in the expression of $\sigma_k$ for $k > 1$ we claim the following.

**Lemma 3.** We have

$$\frac{d^{k-2}}{dt^{k-2}}(1-t^2)^{k+\frac{d-3}{2}}\Big|_{t=0}^{t=1} = \begin{cases} (-1)^{\frac{k}{2}}\binom{k+\frac{d-3}{2}}{\frac{k-2}{2}}(k-2)!, & k \text{ even} \\ 0, & k \text{ odd} \end{cases}$$

*Proof.* Using the generalized binomial theorem and denoting $a \coloneqq k + \frac{d-3}{2}$ we have that for $t = 0$

$$(1-t^2)^a = \sum_{l=0}^{\infty} \binom{a}{l}(-1)^l t^{2l}$$

.

If $2l < k-2$, the $(k-2)-$th derivative is 0 because it is the derivative of a constant. If $2l > k-2$, the $(k-2)-$th derivative is also 0 since we will have $t^s$ for $s > 0$ evaluated at $t = 0$. We conclude

$$-\frac{d^{k-2}}{dt^{k-2}}(1-t^2)^a = -\binom{a}{\frac{k-2}{2}}(-1)^{\frac{k-2}{2}}\left[2\left(\frac{k-2}{2}\right)\right]! \\ = \binom{a}{\frac{k}{2}}(-1)^{\frac{k-2}{2}}(k-2)!$$

Since we only have the term $2l = k-2$ we necessarily have that $k$ is even. Moreover, for $t-1$, we observe that all terms in $\frac{d^{k-2}}{dt^{k-2}}$ will contain a factor of $(1-t^2)^{\frac{d-3}{2}+b}$ where $b \geq 2$ such that in $t = 1$ this will be evaluated to be zero. This concludes the proof. □

**Proposition 1.** For $d \gg 1$ and $k \geq 2$ we have the asymptotics

$$N_k\sigma_k^2 \asymp \begin{cases} \dfrac{\sqrt{d}\ (2k+d)}{2^k k!(d+k)^{5/2}}, & \text{if } k = 2, \\ \dfrac{\sqrt{d}\ (2k+d)}{2^k k!(d+k)^{5/2}}\left[(k-2)(k-3)\cdots\left(\dfrac{k}{2}\right)\right]^2, & \text{if } k > 2. \end{cases}$$

*Proof.* Since we are always considering the setting where $d \gg 1$ we can use the asymptotics of the Gamma function to get the following simplication

$$\frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{d-1}{2}\right)} = \frac{\Gamma(\frac{d-1}{2}+\frac{1}{2})}{\sqrt{\pi}\Gamma\left(\frac{d-1}{2}\right)} \asymp \frac{\Gamma\left(\frac{d-1}{2}\right)\left(\frac{d-1}{2}\right)^{\frac{1}{2}}}{\sqrt{\pi}\Gamma\left(\frac{d-1}{2}\right)} \asymp \sqrt{d}$$

and

$$\frac{\Gamma\left(\frac{d-1}{2}\right)}{\Gamma\left(\frac{d-1}{2}+k\right)} = \frac{1}{\left(\frac{d-1}{2}+(k-1)\right)...\left(\frac{d-1}{2}\right)}.$$

Putting everything together we obtain the following expression for $\sigma_k$ when $k > 1$ in the regime where $d$ is very large

$$\sigma_k \asymp \frac{(-1)^{\frac{k}{2}+1}}{2^k}\frac{\sqrt{d}}{\left(\frac{d-1}{2}+(k-1)\right)...\left(\frac{d-1}{2}\right)}\binom{k+\frac{d-3}{2}}{\frac{k-2}{2}}(k-2)! \tag{C.6}$$

for $k > 1$ even and $\sigma_k = 0$ for $k > 1$ odd.

Using (C.2) and (C.6) we conclude that for $k > 1$ even

$$N_k\sigma_k^2 \asymp \frac{(2k+d-2)(d-2+(k-1))(d-2+(k-2))...(d-1)}{k!} \\ \cdot\frac{d}{4^k}\frac{1}{\left(\frac{d-1}{2}+(k-1)\right)^2\left(\frac{d-1}{2}+(k-2)\right)^2...\left(\frac{d-1}{2}\right)^2}\left(\frac{(k+\frac{d-3}{2})!}{(\frac{k+d-1}{2})!\,(\frac{k-2}{2})!}\right)^2((k-2)!)^2$$

Observing that

$$\left(\frac{d-1}{2}+(k-1)\right)\left(\frac{d-1}{2}+(k-2)\right)...\left(\frac{d-1}{2}\right) = \frac{\left(k+\frac{d-3}{2}\right)!}{\left(\frac{d-3}{2}\right)!}$$

we get

$$N_k\sigma_k^2 \asymp \frac{(2k+d-2)(d-2+(k-1))(d-2+(k-2))...(d-1)}{k!} \\ \cdot\frac{d}{4^k}\left(\frac{\left(\frac{d-3}{2}\right)!\,(k-2)!}{(\frac{k+d-1}{2})!\,(\frac{k-2}{2})!}\right)^2. \tag{C.7}$$

Notice that

$$(d-2+(k-1))(d-2+(k-2))...(d-1) = \frac{(d-3+k)!}{(d-2)!}$$

and that

$$\frac{(d-3+k)!}{\left(\frac{k+d-1}{2}\right)!} = (d+k-3)(d+k-4)...\left(\frac{k+d-1}{2}+1\right)$$
$$= (d+k-3)(d+k-4)...\left(\frac{d+k+1}{2}\right)$$

also

$$\frac{(k-2)!}{\left(\frac{k-2}{2}\right)!} = \begin{cases}(k-2)(k-3)...\left(\frac{k}{2}\right), \text{ if } k>2\\ 1, \text{ if } k=2.\end{cases}$$

such that substituting in the expression for $N_k\sigma_k^2$ and separating the terms depending only on $k$, only on $d$ or on both, we get for $k>2$ even

$$N_k\sigma_k^2 \asymp \frac{(2k+d-2)(d+k-3)(d+k-4)...(\frac{d+k+1}{2})}{\left(\frac{k+d-1}{2}\right)!}$$
$$\cdot \frac{d\left[\left(\frac{d-3}{2}\right)!\right]^2}{(d-2)!}\frac{[(k-2)(k-3)...(\frac{k}{2})]^2}{4^k k!}$$

and for $k=2$

$$N_k\sigma_k^2 \asymp \frac{(2k+d-2)(d+k-3)(d+k-4)...(\frac{d+k+1}{2})}{\left(\frac{k+d-1}{2}\right)!}\frac{d\left[\left(\frac{d-3}{2}\right)!\right]^2}{(d-2)!4^k k!}.$$

Now, since $d$ is very large we will use Stirling approximation on the terms that contains only $d$ or that contains both as follows

$$\frac{d\left[\left(\frac{d-3}{2}\right)\right]^2}{(d-2)!} \asymp \frac{d\left[\left(\frac{d-3}{2}\right)^{\frac{1}{2}}\left(\frac{d-3}{2e}\right)^{\frac{d-3}{2}}\right]^2}{(d-2)^{\frac{1}{2}}(\frac{d-2}{e})^{d-2}}$$
$$\asymp \frac{d\cdot d^{1+d-3}}{d^{\frac{1}{2}+d-2}}\frac{e^{d-2}}{(2e)^{d-3}} \asymp \frac{d^{\frac{1}{2}}}{2^d}$$

and

$$\frac{(2k+d-2)(d+k-3)(d+k-4)...\left(\frac{d+k+1}{2}\right)}{\left(\frac{k+d-1}{2}\right)!} \asymp \frac{(2k+d-2)(d+k-3)!}{[\left(\frac{d+k-1}{2}\right)]^2}$$
$$\asymp \frac{(2k+d)(d+k-3)^{\frac{1}{2}}\left(\frac{d+k-3}{e}\right)^{d+k-3}}{\left[\left(\frac{d+k-1}{2}\right)^{\frac{1}{2}}\left(\frac{d+k-1}{2e}\right)^{\frac{d+k-1}{2}}\right]^2}$$
$$\asymp (2k+d)(d+k)^{-\frac{5}{2}}2^{d+k}.$$

Putting it all together we get

$$N_k\sigma_k^2 \asymp \begin{cases}\frac{\sqrt{d}\ (2k+d)}{2^k k!(d+k)^{5/2}}, & \text{if } k=2,\\ \frac{\sqrt{d}\ (2k+d)}{2^k k!(d+k)^{5/2}}\left[(k-2)(k-3)\cdots\left(\frac{k}{2}\right)\right]^2, & \text{if } k>2.\end{cases} \tag{C.8}$$

as we wanted. □

It will actually be useful later to have the asymptotic expressions for $N_k$ and $\sigma_k^2$ separately. For $N_k$, using Stirling approximation again, we get

$$N_k = \frac{(2k+d-2)(d+k-3)!}{k!(d-2)!}$$
$$\asymp \frac{(2k+d)(d+k)^{\frac{1}{2}}\left(\frac{d+k}{e}\right)^{d+k-3}}{k!d^{\frac{1}{2}}\left(\frac{d}{e}\right)^{d-2}} \tag{C.9}$$
$$\asymp \frac{(2k+d)d^{\frac{3}{2}-d}(d+k)^{d+k-\frac{5}{2}}}{k!e^k}$$

Now, for $\sigma_k^2$, from the second line of (C.7) and using Stirling again, we have for $k > 2$

$$\begin{aligned}\sigma_k^2 &= \frac{d\left[\left(\frac{d-3}{2}\right)!\right]^2}{\left[\left(\frac{d+k-1}{2}\right)!\right]^2}\frac{[(k-2)(k-3)...(k/2)]^2}{4^k} \\ &\asymp \frac{d\left[d^{\frac{1}{2}}\left(\frac{d}{2e}\right)^{\frac{d-3}{2}}\right]^2}{\left[(d+k)^{\frac{1}{2}}\left(\frac{d+k}{2e}\right)^{\frac{d+k-1}{2}}\right]^2}\frac{[(k-2)(k-3)...(k/2)]^2}{4^k} \\ &\asymp \frac{d^{d-1}}{(d+k)^{d+k}}\left(\frac{e}{2}\right)^k[(k-2)(k-3)...(k/2)]^2\end{aligned} \tag{C.10}$$

and for $k = 2$ the following

$$\sigma_2^2 \asymp \frac{d^{d-1}}{(d+k)^{d+k}}\left(\frac{e}{2}\right)^k. \tag{C.11}$$

### C.2 Scaling of $\nu$

Again, since $\nu$ is the key deterministic quantity that appears in the errors we start by calculating its asymptotic behavior.

**Lemma 4 (Order of magnitude of $\nu$ – ReLU activation).** Under our setting, we have

$$\nu = \Theta(d^{-\alpha-1}).$$

*Proof.* We use again (B.8) from Theorem 5, that is,

$$m = \operatorname{tr}\left(\Psi(\Psi+\nu)^{-1}\right)$$

where now $m = M_{\mathrm{TE}} = d^\alpha$.

Calculating

$$d^\alpha = \operatorname{tr}\left(\frac{\Psi}{\Psi+\nu}\right) = \sum_{k\in\mathcal{K}}\frac{N_k\sigma_k^2}{\sigma_k^2+\nu}.$$

since the eigenvalues of $\Psi$ are $\sigma_k^2$ and there are $N_k$ of them.

We want to find upper and lower bounds for $\nu$ to find its order of magnitude with respect to $d$ in the regime where $d$ is very large. Let us separate the terms for $k = 1$ and $k = 2$ from the summation to get

$$d^\alpha = \frac{N_0\sigma_0^2}{\sigma_0^2+\nu} + \frac{N_1\sigma_1^2}{\sigma_1^2+\nu} + \frac{N_2\sigma_2^2}{\sigma_2^2+\nu} + \sum_{\substack{k\in\mathcal{K}\\k\geq 4}}\frac{N_k\sigma_k^2}{\sigma_k^2+\nu}$$

For $k = 0$, we get

$$\frac{N_0\sigma_0^2}{\sigma_0^2+\nu} \asymp \frac{d^{-1}}{d^{-1}+\nu} = \frac{1}{1+d\nu}.$$

For $k = 1$, using (C.4) and (C.5) we obtain

$$\begin{aligned}\frac{N_1\sigma_1^2}{\sigma_1^2+\nu} &= \frac{\frac{1}{4d}}{\left(\frac{1}{2d}\right)^2+\nu} \\ &= \frac{4d^2}{1+4d^2\nu}\frac{1}{4d} \\ &= \frac{d}{1+4d^2\nu}.\end{aligned}$$

Now, for $k = 2$ we use (C.8) and (C.11) to get

$$\frac{N_2\sigma_2^2}{\sigma_2^2+\nu} = \frac{\sqrt{d}(4+d)}{8(d+2)^{\frac{5}{2}}}\frac{1}{\frac{e^2}{4}d^{d-1}(d+2)^{-d-2}+\nu}$$

Finally, for the $k \geq 4$ even

$$\begin{aligned}\sum_{\substack{k\in\mathcal{K}\\k\geq 4}} \frac{N_k\sigma_k^2}{\sigma_k^2+\nu} &= \frac{\sqrt{d}(2k+d)[(k-2)(k-3)...(k/2)]^2}{2^k k!(d+k)^{\frac{5}{2}}\left[\frac{d^{d-1}e^k}{(d+k)^{d+k}2^k}[(k-2)(k-3)...(k/2)]^2+\nu\right]}\\ &= \frac{\sqrt{d}(2k+d)[(k-2)(k-3)...(k/2)]^2}{k!(d+k)^{\frac{5}{2}-d-k}d^{d-1}e^k[(k-2)(k-3)...(k/2)]^2+2^k k!(d+k)^{\frac{5}{2}}\nu}\\ &= \frac{\sqrt{d}(2k+d)(d+k)^{d+k}[(k-2)(k-3)...(k/2)]^2}{k!(d+k)^{\frac{5}{2}}d^{d-1}e^k[(k-2)(k-3)...(k/2)]^2+2^k k!(d+k)^{d+k+\frac{5}{2}}\nu}.\end{aligned}$$

**Upper bound on $\nu$.** In this case we argue that the summation in $k$ is not large.

Ignoring some positive terms in the denominators we can bound

$$\begin{aligned}d^\alpha &\lesssim \frac{1}{d\nu}+\frac{d}{4d^2\nu}+\frac{\sqrt{d}(4+d)}{8(d+2)^{\frac{5}{2}}\nu}+\sum_{\substack{k\in\mathcal{K}\\k\geq 4}}\frac{\sqrt{d}(2k+d)[(k-2)(k-3)...(\frac{k}{2})]^2}{2^k k!(d+k)^{\frac{5}{2}}\nu}\\ &\lesssim \frac{1}{d\nu}+\frac{1}{4d\nu}+\frac{1}{4d\nu}+\frac{\sqrt{d}}{d^{\frac{3}{2}}\nu}\sum_{\substack{k\in\mathcal{K}\\k\geq 4}}\frac{[(k-2)(k-3)...(k/2)]^2}{2^{k-1}k!}.\end{aligned} \tag{C.12}$$

The following claim states that the terms in the summation are not very large and that they actually decrease in $k$.

**Claim 1.** If $k \geq 4$ is an even number, then

$$\frac{[(k-2)(k-3)...(k/2)]^2}{2^{k-1}k!} \leq \frac{1}{2k(k-1)}.$$

**Proof of the Claim.** We have

$$\begin{aligned}\frac{[(k-2)(k-3)...(k/2)]^2}{2^{k-1}k!} &= \frac{[(k-2)!/(k/2-1)!]^2}{k(k-1)(k-2)!2^{k-1}}\\ &= \frac{(k-2)!}{(k/2-1)!k(k-1)2^{k-1}}.\end{aligned}$$

Now, write $k = 2l$ for $l \in \mathbb{N}$ and $l \geq 2$ to get

$$\begin{aligned}\frac{[(k-2)(k-3)...(k/2)]^2}{2^{k-1}k!} &= \frac{(2l-2)!}{[(l-1)!]^2}\frac{1}{2l(2l-1)2^{2l-1}}\\ &= \binom{2l-2}{l-1}\frac{1}{2l(2l-1)2^{2l-1}}.\end{aligned}$$

Now, we use that

$$\binom{2l-2}{l-1} \leq \sum_{i=0}^{2l-2}\binom{2l-2}{i} = 2^{2l-2}$$

since it is the number of subsets of a set of $2l-2$ elements. This gives

$$\frac{[(k-2)(k-3)...(k/2)]^2}{2^{k-1}k!} \leq \frac{2^{2l-2}}{2l(2l-1)2^{2l-1}} = \frac{1}{2k(k-1)}$$

as we wanted. ■

Using the previous claim in (C.12) we get

$$d^\alpha \lesssim \frac{1}{d\nu}+\frac{1}{2d\nu}+\frac{1}{d\nu}\sum_{\substack{k\in\mathcal{K}\\k\geq 4}}\frac{1}{2k(k-1)} \asymp \frac{1}{d\nu}$$

since the sum is convergent.

We conclude that

$$\nu \lesssim d^{-\alpha-1}.$$

**Lower bound on $\nu$.** We have

$$d^{\alpha} \asymp \sum_{k\in\mathcal{K}} \frac{N_k \sigma_k^2}{\sigma_k^2+\nu}$$

such that ignoring all terms in the sum except the term with $k = k_\alpha$ being any even fixed integer with $k > \alpha$ Since all the terms in the sum are positive, we get

$$d^{\alpha} \gtrsim \frac{N_{k_\alpha}\sigma_{k_\alpha}^2}{\sigma_{k_\alpha}^2+\nu}.$$

Since $\alpha = O(1)$, we can choose $k_\alpha = O(1)$ and use from (C.9), (C.10), and (C.11) that

$$N_{k_\alpha} \asymp d^{k_\alpha}, \qquad \sigma_{k_\alpha}^2 \asymp d^{-k_\alpha-1}.$$

Thus,

$$d^{\alpha} \gtrsim \frac{d^{k_\alpha} d^{-k_\alpha-1}}{d^{-k_\alpha-1}+\nu} = \frac{d^{-1}}{d^{-k_\alpha-1}+\nu}.$$

It follows that

$$d^{-k_\alpha-1}+\nu \gtrsim d^{-\alpha-1}.$$

Since $k_\alpha > \alpha$, we have

$$d^{-k_\alpha-1} = o(d^{-\alpha-1}),$$

and therefore, for $d$ large enough,

$$\nu \gtrsim d^{-\alpha-1}.$$

Together with the upper bound we obtain

$$\nu = \Theta(d^{-\alpha-1}).$$

as claimed. □

### C.3 Teacher error scaling

The teacher error, as well as the student, will end up depending on how the teacher capacity coefficient $\alpha$ relates with the target degree $k_0$.

**Remark 1.** Remember that $\beta_{0,l} = 0$ for all $l$ such that every sum in $k \in \mathcal{K}$ that involves $\beta$ will not take $k = 0$ into account.

**Theorem 8.** In our setting of a pure harmonic target of degree $k_0 \geq 1$ with $\|f^*\|_{\mathcal{D}} = 1$, if $M_{\mathrm{TE}} = d^{\alpha}$ for some $\alpha > 0$, then

$$\overline{\mathcal{L}}_{\mathrm{TE}} \asymp \begin{cases} 1, \ \alpha \in (0,\ k_0] \\ d^{-\alpha+k_0}, \ \alpha > k_0. \end{cases}$$

*Proof.* We have the same expression for the teacher deterministic equivalent error as the one for the linear case

$$\overline{\mathcal{L}}_{\mathrm{TE}} = \nu \langle \Psi^{\frac{1}{2}}\beta, (\Psi+\nu)^{-1}\Psi^{\frac{1}{2}}\beta\rangle.$$

Notice that for all $k \leq k_0$ where $k_0 = O(1)$ and $d \gg 1$ we get from (C.9), (C.10), and (C.11) that

$$N_k \asymp d^k$$
$$\sigma_k^2 \asymp d^{-k-1}$$

for $k = 1$ or $k$ even. This is also true for $k = 0$ from equation (C.3) and $N_0 = 1$.

The unit norm hypothesis for the target gives

$$\sum_{k\in\mathcal{K}} \sum_{l=1}^{N_k} \sigma_k^2 \beta_{k,l}^2 = 1.$$

We obtain

$$\begin{aligned}\overline{\mathcal{L}}_{\mathrm{TE}} &= \nu\langle \Psi^{\frac{1}{2}}\beta, (\Psi+\nu)^{-1}\Psi^{\frac{1}{2}}\beta\rangle \\ &= \nu \sum_{k\in\mathcal{K}} \frac{\sigma_k^2}{\sigma_k^2+\nu} \sum_{l=1}^{N_k} \beta_{k,l}^2 \\ &\asymp d^{-\alpha-1} \sum_{k\in\mathcal{K}} \frac{d^{-k-1}}{d^{-k-1}+d^{-\alpha-1}} \sum_{l=1}^{N_k} \beta_{k,l}^2\end{aligned}$$

**Case 1:** $0 < \alpha \leq k_0$**.**

$$\overline{\mathcal{L}}_{\mathrm{TE}} \asymp d^{-\alpha-1} \left\{ \sum_{\substack{k\in\mathcal{K} \\ 0\leq k\leq \alpha}} \frac{d^{-k-1}}{d^{-k-1}} \sum_{l=1}^{N_k} \beta_{k,l}^2 + \sum_{\substack{k\in\mathcal{K} \\ \alpha+1\leq k\leq k_0}} \frac{d^{-k-1}}{d^{-\alpha-1}} \sum_{l=1}^{N_k} \beta_{k,l}^2 \right\}$$

using $\beta_{k,l}^2 \neq 0$ only for $k = k_0$, i.e., the pure harmonic target hypothesis, we get

$$\begin{aligned}\overline{\mathcal{L}}_{\mathrm{TE}} &\asymp d^{-\alpha-1} \frac{d^{-k_0-1}}{d^{-\alpha-1}} \sum_{l=1}^{N_{k_0}} \beta_{k_0,l}^2 \\ &\asymp 1.\end{aligned}$$

**Case 2:** $\alpha > k_0$**.**

$$\begin{aligned}\overline{\mathcal{L}}_{\mathrm{TE}} &\asymp d^{-\alpha-1} \frac{d^{-k_0-1}}{d^{-k_0-1}} \sum_{l=1}^{N_{k_0}} \beta_{k_0,l}^2 \\ &\asymp d^{-\alpha+k_0}.\end{aligned}$$

In conclusion we obtained

$$\overline{\mathcal{L}}_{\mathrm{TE}} \asymp \begin{cases} 1, & \text{if } \alpha \in (0,\ k_0], \\ d^{-\alpha+k_0}, & \text{if } \alpha > k_0 \end{cases} \tag{C.13}$$

as we desired. □

Intuitively, we can interpret this result as follows. When the teacher capacity exponent $\alpha$ is not larger than the target degree $k_0$, the teacher is not able to learn the target accurately, and its error remains of order 1. However, once the teacher is sufficiently large, i.e., $\alpha > k_0$, its error starts decreasing with its capacity. In fact, the key quantity is the difference $\alpha - k_0$, which measures the gap between the teacher capacity and the target complexity.

### C.4 Student error scaling: quadratic improvement

We finally prove the WTS result for the ReLU setting.

*Proof of Theorem 2.* Under the unit norm hypothesis $\|f^*\|_{\mathcal{D}} = 1$, we start with (B.12) that gives the student deterministic equivalent error at $t = T$, with $\overline{P}_0 = \nu(\Psi + \nu I)^{-1}$ and $S_T = (I - e^{-\Psi T})^2$,

$$\overline{\mathcal{L}}_{\mathrm{ST}}(T) = \langle \Psi^{\frac{1}{2}}\beta, [e^{-2\Psi T} - S_T\overline{P}_0 + \overline{P}_0(I - e^{-2\Psi T}) + \overline{\mathcal{P}}(S_T)]\Psi^{\frac{1}{2}}\beta\rangle.$$

We use the stopping-time from the statement of Theorem 2, that is,

$$T = \frac{1}{\sigma_{k_0}^2}(\log M_{\mathrm{TE}} + \log \overline{\mathcal{L}}_{\mathrm{TE}}^{-1}),$$

such that, using (C.13) we get

$$T \asymp \begin{cases} d^{k_0+1}\log(d^{\alpha}), & \text{if } \alpha \in (0,\ k_0] \\ d^{k_0+1}\log(d^{2\alpha-k_0}), & \text{if } \alpha > k_0. \end{cases}$$

We treat each of the four terms of $\overline{\mathcal{L}}_{\mathrm{ST}}(T)$ above separately. Since $\beta_{0,l} = 0$ for all $l$, we do not write $k = 0$ in the index summations below when $\beta$ is in the sum.

**Term 1:** $\langle \Psi^{\frac{1}{2}}\beta, e^{-2\Psi T}\Psi^{\frac{1}{2}}\beta\rangle$

**Case 1:** $0 < \alpha \leq k_0$**.**

$$\begin{aligned} 1) \coloneqq \langle \Psi^{\frac{1}{2}}\beta, e^{-2\Psi T}\Psi^{\frac{1}{2}}\beta\rangle &= \sum_{\substack{k\in\mathcal{K}\\ 1\leq k\leq k_0}} \frac{\sigma_k^2}{e^{2\sigma_k^2 T}} \sum_{l=1}^{N_k} \beta_{k,l}^2 \\ &\asymp \frac{\sigma_{k_0}^2 \sum_{l=1}^{N_{k_0}} \beta_{k_0,l}^2}{\exp\left(2\sigma_{k_0}^2 \sigma_{k_0}^{-2} \log\left(d^\alpha\right)\right)} \\ &\asymp \frac{1}{d^{2\alpha}} \\ &\asymp d^{-2\alpha} \end{aligned} \tag{C.14}$$

**Case 2:** $\alpha > k_0$**.**

$$\begin{aligned} 1) = \langle \Psi^{\frac{1}{2}}\beta, e^{-2\Psi T}\Psi^{\frac{1}{2}}\beta\rangle &\asymp \frac{1}{d^{4\alpha-2k_0}} \\ &\asymp d^{-4\alpha+2k_0} \end{aligned} \tag{C.15}$$

**Term 2:** $\langle \Psi^{\frac{1}{2}}\beta, -(I - e^{-\Psi T})^2\overline{P}_0\Psi^{\frac{1}{2}}\beta\rangle$

We have

$$2) \coloneqq \langle \Psi^{\frac{1}{2}}\beta, -(I - e^{-\Psi T})^2\overline{P}_0\Psi^{\frac{1}{2}}\beta\rangle \asymp \sum_{\substack{k\in\mathcal{K}\\ 1\leq k\leq k_0}} \frac{\sigma_k^2(1 - e^{-\sigma_k^2 T})^2}{\sigma_k^2\nu^{-1} + 1} \sum_{l=1}^{N_k} \beta_{k,l}^2.$$

**Case 1:** $0 < \alpha \leq k_0$**.**

$$\begin{aligned} \langle \Psi^{\frac{1}{2}}\beta, -(I - e^{-\Psi T})^2\overline{P}_0\Psi^{\frac{1}{2}}\beta\rangle &\asymp -\frac{d^{-k_0-1}(1 - d^{-\alpha})^2}{d^{-k_0-1+\alpha+1} + 1} \sum_{l=1}^{N_{k_0}} \beta_{k_0,l}^2 \\ &\asymp -(1 - 2d^{-\alpha} + d^{-2\alpha}) \end{aligned}$$

**Case 2:** $\alpha > k_0$**.**

$$\begin{aligned} \langle \Psi^{\frac{1}{2}}\beta, -(I - e^{-\Psi T})^2\overline{P}_0\Psi^{\frac{1}{2}}\beta\rangle &\asymp -\frac{d^{-k_0-1}(1 - d^{-2\alpha+k_0})^2}{d^{\alpha-k_0}} \sum_{l=1}^{N_{k_0}} \beta_{k_0,l}^2 \\ &\asymp -d^{-\alpha+k_0}(1 - 2d^{-2\alpha+k_0} + d^{-4\alpha+2k_0}) \end{aligned}$$

**Term 3:** $\langle \Psi^{\frac{1}{2}}\beta, \overline{P}_0(I - e^{-2\Psi T})\Psi^{\frac{1}{2}}\beta\rangle$

This term is treated similarly to the previous one

$$3) \coloneqq \langle \Psi^{\frac{1}{2}}\beta, \overline{P}_0(I - e^{-2\Psi T})\Psi^{\frac{1}{2}}\beta\rangle \asymp \sum_{\substack{k\in\mathcal{K}\\ 1\leq k\leq k_0}} \frac{\sigma_k^2(1 - e^{-2\sigma_k^2 T})}{\sigma_k^2\nu^{-1} + 1} \sum_{l=1}^{N_k} \beta_{k,l}^2$$

**Case 1:** $0 < \alpha \leq k_0$**.**

$$\begin{aligned} \langle \Psi^{\frac{1}{2}}\beta, \overline{P}_0(I - e^{-2\Psi T})\Psi^{\frac{1}{2}}\beta\rangle &\asymp d^{-k_0-1}(1 - d^{-2\alpha}) \sum_{l=1}^{N_{k_0}} \beta_{k_0,l}^2 \\ &\asymp (1 - d^{-2\alpha}) \end{aligned}$$

**Case 2:** $\alpha > k_0$**.**

$$\langle \Psi^{\frac{1}{2}}\beta, \overline{P}_0(I - e^{-2\Psi T})\Psi^{\frac{1}{2}}\beta\rangle \asymp d^{-\alpha+k_0}(1 - d^{-4\alpha+2k_0}).$$

Putting together the contributions of terms 2 and 3 we get in first order the following.

**Case 1:** $0 < \alpha \leq k_0$**.**

$$\begin{aligned} 2) + 3) &\asymp (2d^{-\alpha} - 2d^{-2\alpha}) \\ &\lesssim d^{-\alpha}. \end{aligned} \tag{C.16}$$

**Case 2:** $\alpha > k_0$**.**

$$\begin{aligned} 2) + 3) &\asymp (2d^{-2\alpha+k_0} - 2d^{-4\alpha+2k_0})d^{-\alpha+k_0} \\ &\lesssim d^{-3\alpha+2k_0}. \end{aligned} \tag{C.17}$$

**Term 4:** $\langle \Psi^{\frac{1}{2}}\beta, \overline{\mathcal{P}}(S_T)\Psi^{\frac{1}{2}}\beta\rangle$

By Theorem 7 we have

$$\langle \Psi^{\frac{1}{2}}\beta, \overline{\mathcal{P}}(S_T)\Psi^{\frac{1}{2}}\beta\rangle = \left\langle \Psi^{\frac{1}{2}}\beta, \frac{(I - e^{-\Psi T})^2}{(I + \Psi\nu^{-1})^2} + \frac{\frac{\Psi}{(I+\Psi\nu^{-1})^2}\operatorname{tr}\left(\frac{\nu^{-2}\Psi(I-e^{-\Psi T})^2}{(I+\Psi\nu^{-1})^2}\right)}{d^\alpha - \operatorname{tr}\left(\frac{\Psi^2\nu^{-2}}{(I+\Psi\nu^{-1})^2}\right)}\Psi^{\frac{1}{2}}\beta \right\rangle$$

Again, we treat each term separately. The first one is

$$4)(1) \coloneqq \left\langle \Psi^{\frac{1}{2}}\beta, \frac{(I - e^{-\Psi T})^2}{(I + \Psi\nu^{-1})^2}\Psi^{\frac{1}{2}}\beta \right\rangle = \sum_{\substack{k\in\mathcal{K} \\ 1\leq k\leq k_0}} \frac{d^{-k-1}(1 - e^{-\sigma_k^2 T})^2}{(1 + d^{\alpha-k})^2}\sum_{l=1}^{N_k}\beta_{k,l}^2.$$

**Case 1:** $0 < \alpha \leq k_0$**.**

$$\begin{aligned} \left\langle \Psi^{\frac{1}{2}}\beta, \frac{(I - e^{-\Psi T})^2}{(I + \Psi\nu^{-1})^2}\Psi^{\frac{1}{2}}\beta \right\rangle &\asymp d^{-k_0-1}(1 - d^{-\alpha})^2\sum_{l=1}^{N_{k_0}}\beta_{k,l}^2 \\ &\asymp 1. \end{aligned} \tag{C.18}$$

**Case 2:** $\alpha > k_0$**.**

$$\begin{aligned} \left\langle \Psi^{\frac{1}{2}}\beta, \frac{(I - e^{-\Psi T})^2}{(I + \Psi\nu^{-1})^2}\Psi^{\frac{1}{2}}\beta \right\rangle &\asymp \frac{d^{-k_0-1}(1 - d^{-2\alpha+k_0})^2}{d^{2\alpha-2k_0}}\sum_{l=1}^{N_{k_0}}\beta_{k,l}^2 \\ &\asymp d^{-2\alpha+2k_0} \end{aligned} \tag{C.19}$$

For the second term,

$$4)(2) \coloneqq \left\langle \Psi^{\frac{1}{2}}\beta, \frac{\frac{\Psi}{(I+\Psi\nu^{-1})^2}\operatorname{tr}\left(\frac{\nu^{-2}\Psi(I-e^{-\Psi T})^2}{(I+\Psi\nu^{-1})^2}\right)}{d^\alpha - \operatorname{tr}\left(\frac{\Psi^2\nu^{-2}}{(I+\Psi\nu^{-1})^2}\right)}\beta \right\rangle = \sum_{\substack{k\in\mathcal{K} \\ 1\leq k\leq k_0}} \frac{\operatorname{tr}\left(\frac{\nu^{-2}\Psi(I-e^{-\Psi T})^2}{(I+\Psi\nu^{-1})^2}\right)}{d^\alpha - \operatorname{tr}\left(\frac{\Psi^2\nu^{-2}}{(I+\Psi\nu^{-1})^2}\right)}\frac{\sigma_k^4 N_k}{(1 + d^{-k-1+2k_0+1})^2},$$

we first observe that we can simplify the denominator. In fact, by (B.8), we can write

$$\begin{aligned} d^\alpha - \operatorname{tr}\left(\frac{\Psi^2\nu^{-2}}{(I + \Psi\nu^{-1})^2}\right) &= \operatorname{tr}\left(\frac{\Psi\nu^{-1}}{I + \Psi\nu^{-1}}\right) - \operatorname{tr}\left(\frac{\Psi^2\nu^{-2}}{(I + \Psi\nu^{-1})^2}\right) \\ &= \operatorname{tr}\left(\frac{\Psi\nu^{-1} + \Psi^2\nu^{-2} - \Psi^2\nu^{-2}}{(I + \Psi\nu^{-1})^2}\right) \\ &= \operatorname{tr}\left(\frac{\Psi\nu^{-1}}{(I + \Psi\nu^{-1})^2}\right) \end{aligned}$$

such that we will need a lower bound for this trace.

**Lower bound for the denominator trace.**

We want to lower bound

$$\operatorname{tr}\left(\frac{\Psi\nu^{-1}}{(I+\Psi\nu^{-1})^2}\right)=\sum_{k\in\mathcal{K}}\frac{\sigma_k^2 d^{\alpha+1}}{(1+\sigma_k^2 d^{\alpha+1})^2}N_k.$$

Since the terms are positive, a single degree gives a lower bound. Let $k_\alpha$ be a fixed even degree with $k_\alpha>\alpha$, as in the proof of Lemma 4. Then $\sigma_{k_\alpha}^2 d^{\alpha+1}\asymp d^{\alpha-k_\alpha}\to 0$, so

$$\operatorname{tr}\left(\frac{\Psi\nu^{-1}}{(I+\Psi\nu^{-1})^2}\right)\;\geq\;\frac{\sigma_{k_\alpha}^2 d^{\alpha+1}}{(1+\sigma_{k_\alpha}^2 d^{\alpha+1})^2}N_{k_\alpha}\;\asymp\;d^{-k_\alpha-1}d^{\alpha+1}d^{k_\alpha}=d^\alpha.$$

Conversely, $\frac{\Psi\nu^{-1}}{(I+\Psi\nu^{-1})^2}\leq\frac{\Psi\nu^{-1}}{I+\Psi\nu^{-1}}$ and (B.8) give the upper bound $d^\alpha$. We conclude that

$$\operatorname{tr}\left(\frac{\Psi\nu^{-1}}{(I+\Psi\nu^{-1})^2}\right)\asymp d^\alpha.$$

**Upper bound for the numerator trace.**

We want to calculate

$$\operatorname{tr}\left(\frac{\nu^{-2}\Psi(I-e^{-\Psi T})^2}{(I+\Psi\nu^{-1})^2}\right)=\sum_{k\in\mathcal{K}}\frac{d^{2\alpha+2}(1-e^{-\sigma_k^2 T})^2}{(1+\sigma_k^2 d^{\alpha+1})^2}N_k\sigma_k^2.$$

Again, we divide into two cases depending on the value of $\alpha$.

**Case 1:** $0<\alpha\leq k_0$**.** We use the stopping-time

$$T=\frac{1}{\sigma_{k_0}^2}\log\left(d^\alpha\right)$$

let us divide the sum into four terms and calculate each one separately.

We divide the sum into four parts

$$\begin{aligned}\operatorname{tr}\left(\frac{\nu^{-2}\Psi(I-e^{-\Psi T})^2}{(I+\Psi\nu^{-1})^2}\right)=&\underbrace{\sum_{\substack{k\in\mathcal{K}\\0\leq k\leq\alpha}}\frac{d^{2\alpha+2}(1-e^{-\sigma_k^2\sigma_{k_0}^{-2}\log(d^\alpha)})^2}{(1+\sigma_k^2 d^{\alpha+1})^2}N_k\sigma_k^2}_{(i)}\\&+\underbrace{\sum_{\substack{k\in\mathcal{K}\\\alpha\leq k\leq k_0}}\frac{d^{2\alpha+2}(1-e^{-\sigma_k^2\sigma_{k_0}^{-2}\log(d^\alpha)})^2}{(1+\sigma_k^2 d^{\alpha+1})^2}N_k\sigma_k^2}_{(ii)}\\&+\underbrace{\sum_{\substack{k\in\mathcal{K}\\k_0+2\leq k\leq d}}\frac{d^{2\alpha+2}(1-e^{-\sigma_k^2\sigma_{k_0}^{-2}\log(d^\alpha)})^2}{(1+\sigma_k^2 d^{\alpha+1})^2}N_k\sigma_k^2}_{(iii)}\\&+\underbrace{\sum_{\substack{k\in\mathcal{K}\\k>d}}\frac{d^{2\alpha+2}(1-e^{-\sigma_k^2\sigma_{k_0}^{-2}\log(d^\alpha)})^2}{(1+\sigma_k^2 d^{\alpha+1})^2}N_k\sigma_k^2}_{(iv)}.\end{aligned}$$

For the first sum, since $k \leq \alpha$ and $\alpha = O(1)$ then, in particular, $k \ll d$ and we can use $\sigma_k^2 \asymp d^{-k-1}$ and $N_k \asymp d^k$ such that

$$(i) \asymp \sum_{\substack{k \in \mathcal{K} \\ 0 \leq k \leq \alpha}} \frac{d^{2\alpha+2}(1 - e^{-d^{-k+k_0 \log(d^\alpha)}})^2}{(1 + d^{-k-1}d^{\alpha+1})^2} d^{-k-1} d^k$$
$$\asymp d^{2\alpha+1} \sum_{\substack{k \in \mathcal{K} \\ 0 \leq k \leq \alpha}} \frac{(1 - e^{-d^{-k+k_0 \log(d^\alpha)}})^2}{(d^{\alpha-k})^2}$$

where we used that $k < \alpha$ such that the denominator is dominated by the term $d^{\alpha-k}$. Moreover, since $k \leq \alpha \leq k_0$ then the exponential term is too small and we obtain

$$(i) \asymp \sum_{\substack{k \in \mathcal{K} \\ 0 \leq k \leq \alpha}} d^{2k+1} \lesssim d^{2\alpha+1}.$$

Now, for the second term, since $k \leq k_0$ and $k_0 = O(1)$ we can use again that $\sigma_k^2 \asymp d^{-k-1}$ and $d^k$. Also, the exponential is still very small. However, the dominant term in the denominator is now 1 and we get

$$(ii) \asymp d^{2\alpha+2} \sum_{\substack{k \in \mathcal{K} \\ \alpha < k \leq k_0}} d^{-k-1} d^k \asymp d^{2\alpha+1}.$$

For the third term the exponential is non-negligible and we use a Taylor expansion for it. Moreover, since $k \leq d$ and $d$ is large we have

$$\sigma_k^2 \asymp \frac{d^{d-1}}{(d+k)^{d+k}} \left(\frac{e}{2}\right)^k [(k-2)(k-3)...(k/2)]^2$$
$$\gtrsim \frac{d^{d-1}}{(2d)^{d+k}} = d^{-k-1} 2^{-d-k},$$

such that

$$(1 + \sigma_k^2 d^{\alpha+1})^2 \gtrsim (1 + d^{\alpha-k} 2^{-d-k}) \asymp 1. \tag{C.20}$$

since we are in the case where $\alpha \leq k_0 \leq k$.

Now, for $k \leq d$ we have

$$N_k \sigma_k^2 \asymp \frac{\sqrt{d}(2k+d)[(k-2)(k-3)...(k/2)]^2}{2^k k! (d+k)^{\frac{5}{2}}}$$
$$\lesssim \frac{d^{\frac{1}{2}+1}}{d^{\frac{5}{2}}} g(k) = \frac{1}{d} g(k)$$

where

$$g(k) \coloneqq \frac{[(k-2)(k-3)...(k/2)]^2}{2^k k!}$$

and one can check that this function is decreasing in $k$.

Since in the range $2k_0 + 2 \leq k \leq d$ we cannot use $\sigma_k \asymp d^{-k-1}$ we prove the following result which will suffice for our purpose.

**Claim 2.** We have that the sequence $(\sigma_k)_{k \text{ even}}$ is decreasing. Moreover, for evert $k \in \mathbb{N}$ even we have

$$\frac{\sigma_{k+2}^2}{\sigma_k^2} \lesssim \frac{1}{d^2}.$$

**Proof of the Claim.** Using (C.10) we have

$$\frac{\sigma_{k+2}^2}{\sigma_k^2} \asymp \frac{(d+k)^{d+k}}{(d+k+2)^{d+k+2}} \left(\frac{e}{2}\right)^2 \left(\frac{k}{k/2}\right)^2$$
$$\asymp \frac{1}{(d+k+2)^2} \left(\frac{d+k}{d+k+1}\right)^{d+k}$$

Now, using that $\log x \leq x - 1$ for $x = 1 - \frac{1}{d+k+1}$ we obtain

$$\left(\frac{d+k}{d+k+1}\right)^{d+k} \leq e^{-1}$$

such that

$$\frac{\sigma_{k+2}^2}{\sigma_k^2} \lesssim \frac{1}{d^2}$$

as we wanted. ■

With the previous claim and calculations, we get

$$(ii) \lesssim d^{2\alpha+2} \sum_{\substack{k\in\mathcal{K} \\ k_0+2\leq k\leq d}} (d^{-k+k_0}\log(d^\alpha))^2 \frac{1}{d} g(k).$$

Now, we need to subdivide into the cases where $k \ll d$ and $k = cd$ with $c \in (0, 1]$.

Subcase 1.1: $k \ll d$. The function $g(k)$ does not contribute in leading order and we also note that the leading order is $k = k_0 + 2$ such that

$$\begin{aligned}(ii) &\lesssim d^{2\alpha+2} d^{-4} (\log(d^\alpha))^2 d^{-1} \\ &\asymp d^{2k_0-3} (\log(d^\alpha))^2.\end{aligned}$$

Subcase 1.2: $k = cd,\ c \in (0, 1]$. We can use a Stirling approximation for $g(k)$ and get

$$\begin{aligned}g(k) = \frac{\left[\frac{(k-2)!}{(\frac{k}{2}-1)}\right]^2}{2^k k!} &\asymp \frac{\left[k^{\frac{1}{2}} \left(\frac{k}{e}\right)^{k-2} k^{-\frac{1}{2}} \left(\frac{k}{2e}\right)^{-\frac{k}{2}+1}\right]^2}{2^k k^{\frac{1}{2}} \left(\frac{k}{e}\right)^k} \\ &\asymp \frac{k^{2k-4} e^{-2k+4} k^{-k+2} 2^{k-2} e^{k-2}}{2^k k^{\frac{1}{2}} k^k e^{-k}} \\ &\asymp \frac{k^{k-2} e^{-k+2} 2^{k-2}}{k^{k+\frac{1}{2}} e^{-k} 2^k} \\ &\asymp k^{-\frac{5}{2}}\end{aligned}$$

such that

$$\begin{aligned}(iii) &\lesssim \sum_{\substack{k\in\mathcal{K} \\ k_0+2\leq k\leq d \\ k=cd}} d^{2\alpha+2} [d^{-k+k_0}\log(d^\alpha)]^2 d^{-1} k^{-\frac{5}{2}} \\ &\asymp d^{2\alpha+1} (\log(d^\alpha))^2 \sum_{\substack{k\in\mathcal{K} \\ k_0+2\leq k\leq d \\ k=cd}} d^{-2cd+2k_0} (cd)^{-\frac{5}{2}} \\ &\asymp d^{2\alpha+2k_0-\frac{3}{2}} (\log(d^\alpha))^2 \sum_{\substack{k\in\mathcal{K} \\ k_0+2\leq k\leq d \\ k=cd}} d^{-2cd}\end{aligned}$$

which is superexponentially small in d.

Finally, for the fourth term we can use the fact that $\sigma_k^2$ decreases in $k$ from Lemma 2 and (C.20) to conclude that the denominator is again lower bounded by something of order 1. Using Lemma (2) again in the exponential and Taylor expansion we obtain

$$(iv) \lesssim \sum_{\substack{k\in\mathcal{K} \\ k>d}} d^{2\alpha+2} [d^{-k+k_0}\log(d^\alpha)]^2 N_k \sigma_k^2.$$

Now, for $k > d$ we have $d$ and $k$ large such that

$$\begin{aligned} N_k \sigma_k^2 &\asymp \frac{\sqrt{d}(2k+d)}{(d+k)^{\frac{5}{2}}} g(k) \\ &\lesssim d^{\frac{1}{2}-\frac{5}{2}} k k^{-\frac{5}{2}} \\ &\lesssim d^{-2} k^{-\frac{3}{2}} \end{aligned}$$

where we used the fact that we already know $g(k) \asymp k^{-\frac{5}{2}}$ when $k$ is large. Then, we get

$$(iv) \lesssim \sum_{\substack{k \in \mathcal{K} \\ k > d}} d^{2\alpha+2} d^{-2k+2k_0} (\log(d^\alpha))^2 d^{-2} k^{-\frac{3}{2}}$$

we observe that the dominant term is for $k = d+1$ and that the series converges. We obtain

$$\begin{aligned} (iv) &\lesssim d^{2\alpha+2} \sum_{\substack{k \in \mathcal{K} \\ k > d}} d^{2k_0-2k} (\log d^\alpha)^2 d^{-2} k^{-\frac{3}{2}} \\ &\asymp d^{2\alpha+2k_0-2d} (\log(d^\alpha))^2 (d+1)^{-\frac{3}{2}} \end{aligned}$$

where we used that the dominant term is for $k = d+1$.

**Case 2:** $\alpha > k_0$**.** We use the stopping-time

$$T = \frac{1}{\sigma_{k_0}^2} \log(d^{2\alpha-k_0})$$

We divide the sum into four terms again

$$\begin{aligned} \operatorname{tr}\left( \frac{\nu^{-2}\Psi(I-e^{-\Psi T})^2}{(I+\Psi\nu^{-1})^2} \right) = & \underbrace{\sum_{\substack{k \in \mathcal{K} \\ 0 \le k \le k_0}} \frac{d^{2\alpha+2}(1-e^{-\sigma_k^2 \sigma_{k_0}^{-2} \log(d^{2\alpha-k_0})})^2}{(1+\sigma_k^2 d^{\alpha+1})^2} N_k \sigma_k^2}_{(i)} \\ & + \underbrace{\sum_{\substack{k \in \mathcal{K} \\ k_0+2 \le k \le \alpha}} \frac{d^{2\alpha+2}(1-e^{-\sigma_k^2 \sigma_{k_0}^{-2} \log(d^{2\alpha-k_0})})^2}{(1+\sigma_k^2 d^{\alpha+1})^2} N_k \sigma_k^2}_{(ii)} \\ & + \underbrace{\sum_{\substack{k \in \mathcal{K} \\ \alpha+2 \le k \le d}} \frac{d^{2\alpha+2}(1-e^{-\sigma_k^2 \sigma_{k_0}^{-2} \log(d^{2\alpha-k_0})})^2}{(1+\sigma_k^2 d^{\alpha+1})^2} N_k \sigma_k^2}_{(iii)} \\ & + \underbrace{\sum_{\substack{k \in \mathcal{K} \\ k > d}} \frac{d^{2\alpha+2}(1-e^{-\sigma_k^2 \sigma_{k_0}^{-2} \log(d^{2\alpha-k_0})})^2}{(1+\sigma_k^2 d^{\alpha+1})^2} N_k \sigma_k^2}_{(iv)} . \end{aligned}$$

For the first term, since $k \le k_0$ we can use $\sigma_k^2 \asymp d^{-k-1}$ and $N_k \asymp d^k$. Also, using that $k \le k_0 < \alpha$ we get

$$\begin{aligned} (i) &\asymp \sum_{\substack{k \in \mathcal{K} \\ 0 \le k \le k_0}} \frac{d^{2\alpha}(1-e^{-d^{-k+k_0} \log(d^{2\alpha-k_0})})^2}{(1+d^{-k-1}d^{\alpha+1})^2} d^{-k-1} d^k \\ &\asymp \sum_{\substack{k \in \mathcal{K} \\ 0 \le k \le k_0}} \frac{d^{2\alpha+2} d^{-1}}{(d^{\alpha-k})^2} \\ &\asymp d^{2k_0+1}. \end{aligned}$$

For the second term we need to use Taylor expansion in the exponential and we get

$$\begin{aligned}(ii) &\lesssim \sum_{\substack{k\in\mathcal{K}\\ k_0+1\le k\le\alpha}} \frac{d^{2\alpha+1}(d^{-k+k_0}\log(d^{2\alpha-k_0}))^2}{(d^{\alpha-k})^2}\\ &\asymp d^{2\alpha+1}d^{2k_0-2\alpha}(\log(d^{2\alpha-k_0}))^2\\ &\asymp d^{2k_0+1}(\log(d^{2\alpha-k_0}))^2.\end{aligned}$$

For the third term, we have the same reasoning as previous $(iii)$ but now the logarithm term changed and also the dominant term is $k=\alpha+1$.

Subcase 2.1: $k \ll d$. We get

$$\begin{aligned}(iii) &\lesssim d^{2\alpha+2}(d^{k_0-\alpha-1}\log(d^{2\alpha-k_0}))^2 d^{-1}\\ &\lesssim d^{2k_0-1}(\log(d^{2\alpha-k_0}))^2\end{aligned}$$

Subcase 2.2: $k=cd,\ c\in(0,1]$. We get

$$(iii) \lesssim d^{2\alpha+2k_0-\frac{3}{2}}(\log(d^{2\alpha-k_0}))^2 \sum_{\substack{k\in\mathcal{K}\\ k_0+2\le k\le d\\ k=cd}} d^{-2cd}.$$

The fourth term is also the same as before just changing the logarithm term, we get

$$(iv) \lesssim d^{2\alpha+2k_0-2d}(\log(d^{\alpha}))^2.$$

**Combining the trace bounds.**

**Case 1:** $0<\alpha\le k_0$**.**

$$\begin{aligned}4)(2) = \left\langle \Psi^{\frac{1}{2}}\beta, \frac{\frac{\Psi}{(I+\Psi\nu^{-1})^2}\operatorname{tr}\left(\frac{\nu^{-2}\Psi(I-e^{-\Psi T})^2}{(I+\Psi\nu^{-1})^2}\right)}{d^\alpha-\operatorname{tr}\left(\frac{\Psi^2\nu^{-2}}{(I+\Psi\nu^{-1})^2}\right)}\Psi^{\frac{1}{2}}\beta\right\rangle &= \sum_{\substack{k\in\mathcal{K}\\ 1\le k\le k_0}} \frac{d^{2\alpha+1}}{d^\alpha}\frac{\sigma_k^4}{(1+d^{-k+1+\alpha+1})^2}\sum_{l=1}^{N_k}\beta_{k,l}^2\\ &\asymp d^{\alpha+1}d^{-k_0-1}\\ &\asymp d^{\alpha-k_0}.\end{aligned} \tag{C.21}$$

**Case 2:** $\alpha>k_0$**.**

$$\begin{aligned}4)(2) = \left\langle \Psi^{\frac{1}{2}}\beta, \frac{\frac{\Psi}{(I+\Psi\nu^{-1})^2}\operatorname{tr}\left(\frac{\nu^{-2}\Psi(I-e^{-\Psi T})^2}{(I+\Psi\nu^{-1})^2}\right)}{d^\alpha-\operatorname{tr}\left(\frac{\Psi^2\nu^{-2}}{(I+\Psi\nu^{-1})^2}\right)}\Psi^{\frac{1}{2}}\beta\right\rangle &= \sum_{\substack{k\in\mathcal{K}\\ 1\le k\le k_0}} \frac{d^{2k_0+1}}{d^\alpha}\frac{\sigma_k^4}{(1+d^{-k+1+\alpha+1})^2}\sum_{l=1}^{N_k}\beta_{k,l}^2\\ &\asymp \frac{d^{2k_0+1}}{d^\alpha}\frac{d^{-k_0-1}}{d^{2\alpha-2k_0}}\\ &\asymp d^{-3\alpha+3k_0}.\end{aligned} \tag{C.22}$$

Finally, comparing the terms (C.14),(C.16), (C.18), (C.21) and (C.15),(C.17), (C.19), (C.22) we obtain

$$\overline{\mathcal{L}}_{\mathrm{ST}}(T) \asymp \begin{cases}1, & \text{if } \alpha\in(0,\ k_0],\\ d^{-2\alpha+2k_0}, & \text{if } \alpha>k_0.\end{cases}$$

Comparing with (C.13), under the unit norm hypothesis, we obtain

$$\overline{\mathcal{L}}_{\mathrm{ST}}(T) \asymp \overline{\mathcal{L}}_{\mathrm{TE}}^2$$

as we wanted. □

# D MORE GENERAL SETTINGS

We provide a more complete analysis of the weak-to-strong phenomenon. We consider here the case of a general stopping-time with a pure harmonic target function and the case of a general stopping-time with a general target function that can span all degrees up to $k_0$. In both cases, we analyze the interplay between the parameters $\alpha$, $\gamma$, and $k_0$, related to the teacher capacity, the stopping-time, and the target complexity, respectively. We are again assuming $\|f^*\|_{\mathcal{D}} = 1$ and $\beta_{0,l} = 0$ for all $l$. We use $\overline{P}_0 = \nu(\Psi + \nu I)^{-1}$ and $S_{T_\gamma} = (I - e^{-\Psi T_\gamma})^2$ throughout this appendix.

## D.1 GENERAL STOPPING-TIME AND PURE HARMONIC TARGET

With the purpose of understanding how the stopping-time alone influence on the weak-to-strong phenomenon we study what happens when we choose

$$T_\gamma = \frac{1}{\sigma_\gamma^2} \begin{cases} \log(d^\alpha), \ \alpha \in (0, \ k_0] \\ \log(d^{2\alpha - k_0}), \ \alpha > k_0 \end{cases}$$

for a general $\gamma \in \mathcal{K}$ with $\gamma \geq 1$ but still uses a pure harmonic target function of degree $k_0$. This is the same stopping-time of Section C with the substitution of $\sigma_{k_0}^2$ by $\sigma_\gamma^2$ and writing $(\log M_{\mathrm{TE}} + \log \overline{\mathcal{L}}_{\mathrm{TE}}^{-1})$ explicitly using Theorem 8.

In this case, we observe that the teacher error and the denominator trace inside the student error do not change since they do not depend on the stopping-time. Let us calculate each of the student error terms separately.

**Term 1:** $\langle \Psi^{\frac{1}{2}}\beta, e^{-2\Psi T_\gamma}\Psi^{\frac{1}{2}}\beta \rangle$

**Case 1:** $0 < \alpha \leq k_0$**.**

$$\begin{aligned} 1) := \langle \Psi^{\frac{1}{2}}\beta, e^{-2\Psi T_\gamma}\Psi^{\frac{1}{2}}\beta \rangle &= \sum_{\substack{k \in \mathcal{K} \\ 1 \leq k \leq k_0}} \frac{\sigma_k^2}{e^{2\sigma_k^2 T_\gamma}} \sum_{l=1}^{N_k} \beta_{k,l}^2 \\ &\asymp \frac{\sigma_{k_0}^2 \sum_{l=1}^{N_{k_0}} \beta_{k_0,l}^2}{\exp\left(2\sigma_{k_0}^2 d^{\gamma+1} \log(d^\alpha)\right)} \\ &\asymp \frac{1}{(d^\alpha)^{2d^{\gamma-k_0}}}. \end{aligned} \tag{D.1}$$

**Case 2:** $\alpha > k_0$**.**

$$1) = \langle \Psi^{\frac{1}{2}}\beta, e^{-2\Psi T_\gamma}\Psi^{\frac{1}{2}}\beta \rangle \asymp \frac{1}{(d^{2\alpha - k_0})^{2d^{\gamma-k_0}}}. \tag{D.2}$$

**Term 2:** $\langle \Psi^{\frac{1}{2}}\beta, -(I - e^{-\Psi T_\gamma})^2 \overline{P}_0 \Psi^{\frac{1}{2}}\beta \rangle$

We have

$$2) := \langle \Psi^{\frac{1}{2}}\beta, -(I - e^{-\Psi T_\gamma})^2 \overline{P}_0 \Psi^{\frac{1}{2}}\beta \rangle \asymp \sum_{\substack{k \in \mathcal{K} \\ 1 \leq k \leq k_0}} \frac{\sigma_k^2 (1 - e^{-\sigma_k^2 T_\gamma})^2}{\sigma_k^2 \nu^{-1} + 1} \sum_{l=1}^{N_k} \beta_{k,l}^2.$$

**Case 1:** $0 < \alpha \leq k_0$**.**

$$\begin{aligned} \langle \Psi^{\frac{1}{2}}\beta, -(I - e^{-\Psi T_\gamma})^2 \overline{P}_0 \Psi^{\frac{1}{2}}\beta \rangle &\asymp -\frac{d^{-k_0 - 1}(1 - (d^{-\alpha})^{d^{\gamma-k_0}})^2}{d^{-k_0-1+\alpha+1} + 1} \sum_{l=1}^{N_{k_0}} \beta_{k_0,l}^2 \\ &\asymp -(1 - 2(d^{-\alpha})^{d^{\gamma-k_0}} + (d^{-\alpha})^{2d^{\gamma-k_0}}). \end{aligned}$$

**Case 2:** $\alpha > k_0$**.**

$$\begin{aligned}\langle \Psi^{\frac{1}{2}}\beta, -(I-e^{-\Psi T_\gamma})^2 \overline{P}_0 \Psi^{\frac{1}{2}}\beta\rangle &\asymp -\frac{d^{-k_0-1}(1-(d^{-2\alpha+k_0})^{d^{\gamma-k_0}})^2}{d^{\alpha-k_0}} \sum_{l=1}^{N_{k_0}} \beta_{k_0,l}^2 \\ &\asymp -d^{-\alpha+k_0}(1-2(d^{-2\alpha+k_0})^{d^{\gamma-k_0}} + (d^{-2\alpha+k_0})^{2d^{\gamma-k_0}}).\end{aligned}$$

**Term 3:** $\langle \Psi^{\frac{1}{2}}\beta, \overline{P}_0(I-e^{-2\Psi T_\gamma})\Psi^{\frac{1}{2}}\beta\rangle$

This term is treated similarly to the previous one

$$3) \coloneqq \langle \Psi^{\frac{1}{2}}\beta, \overline{P}_0(I-e^{-2\Psi T_\gamma})\Psi^{\frac{1}{2}}\beta\rangle \asymp \sum_{\substack{k\in\mathcal{K}\\ 1\le k\le k_0}} \frac{\sigma_k^2(1-e^{-2\sigma_k^2 T_\gamma})}{\sigma_k^2\nu^{-1}+1} \sum_{l=1}^{N_k} \beta_{k,l}^2.$$

**Case 1:** $0 < \alpha \le k_0$**.**

$$\begin{aligned}\langle \Psi^{\frac{1}{2}}\beta, \overline{P}_0(I-e^{-2\Psi T_\gamma})\Psi^{\frac{1}{2}}\beta\rangle &\asymp d^{-k_0-1}(1-(d^{-\alpha})^{2d^{\gamma-k_0}}) \sum_{l=1}^{N_{k_0}} \beta_{k_0,l}^2 \\ &\asymp (1-(d^{-\alpha})^{2d^{\gamma-k_0}})).\end{aligned}$$

**Case 2:** $\alpha > k_0$**.**

$$\langle \Psi^{\frac{1}{2}}\beta, \overline{P}_0(I-e^{-2\Psi T_\gamma})\Psi^{\frac{1}{2}}\beta\rangle \asymp d^{-\alpha+k_0}(1-(d^{-2\alpha+k_0})^{2d^{\gamma-k_0}}).$$

Putting together the contributions of the second and third term we get in first order the following.

**Case 1:** $0 < \alpha \le k_0$**.**

$$\begin{aligned}2)+3) &\asymp (2(d^{-\alpha})^{d^{\gamma-k_0}} - 2(d^{-\alpha})^{2d^{\gamma-k_0}}) \\ &\lesssim (d^{-\alpha})^{d^{\gamma-k_0}}.\end{aligned} \tag{D.3}$$

**Case 2:** $\alpha > k_0$**.**

$$\begin{aligned}2)+3) &\asymp (2(d^{-2\alpha+k_0})^{d^{\gamma-k_0}} - 2(d^{-2\alpha+k_0})^{2d^{\gamma-k_0}})d^{-\alpha+k_0} \\ &\lesssim (d^{-2\alpha+k_0})^{d^{\gamma-k_0}} d^{-\gamma+k_0}.\end{aligned} \tag{D.4}$$

**Term 4:** $\langle \Psi^{\frac{1}{2}}\beta, \overline{\mathcal{P}}(S_{T_\gamma})\Psi^{\frac{1}{2}}\beta\rangle$

By Theorem 7 we have

$$\langle \Psi^{\frac{1}{2}}\beta, \overline{\mathcal{P}}(S_{T_\gamma})\Psi^{\frac{1}{2}}\beta\rangle = \left\langle \Psi^{\frac{1}{2}}\beta, \frac{(I-e^{-\Psi T_\gamma})^2}{(I+\Psi\nu^{-1})^2} + \frac{\frac{\Psi}{(I+\Psi\nu^{-1})^2}\operatorname{tr}\left(\frac{\nu^{-2}\Psi(I-e^{-\Psi T_\gamma})^2}{(I+\Psi\nu^{-1})^2}\right)}{d^\alpha - \operatorname{tr}\left(\frac{\Psi^2\nu^{-2}}{(I+\Psi\nu^{-1})^2}\right)}\Psi^{\frac{1}{2}}\beta \right\rangle.$$

Again, we treat each term separately. The first one is

$$4)(1) \coloneqq \left\langle \Psi^{\frac{1}{2}}\beta, \frac{(I-e^{-\Psi T_\gamma})^2}{(I+\Psi\nu^{-1})^2}\Psi^{\frac{1}{2}}\beta \right\rangle = \sum_{\substack{k\in\mathcal{K}\\ 1\le k\le k_0}} \frac{(1-e^{-\sigma_k^2 T_\gamma})^2}{(1+d^{\alpha-k})^2} d^{-k-1} \sum_{l=1}^{N_k} \beta_{k,l}^2.$$

**Case 1:** $0 < \alpha \le k_0$**.**

$$\begin{aligned}\left\langle \Psi^{\frac{1}{2}}\beta, \frac{(I-e^{-\Psi T_\gamma})^2}{(I+\Psi\nu^{-1})^2}\Psi^{\frac{1}{2}}\beta \right\rangle &\asymp (1-(d^{-\alpha})^{d^{\gamma-k_0}})^2 d^{-k_0-1} \sum_{l=1}^{N_{k_0}} \beta_{k,l}^2 \\ &\asymp 1.\end{aligned} \tag{D.5}$$

**Case 2:** $\alpha > k_0$**.**

$$\left\langle \Psi^{\frac{1}{2}}\beta, \frac{(I-e^{-\Psi T_\gamma})^2}{(I+\Psi\nu^{-1})^2}\Psi^{\frac{1}{2}}\beta \right\rangle \asymp \frac{(1-(d^{-2\alpha+k_0})^{d^{\gamma-k_0}})^2}{d^{2\alpha-2k_0}} d^{-k_0-1}\sum_{l=1}^{N_{k_0}} \beta_{k,l}^2 \asymp d^{-2\alpha+2k_0} \tag{D.6}$$

The second term is

$$4)(2) := \left\langle \Psi^{\frac{1}{2}}\beta, \frac{\frac{\Psi}{(I+\Psi\nu^{-1})^2}\operatorname{tr}\left(\frac{\nu^{-2}\Psi(I-e^{-\Psi T_\gamma})^2}{(I+\Psi\nu^{-1})^2}\right)}{d^\alpha - \operatorname{tr}\left(\frac{\Psi^2\nu^{-2}}{(I+\Psi\nu^{-1})^2}\right)}\Psi^{\frac{1}{2}}\beta \right\rangle = \left\langle \Psi^{\frac{1}{2}}\beta, \frac{\frac{\Psi}{(I+\Psi\nu^{-1})^2}\operatorname{tr}\left(\frac{\nu^{-2}\Psi(I-e^{-\Psi T_\gamma})^2}{(I+\Psi\nu^{-1})^2}\right)}{\operatorname{tr}\left(\frac{\Psi\nu^{-1}}{(I+\Psi\nu^{-1})^2}\right)}\Psi^{\frac{1}{2}}\beta \right\rangle$$

where we used that

$$\begin{aligned} d^\alpha - \operatorname{tr}\left(\frac{\Psi^2\nu^{-2}}{(I+\Psi\nu^{-1})^2}\right) &= \operatorname{tr}\left(\frac{\Psi\nu^{-1}}{I+\Psi\nu^{-1}}\right) - \operatorname{tr}\left(\frac{\Psi^2\nu^{-2}}{(I+\Psi\nu^{-1})^2}\right) \\ &= \operatorname{tr}\left(\frac{\Psi\nu^{-1}+\Psi^2\nu^{-2}-\Psi^2\nu^{-2}}{(I+\Psi\nu^{-1})^2}\right) \\ &= \operatorname{tr}\left(\frac{\Psi\nu^{-1}}{(I+\Psi\nu^{-1})^2}\right) \end{aligned}$$

and this denominator trace will not change with the new stopping-time.

**Upper bound for the numerator trace.**

We have

$$\operatorname{tr}\left(\frac{\nu^{-2}\Psi(I-e^{-\Psi T_\gamma})^2}{(I+\Psi\nu^{-1})^2}\right) \asymp \sum_{k\in\mathcal{K}} \frac{d^{2\alpha+2}(1-e^{-\sigma_k^2 T_\gamma})^2}{(1+\sigma_k^2 d^{\alpha+1})^2} N_k\sigma_k^2.$$

Now, $k_0$ only appears inside the logarithm such that the division $\alpha \in (0, k_0]$ for comparison with the teacher error will make a difference only inside the logarithm. However, differently from previous calculation, we now divide the sum with respect to $\alpha, \gamma$ and $d$ instead of $\alpha, k_0$ and $d$ as before. Moreover, we need to further subdivide into the case $\alpha \le \gamma$ and $\alpha > \gamma$.

**Case 1:** $0 < \alpha \le k_0$**.**

Subcase 1.1: $\gamma \ge \alpha$.

We subdivide the sum into $4$ terms.

$$\begin{aligned} (i) &\asymp \sum_{\substack{k\in\mathcal{K}\\ 0\le k\le\alpha}} \frac{d^{2\alpha+2}(1-e^{-d^{-k+\gamma}\log(d^\alpha)})^2}{(1+d^{-k-1}d^{\alpha+1})^2} d^{-k-1}d^k \\ &\sim d^{2\alpha+1}\sum_{\substack{k\in\mathcal{K}\\ 0\le k\le\alpha}} \frac{(1-e^{-d^{\gamma-k}\log(d^\alpha)})^2}{(d^{\alpha-k})^2} \end{aligned}$$

since $k \le \alpha \le \gamma$ the exponential is negligible and we obtain

$$(i) \lesssim d\sum_{\substack{k\in\mathcal{K}\\ 0\le k\le\alpha}} d^{2k} \lesssim d^{2\alpha+1}.$$

The second sum is

$$(ii) \asymp \sum_{\substack{k\in\mathcal{K}\\ \alpha<k\le\gamma}} \frac{d^{2\alpha+2}(1-e^{-d^{-k+\gamma}\log(d^\alpha)})^2}{(1+d^{-k-1}d^{\alpha+1})^2} d^{-k-1}d^k$$

such that the exponential is negligible again but the dominant term in the denominator is now 1 and we get

$$\begin{aligned}(ii) &\asymp d^{2\alpha+2} \sum_{\substack{k\in\mathcal{K}\\ \alpha<k\leq\gamma}} d^{-k-1}d^k \\ &\asymp d^{2\alpha+1}\end{aligned}$$

For the third sum we have

$$(iii) \asymp \sum_{\substack{k\in\mathcal{K}\\ \gamma<k\leq d}} \frac{d^{2\alpha+2}(1-e^{-\sigma_k^2 T_\gamma})^2}{(1+\sigma_k^2 d^{\alpha+1})^2} N_k\sigma_k^2$$

Using the same reasoning as for the specific stopping-time case of Appendix C we have

$$(iii) \lesssim d^{2\alpha+2} \sum_{\substack{k\in\mathcal{K}\\ \gamma<k\leq d}} ((2d)^{\gamma-k}\log(d^\alpha))^2 \frac{g(k)}{d}$$

*Subcase 1.1.1:* $k \ll d$. $\lesssim d^{2\alpha-3}(\log(d^\alpha))^2$ by same reasoning as before.

*Subcase 1.1.2:* $k = cd,\ c \in (0,1]$. $\lesssim d^{2\alpha+2\gamma-\frac{3}{2}}(\log(d^\alpha))^2(2d)^{-2\epsilon d}$ by same reasoning as before.

For the last sum we have

$$\begin{aligned}(iv) &\asymp \sum_{\substack{k\in\mathcal{K}\\ k>d}} \frac{d^{2\alpha+2}(1-e^{-\sigma_k^2 T_\gamma})^2}{(1+\sigma_k^2 d^{\alpha+1})^2} N_k\sigma_k^2 \\ &\lesssim d^{2\alpha}2^{2\gamma-2d}d^{2\gamma-2d}(d+1)^{-\frac{3}{2}}(\log(d^\alpha))^2\end{aligned}$$

by same previous reasoning.

Subcase 1.2: $\gamma < \alpha$. We again subdivide the sum into 4 terms.

$$\begin{aligned}(i) &\asymp \sum_{\substack{k\in\mathcal{K}\\ 0\leq k\leq\gamma}} \frac{d^{2\alpha+2}(1-e^{-d^{-k+\gamma}\log(d^\alpha)})^2}{(1+d^{-k-1}d^{\alpha+1})^2} d^{-k-1}d^k \\ &\sim d^{2\alpha+1} \sum_{\substack{k\in\mathcal{K}\\ 0\leq k\leq\alpha}} \frac{(1-e^{-d^{\gamma-k}\log(d^\alpha)})^2}{(d^{\alpha-k})^2}\end{aligned}$$

since $k \leq \gamma$ the exponential is negligible and since $k \leq \gamma \leq \alpha$ the denominator is of order $(d^{\alpha-k})^2$ and we get

$$\begin{aligned}(i) &\asymp d \sum_{\substack{k\in\mathcal{K}\\ 0\leq k\leq\gamma}} d^{2k} \\ &\asymp d^{2\gamma+1}.\end{aligned}$$

The second sum is

$$(ii) \asymp \sum_{\substack{k\in\mathcal{K}\\ \gamma<k\leq\alpha}} \frac{d^{2\alpha+2}(1-e^{-d^{-k+\gamma}\log(d^\alpha)})^2}{(1+d^{-k-1}d^{\alpha+1})^2} d^{-k-1}d^k,$$

but now the dominant term in the denominator is $(d^{\alpha-k})^2$ and the exponential is non-negligible such that we use Taylor expansion to get

$$\begin{aligned}
&\lesssim \sum_{\substack{k\in\mathcal{K}\\ \gamma<k\leq\alpha}} \frac{d^{2\alpha+2}(d^{\gamma-k}\log(d^\alpha))^2}{(d^{\alpha-k})} d^{-1} \\
&\asymp d \sum_{\substack{k\in\mathcal{K}\\ \gamma<k\leq\alpha}} d^{2k}(d^{\gamma-k}\log(d^\alpha))^2 \\
&\asymp d^{2\gamma+1}(\log(d^\alpha))^2
\end{aligned}$$

For the third sum we have

$$(iii) \asymp \sum_{\substack{k\in\mathcal{K}\\ \alpha<k\leq d}} \frac{d^{2\alpha+2}(1-e^{-\sigma_k^2 T_\gamma})^2}{(1+\sigma_k^2 d^{\alpha+1})^2} N_k\sigma_k^2$$

Using the same reasoning as for the specific stopping-time we have

$$(iii) \lesssim d^{2\alpha+2} \sum_{\substack{k\in\mathcal{K}\\ \alpha<k\leq d}} ((2d)^{\gamma-k}\log(d^\alpha))^2 \frac{g(k)}{d}$$

*Subcase 1.2.1:* $k \ll d$. $\lesssim d^{2\gamma-3}(\log(d^\alpha))^2$ by same reasoning as before.

*Subcase 1.2.2:* $k = cd,\ c \in (0,1]$. $\lesssim d^{2\alpha+2\gamma-\frac{3}{2}}(\log(d^\alpha))^2(2d)^{-2\epsilon d}$ by same reasoning as before.

For the last sum we have

$$\begin{aligned}
(iv) &\asymp \sum_{\substack{k\in\mathcal{K}\\ k>d}} \frac{d^{2\alpha+2}(1-e^{-\sigma_k^2 T_\gamma})^2}{(1+\sigma_k^2 d^{\alpha+1})^2} N_k\sigma_k^2 \\
&\lesssim d^{2\alpha}2^{2\gamma-2d}d^{2\gamma-2d}(d+1)^{-\frac{3}{2}}(\log(d^\alpha))^2
\end{aligned}$$

by same previous reasoning.

**Case 2:** $\alpha > k_0$**.**

As already discussed, the only things that change are the terms inside the log, i.e., we should substitute $\log(d^\alpha)$ with $\log(d^{2\alpha-k_0})$.

**Combining the trace bounds.**

Now that we have the new expression for the numerator trace, we come back to the calculation of $4)(2)$.

**Case 1:** $0 < \alpha \leq k_0$**.**

Subcase 1.1: $\gamma \geq \alpha$.

In this case the numerator trace is of order $d^{2\alpha+1}$ and we get

$$\begin{aligned}
4)(2) &\asymp \sum_{\substack{k\in\mathcal{K}\\ 1\leq k\leq k_0}} \frac{d^{2\alpha+1}}{d^\alpha} \frac{\sigma_k^4}{(1+d^{\alpha-k})^2} \sum_{l=1}^{N_k} \beta_{k,l}^2 \\
&\asymp d^{\alpha-k_0}.
\end{aligned} \tag{D.7}$$

Subcase 1.2: $\gamma < \alpha$.

The numerator is now of order $d^{2\gamma+1}$ and we get

$$\begin{aligned}
4)(2) &\asymp \sum_{\substack{k\in\mathcal{K}\\ 1\leq k\leq k_0}} \frac{d^{2\gamma+1}}{d^\alpha} \frac{\sigma_k^4}{(1+d^{\alpha-k})^2} \sum_{l=1}^{N_k} \beta_{k,l}^2 \\
&\asymp d^{2\gamma-\alpha-k_0}.
\end{aligned} \tag{D.8}$$

**Case 2:** $\alpha > k_0$**.**

Subcase 2.1: $\gamma \geq \alpha$.

$$\begin{aligned} 4)(2) &\asymp \sum_{\substack{k\in\mathcal{K}\\ 1\leq k\leq k_0}} \frac{d^{2\alpha+1}}{d^\alpha}\frac{\sigma_k^4}{(1+d^{\alpha-k})^2}\sum_{l=1}^{N_k}\beta_{k,l}^2 \\ &\asymp d^{\alpha+1}\frac{d^{-k_0-1}}{(d^{\alpha-k_0})^2} \\ &\asymp d^{k_0-\alpha}. \end{aligned} \tag{D.9}$$

Subcase 2.2: $\gamma < \alpha$.

$$\begin{aligned} 4)(2) &\asymp d^{2\gamma+1-\alpha}\frac{d^{-k_0-1}}{(d^{\alpha-k_0})^2} \\ &\asymp d^{2\alpha-3\alpha+k_0}\begin{cases}\lesssim d^{3k_0-3\alpha}, & \gamma \leq k_0 < \alpha \\ \asymp d^{k_0-\alpha-\epsilon}, & k_0 \leq \gamma < \alpha.\end{cases} \end{aligned} \tag{D.10}$$

In conclusion, for the general stopping-time and pure harmonic target setting we have that the teacher error does not change, that is, it is again given by

$$\overline{\mathcal{L}}_{\mathrm{TE}} \asymp \begin{cases}1, & \text{if } \alpha \in (0,\ k_0], \\ d^{-\alpha+k_0}, & \text{if } \alpha > k_0,\end{cases}$$

but the student error changes and is given next.

Let us now make another subdivision. We divide the final student error again into the cases where $\alpha \in (0, k_0]$ and $\alpha > k_0$, since these are the natural subdivisions for comparison with the teacher error and are also the subdivisions used in the calculation of all terms of the student error. However, instead of subdividing between $\alpha \leq \gamma$ and $\alpha > \gamma$, as we did for the calculation of $4)(2)$, we now subdivide into the three cases $\gamma < k_0$, $\gamma = k_0$, and $\gamma > k_0$. This is natural since the relation between $\gamma$ and $k_0$ determines the intuitive cutoff induced by the stopping-time relative to the target degree.

**Final student error.**

**Case 1:** $0 < \alpha \leq k_0$**.**

Subcase 1.1: $\gamma < k_0$.

$$\begin{cases} 1)+2)+3) \lesssim 1, \text{ (D.1), (D.3)} \\ 4)(1) \asymp 1, \text{ (D.5)} \\ 4)(2) \asymp d^{\alpha-k_0} \lesssim 1, \ \alpha \leq \gamma, \text{ (D.7)} \\ \text{or } \asymp d^{2\gamma-\alpha-k_0}, \ \alpha > \gamma, \text{ (D.8)} \end{cases}$$

such that $4)(1)$ is the dominant term.

Subcase 1.2: $\gamma \geq k_0$.

In this case $\alpha \leq \gamma$ and we have

$$\begin{cases} 1)+2)+3) \lesssim 1, \text{ (D.1), (D.3)} \\ 4)(1) \asymp 1, \text{ (D.5)} \\ 4)(2) \asymp d^{\alpha-k_0} \lesssim 1, \text{ (D.7)} \end{cases}$$

and again $4)(1)$ dominates.

We conclude that for $\alpha \in (0, k_0]$, the term $4)(1)$ is always dominant such that $\overline{\mathcal{L}}_{\mathrm{ST}}(T_\gamma) \asymp 1$ for $\alpha \leq k_0$ without any further subdivision.

**Case 2:** $\alpha > k_0$**.**

Subcase 2.1: $\gamma < k_0$.

In this case $\alpha > \gamma$ and we have

$$\begin{cases} 1)(*), \text{ (D.2)} \\ 2)+3)(*), \text{ (D.4)} \\ 4)(1) \asymp d^{-2\alpha+2k_0}, \text{ (D.6)} \\ 4)(2) \lesssim d^{2\gamma-3\alpha+k_0}, \text{ (D.10)} \end{cases}$$

(*):

$$\begin{aligned} 1) \asymp \frac{1}{(d^{2\alpha-k_0})^{2d^{\alpha-k_0}}} &\asymp (d^{2\alpha-k_0})^{-2d^{\gamma-k_0}} \\ &\asymp e^{-2(2\alpha-k_0)d^{\gamma-k_0}\log d} \end{aligned}$$

and the exponential approximates 1 when $d$ is large such that $1) \asymp 1$ greater than $4)(1)$. Similar thing happens for $2)+3)$ such that the dominant term is $1)+2)+3)$ and we don't have weak-to-strong generalization in this case.

Subcase 2.2: $\gamma = k_0$.

Again, $\alpha > \gamma$ here and we get

$$\begin{cases} 1) \asymp d^{-4\alpha+2k_0}, \text{ (D.2)} \\ 2)+3) \lesssim d^{-3\alpha+2k_0}, \text{ (D.4)} \\ 4)(1) \asymp d^{-2\alpha+2k_0}, \text{ (D.6)} \\ 4)(2) \asymp d^{-3\alpha+3k_0}, \text{ (D.10)} \end{cases}$$

and $4)(1)$ dominates such that we have weak-to-strong generalization in this case.

Subcase 2.3: $\gamma > k_0$.

$$\begin{cases} 1)(**), \text{ (D.2)} \\ 2)+3)(**), \text{ (D.4)} \\ 4)(1) \asymp d^{2k_0-2\alpha}, \text{ (D.6)} \\ 4)(2) \asymp d^{k_0-\alpha}, \ \alpha \le \gamma, \text{ (D.9)} \\ \text{or} \asymp d^{2\gamma-3\alpha+k_0} \asymp d^{k_0-\alpha-2\epsilon}, \ \alpha > \gamma, \ \epsilon := \alpha-\gamma > 0, \text{ (D.10)} \end{cases}$$

(**):

$$\begin{aligned} 1) &\asymp e^{-2(2\alpha-k_0)d^{\gamma-k_0}\log d} \\ &\lesssim e^{-2(2\alpha-k_0)d^{\gamma-k_0}\log d} \end{aligned}$$

which decays exponentially in $d$ while $4)(1)$ decays polynomially. The same thing happens for $2)+3)$ such that the dominant term in this case is $4)(2)$ for $\alpha \le \gamma$ and we don't have weak-to-strong generalization. For $\alpha > \gamma$, depending on $\epsilon$, the dominant term will be $4)(1)$ (quadratic weak-to-strong) or $4)(2)$ (nonquadratic weak-to-strong) and we characterize below the precise regimes.

#### D.1.1 Summary of general stopping-time and pure harmonic target

The deterministic equivalent of the teacher population error satisfies

$$\overline{\mathcal{L}}_{\mathrm{TE}} \asymp \begin{cases} 1, & \alpha \le k_0, \\ d^{-\alpha+k_0}, & \alpha > k_0. \end{cases}$$

The behavior of the student depends on the position of the stopping-time exponent $\gamma$ relative to $k_0$ and $\alpha$. More precisely, the deterministic equivalent of the student population error satisfies

$$\overline{\mathcal{L}}_{\mathrm{ST}}(T_\gamma) \asymp \begin{cases} 1, & \alpha \le k_0, \\ 1, & \alpha > k_0,\ \gamma < k_0, \\ d^{-2\alpha+2k_0}, & \alpha > k_0,\ k_0 \le \gamma \le \dfrac{\alpha+k_0}{2}, \\ d^{2\gamma-3\alpha+k_0}, & \alpha > k_0,\ \dfrac{\alpha+k_0}{2} < \gamma < \alpha, \\ d^{-\alpha+k_0}, & \alpha > k_0,\ \gamma \ge \alpha. \end{cases} \tag{D.11}$$

Where the quadratic and non-quadratic separation comes from the fact that for the subcase $\gamma > k_0$ we have that the $1) + 2) + 3)$ terms are exponentially small in $d$ and when $\alpha > \gamma$, that is, $\alpha > k_0$, we have

$$\begin{cases} 4)(1) \asymp d^{2k_0-2\alpha} \\ 4)(2) \asymp d^{2\gamma-3\alpha+k_0} \asymp d^{k_0-\alpha-2\epsilon},\ \epsilon = \alpha - \gamma > 0 \end{cases}$$

such that the following happens.

**Quadratic WTS** happens when $k_0 - \alpha - 2\epsilon \le -2\alpha + 2k_0$ because in this case $4)(1)$ dominates. We have

$$\begin{aligned} k_0 - \alpha - 2\epsilon &\le -2\alpha + 2k_0 \\ -3\alpha + 2\gamma &\le -2\alpha + k_0 \\ 2\gamma &\le \alpha + k_0 \\ \gamma &\le \frac{\alpha + k_0}{2}. \end{aligned}$$

**Non-quadratic WTS** happens when $-2\alpha + 2k_0 < k_0 - \alpha - 2\epsilon < k_0 - \alpha$ because in this case $4)(2)$ dominates, while still being asymptotically smaller than the teacher error. The last inequality always holds since $\epsilon > 0$, and for the first one we have

$$\begin{aligned} -2\alpha + 2k_0 &< k_0 - \alpha - 2\epsilon \\ k_0 &< \alpha - 2\epsilon \\ k_0 &< \alpha - 2\alpha + 2\gamma \\ \frac{k_0 + \alpha}{2} &< \gamma. \end{aligned}$$

### D.2 General stopping-time and general target

Let us set the notation

$$\beta_k^2 \coloneqq \sum_{l=1}^{N_k} \beta_{k,l}^2.$$

Then, considering again the general stopping-time

$$T_\gamma = \frac{1}{\sigma_\gamma^2}(\log M_{\mathrm{TE}} + \log \overline{\mathcal{L}}_{\mathrm{TE}}^{-1})$$

and a general target with contributions from $k \in \mathcal{K}$ with $1 \le k \le k_0$ and no restriction on the coefficients beyond the unit norm hypothesis, we obtain the following.

**Teacher error.**

**Case 1:** $0 < \alpha \le k_0$**.**

$$\begin{aligned}\overline{\mathcal{L}}_{\mathrm{TE}} &= \nu\langle \Psi^{\frac{1}{2}}\beta, (\Psi+\nu)^{-1}\Psi^{\frac{1}{2}}\beta\rangle \\ &\asymp d^{-\alpha-1}\left\{\sum_{\substack{k\in\mathcal{K}\\1\le k\le\alpha}}\frac{d^{-k-1}}{d^{-k-1}}\beta_k^2 + \sum_{\substack{k\in\mathcal{K}\\\alpha<k\le k_0}}\frac{d^{-k-1}}{d^{-\alpha-1}}\beta_k^2\right\} \\ &\asymp \sum_{\substack{k\in\mathcal{K}\\1\le k\le k_0}}(d^{-\alpha-1}\mathbb{1}\{1\le k\le\alpha\} + d^{-k-1}\mathbb{1}\{\alpha<k\le k_0\})\beta_k^2.\end{aligned}$$

**Case 2:** $\alpha > k_0$**.**

$$\overline{\mathcal{L}}_{\mathrm{TE}} \asymp d^{-\alpha-1}\left\{\sum_{\substack{k\in\mathcal{K}\\1\le k\le k_0}}\frac{d^{-k-1}}{d^{-k-1}}\beta_k^2\right\} \asymp \sum_{\substack{k\in\mathcal{K}\\1\le k\le k_0}} d^{-\alpha-1}\beta_k^2.$$

**Student error.**

The numerator and denominator traces do not change from the case of general stopping-time and pure harmonic since they do not involve $\beta$.

**Term 1:** $\langle \Psi^{\frac{1}{2}}\beta, e^{-2\Psi T_\gamma}\Psi^{\frac{1}{2}}\beta\rangle$

**Case 1:** $0 < \alpha \le k_0$**.**

$$\begin{aligned}1) \coloneqq \langle \Psi^{\frac{1}{2}}\beta, e^{-2\Psi T_\gamma}\Psi^{\frac{1}{2}}\beta\rangle &= \sum_{\substack{k\in\mathcal{K}\\1\le k\le k_0}}\frac{\sigma_k^2}{e^{2\sigma_k^2 T_\gamma}}\beta_k^2 \\ &\asymp \sum_{\substack{k\in\mathcal{K}\\1\le k\le k_0}}\frac{d^{-k-1}}{\exp\left(2\sigma_k^2 d^{\gamma+1}\log\left[d^{\alpha}(\sum_{\substack{k'\in\mathcal{K}\\1\le k'\le k_0}}(d^{-\alpha-1}\mathbb{1}\{1\le k'\le\alpha\} + d^{-k'-1}\mathbb{1}\{\alpha<k'\le k_0\})\beta_{k'}^2)^{-1}\right]\right)}\beta_k^2 \\ &\asymp \sum_{\substack{k\in\mathcal{K}\\1\le k\le k_0}}\frac{d^{-k-1}}{[d^{\alpha}(\sum_{\substack{k'\in\mathcal{K}\\1\le k'\le k_0}}(d^{-\alpha-1}\mathbb{1}\{1\le k'\le\alpha\} + d^{-k'-1}\mathbb{1}\{\alpha<k'\le k_0\})\beta_{k'}^2)^{-1}]^{2d^{\gamma-k}}}\beta_k^2 \\ &\asymp \sum_{\substack{k\in\mathcal{K}\\1\le k\le k_0}} d^{-k-1-2\alpha d^{\gamma-k}}\left[\sum_{\substack{k'\in\mathcal{K}\\1\le k'\le k_0}}(d^{-\alpha-1}\mathbb{1}\{1\le k'\le\alpha\} + d^{-k'-1}\mathbb{1}\{\alpha<k'\le k_0\})\beta_{k'}^2\right]^{2d^{\gamma-k}}\beta_k^2\end{aligned} \tag{D.12}$$

**Case 2:** $\alpha > k_0$**.**

$$\begin{aligned}1) = \langle \Psi^{\frac{1}{2}}\beta, e^{-2\Psi T_\gamma}\Psi^{\frac{1}{2}}\beta\rangle &\asymp \sum_{\substack{k\in\mathcal{K}\\1\le k\le k_0}}\frac{d^{-k-1}}{[d^{2\alpha+1}(\sum_{\substack{k'\in\mathcal{K}\\1\le k'\le k_0}}\beta_{k'}^2)^{-1}]^{2d^{\gamma-k}}}\beta_k^2 \\ &\asymp \sum_{\substack{k\in\mathcal{K}\\1\le k\le k_0}} d^{-k-1-2(2\alpha+1)d^{\gamma-k}}\left[\sum_{\substack{k'\in\mathcal{K}\\1\le k'\le k_0}}\beta_{k'}^2\right]^{2d^{\gamma-k}}\beta_k^2\end{aligned} \tag{D.13}$$

**Term 2:** $\langle \Psi^{\frac{1}{2}}\beta, -(I - e^{-\Psi T_\gamma})^2\overline{P}_0\Psi^{\frac{1}{2}}\beta\rangle$

**Case 1:** $0 < \alpha \leq k_0$**.**

$$2) \coloneqq \langle \Psi^{\frac{1}{2}}\beta, -(I - e^{-\Psi T_\gamma})^2 \overline{P}_0 \Psi^{\frac{1}{2}}\beta \rangle = -\sum_{\substack{k\in\mathcal{K}\\ 1\leq k\leq k_0}} \frac{\sigma_k^2(1-e^{-\sigma_k^2 T_\gamma})^2}{\sigma_k^2\nu^{-1}+1}\beta_k^2$$

$$\asymp -\sum_{\substack{k\in\mathcal{K}\\ 1\leq k\leq k_0}} (1 - 2\left[d^{-\alpha}\sum_{\substack{k'\in\mathcal{K}\\ 1\leq k'\leq k_0}} (d^{-\alpha-1}\mathbb{1}\{1\leq k'\leq\alpha\} + d^{-k'-1}\mathbb{1}\{\alpha < k' \leq k_0\})\beta_{k'}^2\right]^{d^{\gamma-k}}$$

$$+\left[d^{-\alpha}\sum_{\substack{k'\in\mathcal{K}\\ 1\leq k'\leq k_0}} (d^{-\alpha-1}\mathbb{1}\{1\leq k'\leq\alpha\} + d^{-k'-1}\mathbb{1}\{\alpha < k' \leq k_0\})\beta_{k'}^2\right]^{2d^{\gamma-k}})$$

$$\cdot (d^{-\alpha-1}\mathbb{1}\{1\leq k\leq\alpha\} + d^{-k-1}\mathbb{1}\{\alpha < k \leq k_0\})\beta_k^2.$$

**Case 2:** $\alpha > k_0$**.**

$$2) \asymp -\sum_{\substack{k\in\mathcal{K}\\ 1\leq k\leq k_0}} \left(1 - 2\left[d^{-2\alpha-1}\sum_{\substack{k'\in\mathcal{K}\\ 1\leq k'\leq k_0}} \beta_{k'}^2\right]^{d^{\gamma-k}} + \left[d^{-2\alpha-1}\sum_{\substack{k'\in\mathcal{K}\\ 1\leq k'\leq k_0}} \beta_{k'}^2\right]^{2d^{\gamma-k}}\right) d^{-\alpha-1}\beta_k^2.$$

**Term 3:** $\langle \Psi^{\frac{1}{2}}\beta, \overline{P}_0(I - e^{-2\Psi T_\gamma})\Psi^{\frac{1}{2}}\beta\rangle$

**Case 1:** $0 < \alpha \leq k_0$**.**

$$3) \asymp \sum_{\substack{k\in\mathcal{K}\\ 1\leq k\leq k_0}} \left(1 - \left[d^{-\alpha}\sum_{\substack{k'\in\mathcal{K}\\ 1\leq k'\leq k_0}} (d^{-\alpha-1}\mathbb{1}\{1\leq k'\leq\alpha\} + d^{-k'-1}\mathbb{1}\{\alpha < k' \leq k_0\})\beta_{k'}^2\right]^{2d^{\gamma-k}}\right)$$

$$\cdot (d^{-\alpha-1}\mathbb{1}\{1\leq k\leq\alpha\} + d^{-k-1}\mathbb{1}\{\alpha < k \leq k_0\})\beta_k^2.$$

**Case 2:** $\alpha > k_0$**.**

$$3) \asymp \sum_{\substack{k\in\mathcal{K}\\ 1\leq k\leq k_0}} \left(1 - \left[d^{-2\alpha-1}\sum_{\substack{k'\in\mathcal{K}\\ 1\leq k'\leq k_0}} \beta_{k'}^2\right]^{2d^{\gamma-k}}\right) d^{-\alpha-1}\beta_k^2.$$

Now, summing terms 2 and 3 we obtain as follows.

**Case 1:** $0 < \alpha \leq k_0$**.**

$$2)+3) \asymp \sum_{\substack{k\in\mathcal{K}\\ 1\leq k\leq k_0}} (2\left[d^{-\alpha}\sum_{\substack{k'\in\mathcal{K}\\ 1\leq k'\leq k_0}} (d^{-\alpha-1}\mathbb{1}\{1\leq k'\leq\alpha\} + d^{-k'-1}\mathbb{1}\{\alpha < k' \leq k_0\})\beta_{k'}^2\right]^{d^{\gamma-k}}$$

$$-2\left[d^{-\alpha}\sum_{\substack{k'\in\mathcal{K}\\ 1\leq k'\leq k_0}} (d^{-\alpha-1}\mathbb{1}\{1\leq k'\leq\alpha\} + d^{-k'-1}\mathbb{1}\{\alpha < k' \leq k_0\})\beta_{k'}^2\right]^{2d^{\gamma-k}})$$

$$\cdot (d^{-\alpha-1}\mathbb{1}\{1\leq k\leq\alpha\} + d^{-k-1}\mathbb{1}\{\alpha < k \leq k_0\})\beta_k^2. \tag{D.14}$$

**Case 2:** $\alpha > k_0$**.**

$$2)+3) \asymp \sum_{\substack{k\in\mathcal{K}\\1\le k\le k_0}} \left( 2\left[ d^{-2\alpha-1} \sum_{\substack{k'\in\mathcal{K}\\1\le k'\le k_0}} \beta_{k'}^2 \right]^{d^{\gamma-k}} - 2\left[ d^{-2\alpha-1} \sum_{\substack{k'\in\mathcal{K}\\1\le k'\le k_0}} \beta_{k'}^2 \right]^{2d^{\gamma-k}} \right) d^{-\alpha-1}\beta_k^2. \tag{D.15}$$

**Term 4:** $\langle \Psi^{\frac{1}{2}}\beta, \overline{\mathcal{P}}(S_{T_\gamma})\Psi^{\frac{1}{2}}\beta\rangle$

For the first term,

$$4)(1) := \left\langle \Psi^{\frac{1}{2}}\beta, \frac{(I-e^{-\Psi T_\gamma})^2}{(I+\Psi\nu^{-1})^2}\Psi^{\frac{1}{2}}\beta \right\rangle = \sum_{\substack{k\in\mathcal{K}\\1\le k\le k_0}} \frac{(1-e^{-\sigma_k^2 T_\gamma})^2}{(1+d^{\alpha-k})^2} d^{-k-1}\beta_k^2.$$

**Case 1:** $0 < \alpha \le k_0$**.**

$$4)(1) \asymp \sum_{\substack{k\in\mathcal{K}\\1\le k\le k_0}} \left( 1 - \left[ d^{-\alpha} \sum_{\substack{k'\in\mathcal{K}\\1\le k'\le k_0}} (d^{-\alpha-1}\mathbb{1}\{1\le k'\le\alpha\} + d^{-k'-1}\mathbb{1}\{\alpha < k'\le k_0\})\beta_{k'}^2 \right]^{d^{\gamma-k}} \right)^2$$
$$\cdot (d^{-2\alpha+k-1}\mathbb{1}\{1\le k\le\alpha\} + d^{-k-1}\mathbb{1}\{\alpha<k\le k_0\})\beta_k^2. \tag{D.16}$$

**Case 2:** $\alpha > k_0$**.**

$$4)(1) \asymp \sum_{\substack{k\in\mathcal{K}\\1\le k\le k_0}} \left( 1 - \left[ d^{-2\alpha-1} \sum_{\substack{k'\in\mathcal{K}\\1\le k'\le k_0}} \beta_{k'}^2 \right]^{d^{\gamma-k}} \right)^2 d^{-2\alpha+k-1}\beta_k^2. \tag{D.17}$$

For the second term,

$$4)(2) := \left\langle \Psi^{\frac{1}{2}}\beta, \frac{\frac{\Psi}{(I+\Psi\nu^{-1})^2}\operatorname{tr}\left(\frac{\nu^{-2}\Psi(I-e^{-\Psi T_\gamma})^2}{(I+\Psi\nu^{-1})^2}\right)}{d^\alpha - \operatorname{tr}\left(\frac{\Psi^2\nu^{-2}}{(I+\Psi\nu^{-1})^2}\right)}\Psi^{\frac{1}{2}}\beta \right\rangle,$$

we have to subdivide into cases as follows.

**Case 1:** $0 < \alpha \le k_0$**.**

Subcase 1.1: $\gamma \ge \alpha$.

In this case the numerator trace is of order $d^{2\alpha+1}$ and we get

$$\begin{aligned} 4)(2) &\asymp \sum_{\substack{k\in\mathcal{K}\\1\le k\le k_0}} \frac{d^{2\alpha+1}}{d^\alpha}\frac{d^{-2k-2}}{(1+d^{\alpha-k})^2}\beta_k^2 \\ &\asymp \sum_{\substack{k\in\mathcal{K}\\1\le k\le k_0}} d^{\alpha-2k-1}(d^{-2\alpha+2k}\mathbb{1}\{1\le k\le\alpha\} + \mathbb{1}\{\alpha<k\le k_0\})\beta_k^2 \\ &\asymp \sum_{\substack{k\in\mathcal{K}\\1\le k\le k_0}} (d^{-\alpha-1}\mathbb{1}\{1\le k\le\alpha\} + d^{\alpha-2k-1}\mathbb{1}\{\alpha<k\le k_0\})\beta_k^2 \end{aligned}$$

Subcase 1.2: $\gamma < \alpha$.

The numerator is now of order $d^{2\gamma+1}$ and we get

$$\begin{aligned} 4)(2) &\asymp \sum_{\substack{k\in\mathcal{K}\\ 1\le k\le k_0}} \frac{d^{2\gamma+1}}{d^\alpha}\frac{d^{-2k-2}}{(1+d^{\alpha-k})^2}\beta_k^2 \\ &\asymp \sum_{\substack{k\in\mathcal{K}\\ 1\le k\le k_0}} d^{2\gamma-\alpha-2k-1}(d^{-2\alpha+2k}\mathbb{1}\{1\le k\le \alpha\}+\mathbb{1}\{\alpha<k\le k_0\})\beta_k^2 \\ &\asymp \sum_{\substack{k\in\mathcal{K}\\ 1\le k\le k_0}} (d^{2\gamma-3\alpha-1}\mathbb{1}\{1\le k\le \alpha\}+d^{2\gamma-\alpha-2k-1}\mathbb{1}\{\alpha<k\le k_0\})\beta_k^2 \end{aligned} \tag{D.18}$$

**Case 2:** $\alpha > k_0$**.**

Subcase 2.1: $\gamma \ge \alpha$.

$$\begin{aligned} 4)(2) &\asymp \sum_{\substack{k\in\mathcal{K}\\ 1\le k\le k_0}} \frac{d^{2\alpha+1}}{d^\alpha}\frac{d^{-2k-2}}{(1+d^{\alpha-k})^2}\beta_k^2 \\ &\asymp \sum_{\substack{k\in\mathcal{K}\\ 1\le k\le k_0}} d^{\alpha-2k-1}d^{-2\alpha+2k}\beta_k^2 \\ &\asymp d^{-\alpha-1}\sum_{\substack{k\in\mathcal{K}\\ 1\le k\le k_0}} \beta_k^2. \end{aligned} \tag{D.19}$$

Subcase 2.2: $\gamma < \alpha$.

$$\begin{aligned} 4)(2) &\asymp \sum_{\substack{k\in\mathcal{K}\\ 1\le k\le k_0}} \frac{d^{2\gamma+1}}{d^\alpha}\frac{d^{-2k-2}}{(d^{\alpha-k})^2}\beta_k^2 \\ &\asymp \sum_{\substack{k\in\mathcal{K}\\ 1\le k\le k_0}} d^{2\gamma-\alpha-2k-1}d^{-2\alpha+2k}\beta_k^2 \\ &\asymp d^{2\gamma-3\alpha-1}\sum_{\substack{k\in\mathcal{K}\\ 1\le k\le k_0}} \beta_k^2. \end{aligned} \tag{D.20}$$

### D.2.1 Summary of general stopping-time and general target

**Regime 1:** $0 < \alpha \le k_0,\ \gamma < \alpha$**.** In this regime, the teacher error satisfies

$$\overline{\mathcal{L}}_{\mathrm{TE}} \asymp \sum_{\substack{k\in\mathcal{K}\\ 1\le k\le k_0}} \left(d^{-\alpha-1}\mathbb{1}\{1\le k\le\alpha\}+d^{-k-1}\mathbb{1}\{\alpha<k\le k_0\}\right)\beta_k^2.$$

For the student, from (D.12) we have

$$1) \asymp \sum_{\substack{k\in\mathcal{K}\\ 1\le k\le k_0}} d^{-k-1}\left[d^{-\alpha}\sum_{\substack{k'\in\mathcal{K}\\ 1\le k'\le k_0}}\left(d^{-\alpha-1}\mathbb{1}\{1\le k'\le\alpha\}+d^{-k'-1}\mathbb{1}\{\alpha<k'\le k_0\}\right)\beta_{k'}^2\right]^{2d^{\gamma-k}}\beta_k^2.$$

For $k \le \gamma$, we have $d^{\gamma-k} \ge 1$ and

$$d^{-\alpha}\sum_{\substack{k'\in\mathcal{K}\\ 1\le k'\le k_0}}\left(d^{-\alpha-1}\mathbb{1}\{1\le k'\le\alpha\}+d^{-k'-1}\mathbb{1}\{\alpha<k'\le k_0\}\right)\beta_{k'}^2 \asymp d^{-\alpha}\overline{\mathcal{L}}_{\mathrm{TE}} = o(1),$$

such that the corresponding contributions in (D.12) are negligible. On the other hand, for $k > \gamma$, we have $d^{\gamma-k}$ going to zero, such that

$$\left[ d^{-\alpha} \sum_{\substack{k' \in \mathcal{K} \\ 1 \le k' \le k_0}} \left( d^{-\alpha-1} \mathbb{1}\{1 \le k' \le \alpha\} + d^{-k'-1} \mathbb{1}\{\alpha < k' \le k_0\} \right) \beta_{k'}^2 \right]^{2d^{\gamma-k}} \to 1.$$

Then, for the first term

$$1) \asymp \sum_{\substack{k \in \mathcal{K} \\ \gamma < k \le k_0}} d^{-k-1} \beta_k^2 + o(\overline{\mathcal{L}}_{\mathrm{TE}}).$$

For terms $2) + 3)$, from (D.14),

$$\begin{aligned} 2) + 3) \asymp \sum_{\substack{k \in \mathcal{K} \\ 1 \le k \le k_0}} ( & 2 \left[ d^{-\alpha} \sum_{\substack{k' \in \mathcal{K} \\ 1 \le k' \le k_0}} \left( d^{-\alpha-1} \mathbb{1}\{1 \le k' \le \alpha\} + d^{-k'-1} \mathbb{1}\{\alpha < k' \le k_0\} \right) \beta_{k'}^2 \right]^{d^{\gamma-k}} \\ & - 2 \left[ d^{-\alpha} \sum_{\substack{k' \in \mathcal{K} \\ 1 \le k' \le k_0}} \left( d^{-\alpha-1} \mathbb{1}\{1 \le k' \le \alpha\} + d^{-k'-1} \mathbb{1}\{\alpha < k' \le k_0\} \right) \beta_{k'}^2 \right]^{2d^{\gamma-k}} ) \\ & \left( d^{-\alpha-1} \mathbb{1}\{1 \le k \le \alpha\} + d^{-k-1} \mathbb{1}\{\alpha < k \le k_0\} \right) \beta_k^2. \end{aligned}$$

For the same previous reasoning, for $k \le \gamma$, both powers converge to 0 since $d^{\gamma-k} \ge 1$. For $k > \gamma$, both powers converge to 1 such that their difference also converges to 0. Thus, we obtain

$$2) + 3) = o(\overline{\mathcal{L}}_{\mathrm{TE}}).$$

For $4)(1)$, we have from (D.16) that

$$\begin{aligned} 4)(1) \asymp \sum_{\substack{k \in \mathcal{K} \\ 1 \le k \le k_0}} & \left( 1 - \left[ d^{-\alpha} \sum_{\substack{k' \in \mathcal{K} \\ 1 \le k' \le k_0}} \left( d^{-\alpha-1} \mathbb{1}\{1 \le k' \le \alpha\} + d^{-k'-1} \mathbb{1}\{\alpha < k' \le k_0\} \right) \beta_{k'}^2 \right]^{d^{\gamma-k}} \right)^2 \\ & \left( d^{-2\alpha+k-1} \mathbb{1}\{1 \le k \le \alpha\} + d^{-k-1} \mathbb{1}\{\alpha \le k \le k_0\} \right) \beta_k^2. \end{aligned}$$

For $k \le \gamma$, the power inside the parentheses is negligible and the term inside becomes 1. Since $k \le \gamma < \alpha$, then

$$d^{-2\alpha+k-1} \le d^{\gamma-\alpha} d^{-\alpha-1} = o(d^{-\alpha-1}),$$

and the terms in the sum with $k \le \gamma$ are of order $o(\overline{\mathcal{L}}_{\mathrm{TE}})$. For $k > \gamma$, the power inside the parentheses converges to 1, such that the whole term converges to 0. Then,

$$4)(1) = o(\overline{\mathcal{L}}_{\mathrm{TE}}).$$

Finally, since $\gamma < \alpha$, from (D.18) we have

$$4)(2) \asymp \sum_{\substack{k \in \mathcal{K} \\ 1 \le k \le k_0}} \left( d^{2\gamma-3\alpha-1} \mathbb{1}\{1 \le k \le \alpha\} + d^{2\gamma-\alpha-2k-1} \mathbb{1}\{\alpha < k \le k_0\} \right) \beta_k^2.$$

Here we do a different case analysis. For $k \le \alpha$, the ratio between the coefficient in $4)(2)$ and the corresponding coefficient in the teacher error is

$$\frac{d^{2\gamma-3\alpha-1}}{d^{-\alpha-1}} = d^{2\gamma-2\alpha} = o(1).$$

For $k > \alpha$, the corresponding ratio is

$$\frac{d^{2\gamma-\alpha-2k-1}}{d^{-k-1}} = d^{2\gamma-\alpha-k} \leq d^{2\gamma-2\alpha} = o(1).$$

Thus,

$$4)(2) = o(\overline{\mathcal{L}}_{\mathrm{TE}}).$$

We conclude that for Case 1, we obtain

$$\overline{\mathcal{L}}_{\mathrm{ST}}(T_\gamma) \asymp \sum_{\substack{k\in\mathcal{K}\\ \gamma<k\leq k_0}} d^{-k-1}\beta_k^2 + o(\overline{\mathcal{L}}_{\mathrm{TE}}).$$

Therefore, weak-to-strong generalization occurs if and only if

$$\sum_{\substack{k\in\mathcal{K}\\ \gamma<k\leq k_0}} d^{-k-1}\beta_k^2 = o\left(\overline{\mathcal{L}}_{\mathrm{TE}}\right).$$

and then it depends on the target coefficients sizes $\beta_k^2$ for $k > \gamma$.

**Regime 2:** $0 < \alpha \leq k_0$, $\gamma \geq \alpha$**.** Observe that in the previous case we only used that $\gamma < \alpha$ when analyzing terms $4)(1)$ and $4)(2)$ such that the other terms remain the same. Thus, we get again

$$\overline{\mathcal{L}}_{\mathrm{TE}} \asymp \sum_{\substack{k\in\mathcal{K}\\ 1\leq k\leq k_0}} \left(d^{-\alpha-1}\mathbb{1}\{1 \leq k \leq \alpha\} + d^{-k-1}\mathbb{1}\{\alpha < k \leq k_0\}\right)\beta_k^2,$$

and also

$$1) \asymp \sum_{\substack{k\in\mathcal{K}\\ \gamma<k\leq k_0}} d^{-k-1}\beta_k^2 + o(\overline{\mathcal{L}}_{\mathrm{TE}}),$$

and

$$2) + 3) = o(\overline{\mathcal{L}}_{\mathrm{TE}}).$$

For the first part of term $4)$, from (D.16),

$$4)(1) \asymp \sum_{\substack{k\in\mathcal{K}\\ 1\leq k\leq k_0}} \left(1 - \left[d^{-\alpha} \sum_{\substack{k'\in\mathcal{K}\\ 1\leq k'\leq k_0}} \left(d^{-\alpha-1}\mathbb{1}\{1 \leq k' \leq \alpha\} + d^{-k'-1}\mathbb{1}\{\alpha < k' \leq k_0\}\right)\beta_{k'}^2\right]^{d^{\gamma-k}}\right)^2$$
$$\left(d^{-2\alpha+k-1}\mathbb{1}\{1 \leq k \leq \alpha\} + d^{-k-1}\mathbb{1}\{\alpha < k \leq k_0\}\right)\beta_k^2.$$

For $k \leq \gamma$, the power inside the parentheses is negligible, whereas for $k > \gamma$ it converges to 1. Since $\gamma \geq \alpha$, this gives

$$4)(1) \asymp \sum_{\substack{k\in\mathcal{K}\\ 1\leq k\leq k_0}} d^{-2\alpha+k-1}\mathbb{1}\{1 \leq k \leq \alpha\}\beta_k^2 + \sum_{\substack{k\in\mathcal{K}\\ 1\leq k\leq k_0}} d^{-k-1}\mathbb{1}\{\alpha < k \leq \gamma\}\beta_k^2 + o(\overline{\mathcal{L}}_{\mathrm{TE}}),$$

that is we divided the last indicator as

$$\mathbb{1}\{\alpha < k \leq k_0\} = \mathbb{1}\{\alpha < k \leq \gamma\} + \mathbb{1}\{\gamma < k \leq k_0\}$$

ans used that the last contribution is $o(\overline{\mathcal{L}}_{\mathrm{TE}})$.

In particular, for $\alpha < k \leq \gamma$, term $4)(1)$ has exactly the same coefficient $d^{-k-1}$ as the corresponding term in the teacher error. And for $1 \leq k \leq \alpha$ they are smaller than the teacher since in this case

$$d^{-2\alpha+k-1} \leq d^{-\alpha-1}.$$

Finally, since $\gamma \geq \alpha$, from (D.18) we have

$$4)(2) \asymp \sum_{\substack{k\in\mathcal{K}\\ 1\leq k\leq k_0}} \left(d^{-\alpha-1}\mathbb{1}\{1 \leq k \leq \alpha\} + d^{\alpha-2k-1}\mathbb{1}\{\alpha < k \leq k_0\}\right)\beta_k^2.$$

For $k \leq \alpha$, term $4)(2)$ has exactly the same coefficient as the corresponding term in the teacher error,

$$\sum_{\substack{k\in\mathcal{K}\\1\leq k\leq k_0}} d^{-\alpha-1}\mathbb{1}\{1\leq k\leq \alpha\}\beta_k^2.$$

For $k > \alpha$, the second part of (D.18) is no larger than the corresponding teacher contribution since

$$\frac{d^{\alpha-2k-1}}{d^{-k-1}} = d^{\alpha-k} = o(1).$$

Therefore, the teacher-scale contribution for $k \leq \alpha$ is reproduced by $4)(2)$. For $\alpha < k \leq \gamma$, it is reproduced by $4)(1)$ as said above, while for $k > \gamma$ (if applicable) it is reproduced by $1)$. More precisely,

$$\overline{\mathcal{L}}_{\mathrm{ST}}(T_\gamma) \asymp \sum_{\substack{k\in\mathcal{K}\\1\leq k\leq k_0}} d^{-\alpha-1}\mathbb{1}\{1\leq k\leq \alpha\}\beta_k^2 + \sum_{\substack{k\in\mathcal{K}\\1\leq k\leq k_0}} d^{-k-1}\mathbb{1}\{\alpha< k\leq \gamma\}\beta_k^2 + \sum_{\substack{k\in\mathcal{K}\\1\leq k\leq k_0}} d^{-k-1}\mathbb{1}\{\gamma< k\leq k_0\}\beta_k^2.$$

Since $\gamma \geq \alpha$, the last two sums together give the entire contribution with $k > \alpha$, and hence

$$\overline{\mathcal{L}}_{\mathrm{ST}}(T_\gamma) \asymp \sum_{\substack{k\in\mathcal{K}\\1\leq k\leq k_0}} \left(d^{-\alpha-1}\mathbb{1}\{1\leq k\leq \alpha\} + d^{-k-1}\mathbb{1}\{\alpha< k\leq k_0\}\right)\beta_k^2 \asymp \overline{\mathcal{L}}_{\mathrm{TE}}.$$

Hence, there is no WTS when $\alpha \leq k_0$ and $\gamma \geq \alpha$.

Now, for the cases where $\alpha > k_0$ we will need to do a more complete subcase analysis. Instead of only comparing $\alpha$ and $\gamma$ as before we also need to compare how $\gamma$ relates with $k_0$.

**Regime 3:** $\alpha > k_0,\ \gamma < k_0$**.** In this regime, we have

$$\overline{\mathcal{L}}_{\mathrm{TE}} \asymp d^{-\alpha-1} \sum_{\substack{k\in\mathcal{K}\\1\leq k\leq k_0}} \beta_k^2.$$

Since $1 \leq \gamma < k_0$, we now separate the target degrees into $k \leq \gamma$ and $k > \gamma$. From (D.13),

$$1) \asymp \sum_{\substack{k\in\mathcal{K}\\1\leq k\leq k_0}} d^{-k-1}\left[d^{-2\alpha-1}\sum_{\substack{k'\in\mathcal{K}\\1\leq k'\leq k_0}}\beta_{k'}^2\right]^{2d^{\gamma-k}}\beta_k^2.$$

For $k \leq \gamma$, we have $d^{\gamma-k} \geq 1$ and, since in this case $d^{-2\alpha-1}\sum_{\substack{k'\in\mathcal{K}\\1\leq k'\leq k_0}}\beta_{k'}^2 \asymp d^{-\alpha}\overline{\mathcal{L}}_{\mathrm{TE}} = o(\overline{\mathcal{L}}_{\mathrm{TE}})$, the corresponding contributions of $1)$ are negligible. On the other hand, for $k > \gamma$, we have $d^{\gamma-k}$ going to zero. Then, (D.13) gives

$$1) \asymp \sum_{\substack{k\in\mathcal{K}\\\gamma<k\leq k_0}} d^{-k-1}\beta_k^2 + o(\overline{\mathcal{L}}_{\mathrm{TE}}).$$

For $2) + 3)$, (D.15) gives

$$2)+3) \asymp \sum_{\substack{k\in\mathcal{K}\\1\leq k\leq k_0}} \left(2\left[d^{-2\alpha-1}\sum_{\substack{k'\in\mathcal{K}\\1\leq k'\leq k_0}}\beta_{k'}^2\right]^{d^{\gamma-k}} - 2\left[d^{-2\alpha-1}\sum_{\substack{k'\in\mathcal{K}\\1\leq k'\leq k_0}}\beta_{k'}^2\right]^{2d^{\gamma-k}}\right)d^{-\alpha-1}\beta_k^2.$$

Again, for $k \leq \gamma$, both powers are negligible and for $k > \gamma$, both powers converge to 1 and their leading-order contributions cancel. Thus,

$$2)+3) = o(\overline{\mathcal{L}}_{\mathrm{TE}}).$$

For the first part of term $4)$, from (D.17),

$$4)(1) \asymp \sum_{\substack{k\in\mathcal{K}\\1\leq k\leq k_0}} \left(1 - \left[d^{-2\alpha-1}\sum_{\substack{k'\in\mathcal{K}\\1\leq k'\leq k_0}}\beta_{k'}^2\right]^{d^{\gamma-k}}\right)^2 d^{-2\alpha+k-1}\beta_k^2.$$

For $k \le \gamma$, the power inside the parentheses is negligible, and the corresponding contribution satisfies

$$\sum_{\substack{k\in\mathcal{K}\\1\le k\le\gamma}} d^{-2\alpha+k-1}\beta_k^2 \le d^{\gamma-\alpha}d^{-\alpha-1}\sum_{\substack{k\in\mathcal{K}\\1\le k\le\gamma}}\beta_k^2 = o(\overline{\mathcal{L}}_{\mathrm{TE}}),$$

since, in particular, we have $\alpha > \gamma$ in the case considered. For $k > \gamma$, the power converges to 1. Thus,

$$4)(1) = o(\overline{\mathcal{L}}_{\mathrm{TE}}).$$

Finally, since $\gamma < \alpha$, from (D.20) we have

$$4)(2) \asymp d^{2\gamma-3\alpha-1}\sum_{\substack{k\in\mathcal{K}\\1\le k\le k_0}}\beta_k^2 \asymp d^{2\gamma-2\alpha}\overline{\mathcal{L}}_{\mathrm{TE}} = o(\overline{\mathcal{L}}_{\mathrm{TE}}).$$

In conclusion, we obtain

$$\overline{\mathcal{L}}_{\mathrm{ST}}(T_\gamma) \asymp \sum_{\substack{k\in\mathcal{K}\\\gamma<k\le k_0}} d^{-k-1}\beta_k^2 + o(\overline{\mathcal{L}}_{\mathrm{TE}}).$$

Consequently, as in Case 1, in this regime WTS occurs if and only if the first term above is $o(\overline{\mathcal{L}}_{\mathrm{TE}})$, that is,

$$\sum_{\substack{k\in\mathcal{K}\\\gamma<k\le k_0}} d^{-k-1}\beta_k^2 = o\left(d^{-\alpha-1}\sum_{\substack{k\in\mathcal{K}\\1\le k\le k_0}}\beta_k^2\right).$$

**Regime 4:** $\alpha > k_0,\ k_0 \le \gamma \le \alpha$**.** In this regime, we have

$$\overline{\mathcal{L}}_{\mathrm{TE}} \asymp d^{-\alpha-1}\sum_{\substack{k\in\mathcal{K}\\1\le k\le k_0}}\beta_k^2.$$

Moreover, since $k \le k_0 \le \gamma$, we have $d^{\gamma-k} \ge 1$ for every $k \in \mathcal{K}$ with $1 \le k \le k_0$. Since

$$d^{-2\alpha-1}\sum_{\substack{k'\in\mathcal{K}\\1\le k'\le k_0}}\beta_{k'}^2 \asymp d^{-\alpha}\overline{\mathcal{L}}_{\mathrm{TE}},$$

we have the first term of the student (D.13) satisfying

$$1) \lesssim \left[d^{-2\alpha-1}\sum_{\substack{k'\in\mathcal{K}\\1\le k'\le k_0}}\beta_{k'}^2\right]^2 \sum_{\substack{k\in\mathcal{K}\\1\le k\le k_0}} d^{-k-1}\beta_k^2 \lesssim d^{-\alpha-2}\overline{\mathcal{L}}_{\mathrm{TE}}\sum_{\substack{k\in\mathcal{K}\\1\le k\le k_0}}\beta_k^2 \lesssim d^{-1}\overline{\mathcal{L}}_{\mathrm{TE}}^2 = o(\overline{\mathcal{L}}_{\mathrm{TE}}^2)$$

From (D.15), we get

$$2)+3) \lesssim \left[d^{-2\alpha-1}\sum_{\substack{k'\in\mathcal{K}\\1\le k'\le k_0}}\beta_{k'}^2\right] d^{-\alpha-1}\sum_{\substack{k\in\mathcal{K}\\1\le k\le k_0}}\beta_k^2 \asymp d^{-\alpha}\overline{\mathcal{L}}_{\mathrm{TE}}^2 = o(\overline{\mathcal{L}}_{\mathrm{TE}}^2).$$

For the first part of term $4)$, (D.17) gives

$$4)(1) \asymp \sum_{\substack{k\in\mathcal{K}\\1\le k\le k_0}}\left(1-\left[d^{-2\alpha-1}\sum_{\substack{k'\in\mathcal{K}\\1\le k'\le k_0}}\beta_{k'}^2\right]^{d^{\gamma-k}}\right)^2 d^{-2\alpha+k-1}\beta_k^2.$$

The power inside the parentheses is again negligible and we obtain

$$4)(1) \asymp \sum_{\substack{k\in\mathcal{K}\\1\le k\le k_0}} d^{-2\alpha+k-1}\beta_k^2 \lesssim d^{-2\alpha+k_0-1}\sum_{\substack{k\in\mathcal{K}\\1\le k\le k_0}}\beta_k^2 = d^{k_0-\alpha}\overline{\mathcal{L}}_{\mathrm{TE}} = o(\overline{\mathcal{L}}_{\mathrm{TE}})$$

Finally, from (D.20) we have

$$4)(2) \asymp d^{2\gamma-3\alpha-1} \sum_{\substack{k\in\mathcal{K}\\ 1\le k\le k_0}} \beta_k^2 \asymp d^{2\gamma-2\alpha}\overline{\mathcal{L}}_{\mathrm{TE}} = o(\overline{\mathcal{L}}_{\mathrm{TE}})$$

Combining these estimates, we obtain that the dominant term is $4)(1)$ such that we obtain

$$\overline{\mathcal{L}}_{\mathrm{ST}}(T_\gamma) \asymp d^{k_0-\alpha}\overline{\mathcal{L}}_{\mathrm{TE}} = o(\overline{\mathcal{L}}_{\mathrm{TE}}).$$

Then, we have WTS for this case. The quadratic and non-quadratic distinction as done in the pure harmonic case, would require a more careful analysis and will end up depending on the target coefficients.

**Regime 5:** $\alpha > k_0,\ \gamma \ge \alpha$**.** In this regime, we have again

$$\overline{\mathcal{L}}_{\mathrm{TE}} \asymp d^{-\alpha-1} \sum_{\substack{k\in\mathcal{K}\\ 1\le k\le k_0}} \beta_k^2.$$

Moreover, since $k \le k_0 < \alpha \le \gamma$, we have again that $\gamma - k > 0$ and $d^{\gamma-k} \ge 1$ for every $k \in \mathcal{K}$ with $1 \le k \le k_0$ such that we obtain the same conclusion as in the previous case for the first term,that is,

$$1) = o(\overline{\mathcal{L}}_{\mathrm{TE}}).$$

Similarly, we have the same previous conclusion for (D.15)

$$2) + 3) = o(\overline{\mathcal{L}}_{\mathrm{TE}}).$$

And the same thing for the first part of term $4)$,

$$4)(1) = o(\overline{\mathcal{L}}_{\mathrm{TE}}),$$

Finally, since $\alpha \le \gamma$ now we have to use (D.19) for the second part of $4)$ and this gives

$$4)(2) \asymp d^{-\alpha-1} \sum_{\substack{k\in\mathcal{K}\\ 1\le k\le k_0}} \beta_k^2 \asymp \overline{\mathcal{L}}_{\mathrm{TE}}.$$

Therefore, term $4)(2)$ is the dominant contribution to the student error, and we obtain

$$\overline{\mathcal{L}}_{\mathrm{ST}}(T_\gamma) \asymp \overline{\mathcal{L}}_{\mathrm{TE}}.$$

Then, there is no WTS in this last case.

All regimes are summarized in the table below.

| Regime | Comparison | Conclusion |
|---|---|---|
| $\alpha \le k_0,\ \gamma < \alpha$ | WTS $\iff \sum_{\substack{k\in\mathcal{K}\\ \gamma<k\le k_0}} d^{-k-1}\beta_k^2 = o(\overline{\mathcal{L}}_{\mathrm{TE}})$ | Coefficient-dependent |
| $\alpha \le k_0,\ \gamma \ge \alpha$ | $\overline{\mathcal{L}}_{\mathrm{ST}} \asymp \overline{\mathcal{L}}_{\mathrm{TE}}$ | No WTS |
| $\alpha > k_0,\ \gamma < k_0$ | WTS $\iff \sum_{\substack{k\in\mathcal{K}\\ \gamma<k\le k_0}} d^{-k-1}\beta_k^2 = o\left(d^{-\alpha-1}\sum_{\substack{k\in\mathcal{K}\\ 1\le k\le k_0}}\beta_k^2\right)$ | Coefficient-dependent |
| $\alpha > k_0,\ k_0 \le \gamma < \alpha$ | $\overline{\mathcal{L}}_{\mathrm{ST}} \asymp d^{k_0-\alpha}\overline{\mathcal{L}}_{\overline{\mathcal{L}}} = o(\mathcal{L}_{\mathrm{TE}})$ | WTS |
| $\alpha > k_0,\ \gamma \ge \alpha$ | $\overline{\mathcal{L}}_{\mathrm{ST}} \asymp \overline{\mathcal{L}}_{\mathrm{TE}}$ | No WTS |

Table 2: Weak-to-strong generalization regimes determined by the teacher width exponent $\alpha$, stopping-time parameter $\gamma$, and maximal target degree $k_0$. Here WTS means $\overline{\mathcal{L}}_{\mathrm{ST}} = o(\overline{\mathcal{L}}_{\mathrm{TE}})$ as $d \to \infty$.

**Remark 2** (The two coefficient-dependent conditions are one condition)**.** In the first and third rows of the table, weak-to-strong generalization occurs if and only if $\sum_{\substack{k\in\mathcal{K}\\ \gamma<k\le k_0}} d^{-k-1}\beta_k^2 = o(\overline{\mathcal{L}}_{\mathrm{TE}})$. In the third row this is how the condition is written, since $\overline{\mathcal{L}}_{\mathrm{TE}} \asymp d^{-\alpha-1}\sum_{\substack{k\in\mathcal{K}\\ 1\le k\le k_0}} \beta_k^2$ there. In the first row $\overline{\mathcal{L}}_{\mathrm{TE}} \asymp \sum_{\substack{k\in\mathcal{K}\\ 1\le k\le \alpha}} d^{-\alpha-1}\beta_k^2 + \sum_{\substack{k\in\mathcal{K}\\ \alpha<k\le k_0}} d^{-k-1}\beta_k^2$, and the condition compares the same held-out energy $\sum_{\substack{k\in\mathcal{K}\\ \gamma<k\le k_0}} d^{-k-1}\beta_k^2 = \sum_{\substack{k\in\mathcal{K}\\ \gamma<k\le k_0}} \|f_k^*\|_{\mathcal{D}}^2$ with this different expression. The two rows therefore state a single condition, (6.3) in the main text, with the teacher error taking its two forms (6.4).

## E TWO MIXED-TARGET EXAMPLES

This appendix proves the two statements about the two-mode target (6.5) made in Section 6. Both concern the deterministic equivalents (3.2) and (3.3), and both are uniform over all stopping times $t \ge 0$, not only over the harmonic stopping times $T_\gamma$ of Appendix D. They are stated up to constants, which is all that the conclusions of Section 6 require; the leading constants, which the figures track, are recorded in Remark 3. Transferring the conclusions to the random spherical model requires Assumption 1 at the relevant error scales, uniformly in time; the figures of this appendix are evaluations of the deterministic equivalents, not simulations.

**The errors of a mixed target are sums over its harmonic components.** Since $\Psi$ is the scalar $\sigma_k^2$ on the degree-$k$ harmonic block, every matrix in (3.2) and (3.3) is a function of $\Psi$, hence scalar on each block, and the two errors are sums over the harmonic components $f_k^\star$ of the target, weighted by their energies $\|f_k^\star\|_{\mathcal{D}}^2$, the squared norms of the blocks of $\Psi^{1/2}\beta$:

$$\overline{\mathcal{L}}_{\mathrm{TE}} = \sum_k \|f_k^\star\|_{\mathcal{D}}^2 \,\frac{\nu}{\sigma_k^2+\nu}, \qquad \overline{\mathcal{L}}_{\mathrm{ST}}(t) = \sum_k \|f_k^\star\|_{\mathcal{D}}^2 \,\overline{\ell}_k(t), \tag{E.1}$$

where $\overline{\ell}_k(t)$ is the deterministic equivalent of the student error for a unit target of degree $k$. Reading off (3.3) on the degree-$k$ block,

$$\overline{\ell}_k(t) = \Big(\frac{\nu+\sigma_k^2 e^{-\sigma_k^2 t}}{\sigma_k^2+\nu}\Big)^2 + \frac{\nu\sigma_k^2}{(\sigma_k^2+\nu)^2}\,\overline{H}(t), \qquad \overline{H}(t) = \frac{\mathrm{tr}[\Psi(\Psi+\nu I)^{-2}(I-e^{-\Psi t})^2]}{\mathrm{tr}[\Psi(\Psi+\nu I)^{-2}]} \in [0,1]. \tag{E.2}$$

The first term is the bias $\overline{B}(t)$ and the second the variance $\overline{V}(t)$ of the unit degree-$k$ target; $\overline{H}$ is the trace ratio of (3.3), common to all degrees, which increases from 0 at $t=0$ to 1 at $t=\infty$. This is the additivity used in Section 6: the held-out energy $\mathcal{E}_{>\gamma}$ of (6.2) and the two forms of the teacher error (6.4) are (E.1) together with Lemma 5 below. Two consequences of (E.2) are used repeatedly: every term is nonnegative, so a lower bound may drop any of them, and $\overline{\ell}_k(t) \le 1 + \frac{1}{4} = \frac{5}{4}$ for all $t$, since the bias is at most 1 and the variance at most $\frac{1}{4}\overline{H} \le \frac{1}{4}$.

Throughout, $m = d^\alpha$ with $\alpha > 1$ fixed, so that $\nu \asymp d^{-\alpha-1}$ by Lemma 4. Appendix C uses this estimate through the pure-target calculations; here we need it in the following packaged form.

**Lemma 5** (Consequences of the order of $\nu$)**.** Fix $\alpha > 1$ and $m = d^\alpha$, and let $k \in \mathcal{K}$ be a fixed degree. With constants depending only on $\alpha$ and $k$,

$$\text{(i)}\quad \frac{\nu}{\sigma_k^2+\nu} \asymp d^{k-\alpha} \quad (k<\alpha), \qquad \text{(ii)}\quad \frac{\nu}{\sigma_k^2+\nu} \longrightarrow 1 \quad (k>\alpha),$$

$$\text{(iii)}\quad \mathrm{tr}\,\Psi \asymp d^{-1}, \qquad \text{(iv)}\quad \nu\,\mathrm{tr}[\Psi(\Psi+\nu I)^{-2}] \asymp m.$$

*Proof. Items (i) and (ii).* By Lemma 4, $\nu \asymp d^{-\alpha-1}$, while $\sigma_k^2 \asymp d^{-k-1}$ for fixed $k$. Hence

$$\frac{\sigma_k^2}{\nu} \asymp d^{\alpha-k}, \qquad \text{and} \qquad \frac{\nu}{\sigma_k^2+\nu} = \frac{1}{1+\sigma_k^2/\nu}.$$

If $k<\alpha$ the ratio $\sigma_k^2/\nu$ tends to infinity, and the right-hand side is $\asymp \nu/\sigma_k^2 \asymp d^{k-\alpha}$, which is (i). If $k>\alpha$ the ratio tends to zero, and the right-hand side tends to one, which is (ii).

*Item (iii).* The degree-one term alone gives the lower bound, by (C.5):

$$\operatorname{tr}\Psi = \sum_{k\in\mathcal{K}} N_k\sigma_k^2 \;\geq\; N_1\sigma_1^2 = \frac{1}{4d}.$$

For the upper bound, the trace of the kernel operator is the expectation of its diagonal, and $(r)_+^2 \leq r^2$:

$$\operatorname{tr}\Psi = \mathbb{E}_x\, K(x,x) = \mathbb{E}\,(\langle u,x\rangle)_+^2 \;\leq\; \mathbb{E}\,\langle u,x\rangle^2 = \frac{1}{d}.$$

*Item (iv), upper bound.* Since $\frac{\nu\sigma_k^2}{(\sigma_k^2+\nu)^2} \leq \frac{\sigma_k^2}{\sigma_k^2+\nu}$ for every $k$, the fixed-point equation (3.1) gives

$$\nu\operatorname{tr}[\Psi(\Psi+\nu I)^{-2}] = \sum_{k\in\mathcal{K}} N_k \frac{\nu\sigma_k^2}{(\sigma_k^2+\nu)^2} \;\leq\; \sum_{k\in\mathcal{K}} N_k \frac{\sigma_k^2}{\sigma_k^2+\nu} = m.$$

*Item (iv), lower bound.* Let $J$ be the largest degree in $\mathcal{K}$ with $J \leq \alpha$, and let $D_J = \sum_{k\leq J} N_k \asymp d^J$ be the number of eigenvalues of $\Psi$ of degree at most $J$. The first degree in $\mathcal{K}$ above $J$ exceeds $\alpha$, and $\sigma_k^2$ decreases in $k$ (Claim 2), so by Lemma 4

$$\sup_{k>J} \frac{\sigma_k^2}{\nu} \longrightarrow 0, \qquad \text{hence} \qquad \inf_{k>J} \frac{\nu}{\sigma_k^2+\nu} \;\geq\; 1-o(1).$$

We bound the trace from below by retaining the degrees above $J$ and dropping the degrees up to $J$, which is valid because every summand is nonnegative:

$$\nu\operatorname{tr}[\Psi(\Psi+\nu I)^{-2}] = \sum_{k\in\mathcal{K}} N_k \,\frac{\sigma_k^2}{\sigma_k^2+\nu}\,\frac{\nu}{\sigma_k^2+\nu} \;\geq\; (1-o(1)) \sum_{k>J} N_k \frac{\sigma_k^2}{\sigma_k^2+\nu}.$$

By the fixed-point equation (3.1), the last sum is $m$ minus the sum over $k \leq J$, and each summand there is at most $N_k$:

$$\sum_{k>J} N_k \frac{\sigma_k^2}{\sigma_k^2+\nu} \;\geq\; m - D_J.$$

Finally, $D_J/m \to 0$ if $J < \alpha$, since $D_J \asymp d^J$, and $D_J/m \to 1/J! \leq \frac{1}{2}$ if $J = \alpha \geq 2$, by (C.9). In both cases $m - D_J \gtrsim m$, which completes the proof. □

The pure-target estimates of Appendices C and D bound $\overline{H}$ at the harmonic stopping times $T_\gamma$. The all-time statements below need the same bound at an arbitrary multiple of a harmonic time scale, which is the following lemma. Its proof is the upper bound on term 4)(2) of Appendix D.1, with the stopping time left free.

**Lemma 6** (Variance filter bound)**.** Fix $\alpha > 1$, $m = d^\alpha$, and a harmonic degree $j \in \mathcal{K}$ with $j < \alpha$. There is a constant $C_{\alpha,j}$ such that for all $x \geq 0$ and all large $d$,

$$\overline{H}(x/\sigma_j^2) \leq C_{\alpha,j}\,(1+x^2)\,d^{2j-2\alpha}. \tag{E.3}$$

*Proof.* Multiply numerator and denominator of $\overline{H}$ by $\nu$. By Lemma 5 the denominator $\nu\operatorname{tr}[\Psi(\Psi+\nu I)^{-2}]$ is of order $m = d^\alpha$, and the numerator is

$$\sum_{k\in\mathcal{K}} N_k \,\frac{\nu\sigma_k^2}{(\sigma_k^2+\nu)^2}\,(1-e^{-\sigma_k^2 t})^2, \qquad t = x/\sigma_j^2.$$

For $k \leq j < \alpha$, bound the filter by 1 and $\frac{\nu\sigma_k^2}{(\sigma_k^2+\nu)^2} \leq \frac{\nu}{\sigma_k^2+\nu} \asymp d^{k-\alpha}$ (Lemma 5), so with $N_k \asymp d^k$ these degrees contribute at most a constant times $d^{2j-\alpha}$. For $k > j$, use $1-e^{-\sigma_k^2 t} \leq \sigma_k^2 t = x\sigma_k^2/\sigma_j^2$ and $\frac{\nu\sigma_k^2}{(\sigma_k^2+\nu)^2}\sigma_k^4 \leq \nu\sigma_k^2$, so these degrees contribute at most

$$\frac{x^2}{\sigma_j^4}\,\nu\sum_k N_k\sigma_k^2 = \frac{x^2}{\sigma_j^4}\,\nu\operatorname{tr}\Psi \asymp x^2\,d^{2j+2}\,d^{-\alpha-1}\,d^{-1} = x^2 d^{2j-\alpha},$$

by $\sigma_j^2 \asymp d^{-j-1}$, $\nu \asymp d^{-\alpha-1}$ and $\operatorname{tr}\Psi \asymp d^{-1}$ from Lemmas 4 and 5. Dividing by the denominator gives (E.3). □

### E.1 A SMALL COMPONENT ABOVE THE TEACHER'S CAPACITY: NO POLYNOMIAL IMPROVEMENT

**Proposition 2** (No polynomial improvement at any stopping time)**.** Let $m = d^3$ and let $f_d^\star$ be the two-mode target (6.5) with $w_d = c/d$ for a constant $c > 0$. Then

$$\overline{\mathcal{L}}_{\mathrm{TE}} \asymp d^{-1}, \qquad \inf_{t \geq 0} \overline{\mathcal{L}}_{\mathrm{ST}}(t) \asymp d^{-1}, \tag{E.4}$$

and the upper bound is attained at $T_2 = 4\log(d)/\sigma_2^2$. In particular, both errors vanish, but for every $\eta > 0$ the best student error is not $O(d^{-\eta}\,\overline{\mathcal{L}}_{\mathrm{TE}})$: the student improves on the teacher by at most a constant factor, whatever the stopping time.

*Proof.* Here $\alpha = 3$, and Lemmas 4 and 5 give $\nu \asymp d^{-4}$, $\frac{\nu}{\sigma_2^2+\nu} \asymp d^{-1}$, and, since the degree-four eigenvalue lies below the capacity edge, $\frac{\nu}{\sigma_4^2+\nu} \to 1$. By (E.1), $\overline{\mathcal{L}}_{\mathrm{TE}} = (1-w_d)\frac{\nu}{\sigma_2^2+\nu} + w_d\frac{\nu}{\sigma_4^2+\nu}$, and both summands are of order $d^{-1}$.

*Lower bound.* We bound $\overline{\mathcal{L}}_{\mathrm{ST}}(t)$ from below by retaining only the degree-four bias, without the exponential inside it, and dropping every other term of (E.1)–(E.2); this is valid because all of these terms are nonnegative, and it is the first inequality of Theorem 1 on one block. For every $t \geq 0$,

$$\overline{\mathcal{L}}_{\mathrm{ST}}(t) \;\geq\; w_d\,\overline{\ell}_4(t) \;\geq\; w_d\Big(\frac{\nu}{\sigma_4^2+\nu}\Big)^2 \;\asymp\; d^{-1}. \tag{E.5}$$

This is the energy of the degree-four mode, which the teacher does not learn and which the student therefore never learns either.

*Upper bound.* At $t = T_2$, that is $x = 4\log d$ on the degree-two time scale, the degree-two bias in (E.2) is $\big(\frac{\nu}{\sigma_2^2+\nu} + O(d^{-4})\big)^2 \asymp d^{-2}$, and by Lemma 6 with $j = 2$ the degree-two variance is at most $\frac{\nu}{\sigma_2^2+\nu}\overline{H}(T_2) \lesssim d^{-1}(\log d)^2 d^{-2}$. Hence $\overline{\ell}_2(T_2) \asymp d^{-2}$, and with $\overline{\ell}_4 \leq \frac{5}{4}$,

$$\overline{\mathcal{L}}_{\mathrm{ST}}(T_2) \leq \overline{\ell}_2(T_2) + \tfrac{5}{4}w_d \lesssim d^{-1}.$$

Together with (E.5) this proves (E.4). □

In the classification of Appendix D.2 this target has $k_0 = 4 > \alpha = 3$, and the stop $T_2$ falls in the first row of the table there: the held-out energy $\mathcal{E}_{>2} = w_d$ is of the same order as $\overline{\mathcal{L}}_{\mathrm{TE}}$, so the condition (6.3) fails. The proposition adds that no other stopping time does better, because the obstruction (E.5) is a floor at every time.

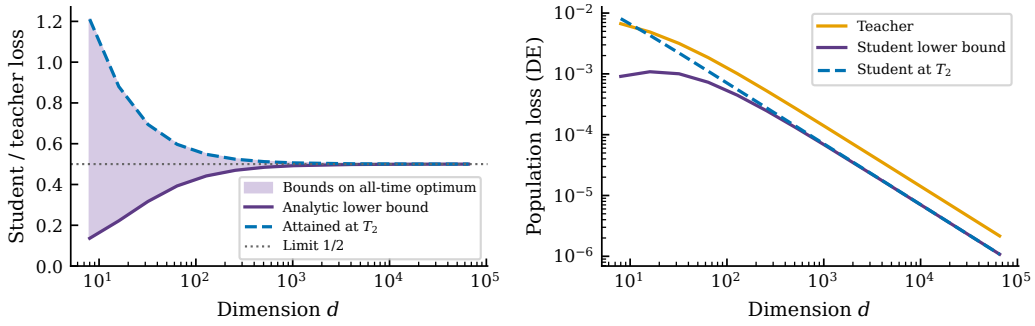



Figure 5: Proposition 2 for the deterministic equivalent. The floor $w_d\big(\frac{\nu}{\sigma_4^2+\nu}\big)^2/\overline{\mathcal{L}}_{\mathrm{TE}}$ of (E.5) and the attained ratio $\overline{\mathcal{L}}_{\mathrm{ST}}(T_2)/\overline{\mathcal{L}}_{\mathrm{TE}}$ bracket the all-time optimum. The figure uses $c = (\pi-3)/2$, for which both tend to $1/2$ (Remark 3). The companion panel shows that the teacher error and the two student bounds all vanish at rate $d^{-1}$. These are deterministic calculations, not random-teacher observations.

### E.2 Two modes below capacity with incompatible stopping times: no quadratic improvement

**Proposition 3** (No quadratic improvement at any stopping time)**.** Let $m = d^5$ and let $f_d^\star$ be the two-mode target (6.5) with $w_d = Bd^{-5}$ for a constant $B > 0$. Then

$$\overline{\mathcal{L}}_{\mathrm{TE}} \asymp d^{-3}, \qquad \inf_{t\geq 0} \overline{\mathcal{L}}_{\mathrm{ST}}(t) \asymp d^{-5}, \tag{E.6}$$

and the upper bound is attained at $t = x/\sigma_4^2$ for every fixed $x > 0$, with constants depending on $x$. Consequently

$$\frac{\inf_t \overline{\mathcal{L}}_{\mathrm{ST}}(t)}{\overline{\mathcal{L}}_{\mathrm{TE}}} \asymp d^{-2} \longrightarrow 0, \qquad \frac{\inf_t \overline{\mathcal{L}}_{\mathrm{ST}}(t)}{\overline{\mathcal{L}}_{\mathrm{TE}}^2} \asymp d \longrightarrow \infty : \tag{E.7}$$

weak-to-strong generalization occurs, but no stopping time gives quadratic improvement. The best improvement exponent is $5/3$.

*Proof.* Here $\alpha = 5$, and Lemmas 4 and 5 give

$$\nu \asymp d^{-6}, \qquad \frac{\nu}{\sigma_2^2+\nu} \asymp d^{-3}, \qquad \frac{\nu}{\sigma_4^2+\nu} \asymp d^{-1}, \qquad \nu \operatorname{tr}[\Psi(\Psi+\nu I)^{-2}] \asymp d^5.$$

By (E.1), $\overline{\mathcal{L}}_{\mathrm{TE}} = (1-w_d)\frac{\nu}{\sigma_2^2+\nu} + w_d \frac{\nu}{\sigma_4^2+\nu} \asymp d^{-3}$, the second summand being of order $d^{-6}$.

*Lower bound.* Fix $t \geq 0$ and write $x = \sigma_4^2 t$. We bound $\overline{\mathcal{L}}_{\mathrm{ST}}(t)$ from below by retaining two terms of (E.1)–(E.2) and dropping all the others, which is valid because every term is nonnegative. The first retained term is the degree-four bias, in which $\frac{\nu+\sigma_4^2 e^{-x}}{\sigma_4^2+\nu} \geq \frac{(\nu+\sigma_4^2)e^{-x}}{\sigma_4^2+\nu} = e^{-x}$, so

$$w_d\, \overline{\ell}_4(t) \;\geq\; w_d\, e^{-2x}.$$

The second retained term is the degree-two variance, which we bound from below by retaining only the degree-four summand of the numerator trace of $\overline{H}$ and dropping the other summands, again valid because they are nonnegative:

$$(1-w_d)\,\overline{\ell}_2(t) \;\geq\; (1-w_d)\,\frac{\nu\sigma_2^2}{(\sigma_2^2+\nu)^2}\,\frac{N_4\,\frac{\nu\sigma_4^2}{(\sigma_4^2+\nu)^2}}{\nu\operatorname{tr}[\Psi(\Psi+\nu I)^{-2}]}\,(1-e^{-x})^2 =: A_d\,(1-e^{-x})^2,$$

where, by the estimates above and $N_4 \asymp d^4$,

$$A_d \asymp d^{-3}\cdot\frac{d^4\cdot d^{-1}}{d^5} = d^{-5}.$$

Combining the two retained terms, for every $t \geq 0$,

$$\overline{\mathcal{L}}_{\mathrm{ST}}(t) \;\geq\; w_d\, e^{-2x} + A_d\,(1-e^{-x})^2 \;\geq\; \min_{0\leq y\leq 1}\left[w_d y^2 + A_d(1-y)^2\right] = \frac{w_d A_d}{w_d + A_d} \;\asymp\; d^{-5}, \tag{E.8}$$

since $w_d$ and $A_d$ are both of order $d^{-5}$. The minimum over $y = e^{-x}$ is computed by completing the square.

*Upper bound.* Let $t = x/\sigma_4^2$ with $x > 0$ fixed. The degree-four component contributes at most $\frac{5}{4}w_d \lesssim d^{-5}$. For the degree-two component, $\sigma_2^2 t = x\sigma_2^2/\sigma_4^2 \asymp xd^2$, so the exponential in its bias is negligible and the bias is $\asymp \left(\frac{\nu}{\sigma_2^2+\nu}\right)^2 \asymp d^{-6}$; its variance is at most $\frac{\nu}{\sigma_2^2+\nu}\overline{H}(t) \lesssim d^{-3}(1+x^2)d^{-2}$ by Lemma 6 with $j = 4$. Hence $\overline{\mathcal{L}}_{\mathrm{ST}}(x/\sigma_4^2) \lesssim d^{-5}$, which with (E.8) proves (E.6); (E.7) follows by division. □

Each pure mode separately admits quadratic improvement under its own stopping rule (Theorem 4 with $k_0 = 2$ or $k_0 = 4$ and $\alpha = 5$). The mixture cannot: learning the degree-four signal requires $\sigma_4^2 t$ of order one, and the same filter $1 - e^{-\sigma_4^2 t}$ admits the degree-four error that the teacher makes on the dominant degree-two mode (E.8). The harmonic stop $T_4$ of Section 6 also attains the order $d^{-5}$: this is the guarantee (6.1) for $\gamma = 4$, whose upper bound $d^{k_0-\alpha}\overline{\mathcal{L}}_{\mathrm{TE}} \asymp d^{-4}$ is not attained by this target because its degree-four energy is small.

**Remark 3** (Leading constants used in the figures)**.** The constants $c$ and $B$ play no role in Propositions 2 and 3, and the propositions are stated up to constants. The figures of this appendix, Figure 1 and Figure 4 use the particular values $c = (\pi - 3)/2$ and $B = C^2/24$ with $C = (12\pi - 37)/24$, for which the two competing error terms in each example are equal at leading order, and they are normalized by the leading constants listed in Table 3. These constants are obtained by carrying the constants through the proofs above, using the exact expressions for $N_k$ and $\sigma_k^2$ of Appendix C.1, a routine calculation that AI tools can perform; we record them here as claimed, for the reproducibility of the figures, and do not prove them. The numerical evaluations of Appendix G agree with them.

| Target, width | $\overline{\mathcal{L}}_{\mathrm{TE}}$ | $\inf_{t\geq 0} \overline{\mathcal{L}}_{\mathrm{ST}}(t)$ | Shown in the figures |
|---|---|---|---|
| Pure $Y_{2,d}$, $m = d^3$ | $\sim c/d$ | $\overline{\mathcal{L}}_{\mathrm{ST}}(T_2) \sim c^2/d^2$ | Figure 1: ratio $\to 1$ |
| Proposition 2, $m = d^3$ | $\sim 2c/d$ | $\sim c/d$, attained at $T_2$ | Figure 5: both ratios $\to 1/2$ |
| Proposition 3, $m = d^5$ | $\sim Cd^{-3}$ | $\sim \dfrac{C^2}{48} d^{-5}$, attained at $\sigma_4^2 t = \log 2$ | Figures 4, 7: $\inf_t \overline{\mathcal{L}}_{\mathrm{ST}}(t)/\overline{\mathcal{L}}_{\mathrm{TE}} \sim C/(48d^2)$, $\inf_t \overline{\mathcal{L}}_{\mathrm{ST}}(t)/\overline{\mathcal{L}}_{\mathrm{TE}}^2 \sim d/48$, $\inf_t \overline{\mathcal{L}}_{\mathrm{ST}}(t)/\overline{\mathcal{L}}_{\mathrm{TE}}^{5/3} \to C^{1/3}/48$ |

Table 3: Claimed leading constants behind the figures, with $c = (\pi - 3)/2$, $C = (12\pi - 37)/24$ and $B = C^2/24$. The ratio in the last column is $\overline{\mathcal{L}}_{\mathrm{ST}}(T_2)/\overline{\mathcal{L}}_{\mathrm{TE}}^2$ in the first row, the two student-to-teacher ratios of Figure 5 in the second, and the three ratios named in the third.

# F Other activations

The proofs of Appendix C compute the ReLU coefficients $\sigma_k$ exactly, but the conclusion of Theorem 2 uses only a few coarse properties of the spectrum: the orders of $N_k$ and $\sigma_k^2$ at finitely many fixed degrees, a bound on the trace of $\Psi$, and the presence of the target degree in the kernel. In this appendix we show that these properties hold for a broad class of activations, and we prove the analogue of Theorem 2 for them (Theorem 9). The tanh and logistic sigmoid activations are worked out as examples. As elsewhere, all statements concern the deterministic equivalents (3.2) and (3.3); the transfer to the random spherical model requires the analogue of Assumption 1 for the activation in question.

**Scaling.** For $u, x$ uniform on $\mathbb{S}^{d-1}$ the preactivation $\langle u, x\rangle$ is of order $d^{-1/2}$. ReLU is positively homogeneous, so this scale only multiplies the kernel by a constant: $(\langle u, x\rangle)_+ = d^{-1/2}(\sqrt{d}\,\langle u, x\rangle)_+$. For a general activation the scale matters, and in this appendix we use the features $\sigma(\sqrt{d}\,\langle u, x\rangle)$, whose preactivations are approximately standard Gaussian. For ReLU this multiplies $\Psi$ by $d$. Under $\Psi \mapsto c\Psi$ the fixed point $\nu$ of (3.1) becomes $c\nu$, the stopping time (4.2) becomes $T/c$, and both $\overline{\mathcal{L}}_{\mathrm{TE}}$ and $\overline{\mathcal{L}}_{\mathrm{ST}}(T)$ are unchanged, so the rescaling has no effect on the results.

**Harmonic coefficients.** Let $t = \langle u, x\rangle$ for independent uniform $u, x \in \mathbb{S}^{d-1}$, and let $P_k$ be the degree-$k$ Gegenbauer polynomial normalized by $P_k(1) = 1$. These polynomials are orthogonal for the law of $t$, with $\mathbb{E}[P_k(t)^2] = 1/N_k$, and the addition theorem states that

$$\sum_{l=1}^{N_k} \phi_{k,l}(u)\phi_{k,l}(x) = N_k P_k(\langle u, x\rangle).$$

Expanding $\sigma(\sqrt{d}\,t)$ in the $P_k$ and applying the addition theorem gives the Funk–Hecke expansion

$$\sigma(\sqrt{d}\,\langle u, x\rangle) = \sum_{k\geq 0} \sigma_k \sum_{l=1}^{N_k} \phi_{k,l}(u)\phi_{k,l}(x), \qquad \sigma_k = \mathbb{E}\big[\sigma(\sqrt{d}\,t)P_k(t)\big]. \tag{F.1}$$

Thus the feature $\sigma(\sqrt{d}\,\langle u, \cdot\rangle)$ has coordinates $\sigma_k\phi_{k,l}(u)$, and $\Psi$ has blocks $\sigma_k^2 I_{N_k}$, as in (4.1). For ReLU, $\sigma_k$ is $\sqrt{d}$ times the coefficient of Appendix C.

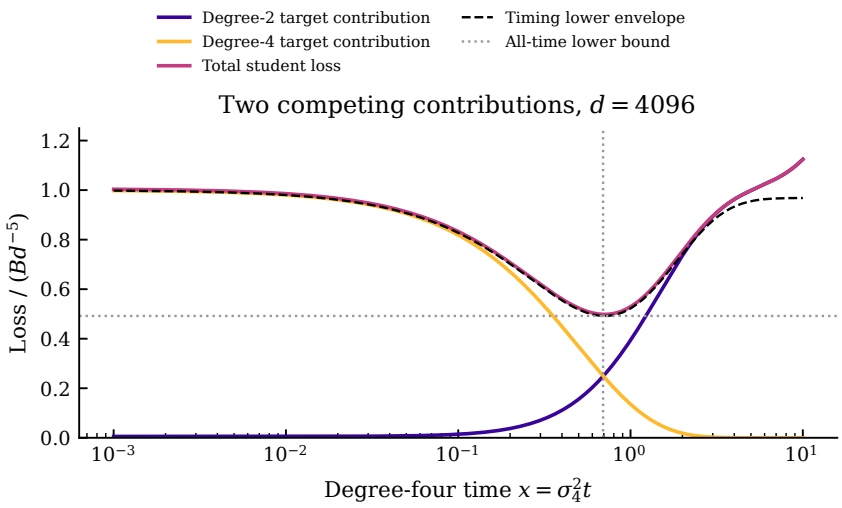



Figure 6: The timing conflict of Proposition 3 at $d = 4096$, with $B = C^2/24$ as in Remark 3 and loss divided by $Bd^{-5}$. The degree-four signal contribution decreases while the error inherited from the dominant degree-two mode increases. Their sum lies above the dashed envelope of (E.8). Dotted lines mark its all-time lower bound and the asymptotically optimal time $\sigma_4^2 t = \log 2$. The plot focuses on the degree-four learning window; all curves are theoretical.

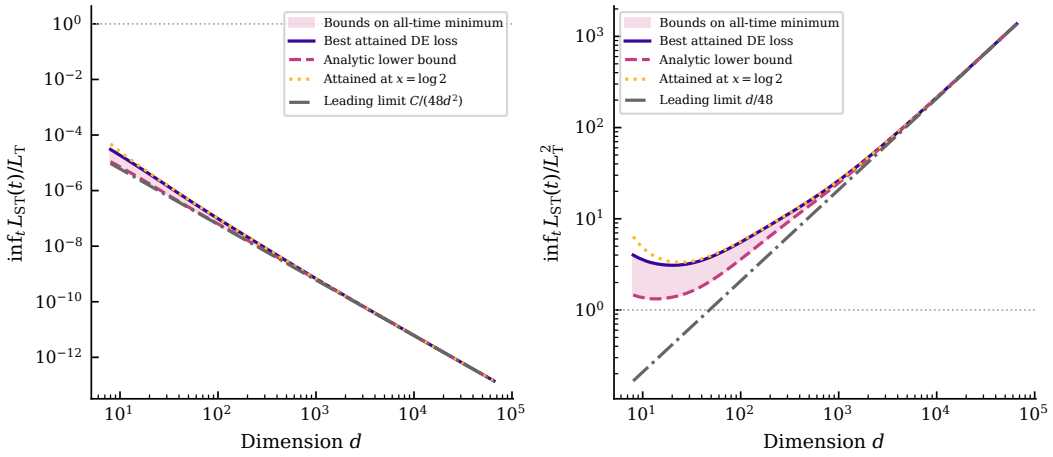



Figure 7: Bounds on the all-time minimum in Proposition 3, with $B = C^2/24$ as in Remark 3. The shaded band lies between the analytic lower bound (E.8) and the best numerically attained loss. The explicit stop $\sigma_4^2 t = \log 2$ is another upper bound. The leading ratios $C/(48d^2)$ and $d/48$ of Remark 3 are shown separately: WTS occurs, but no stopping sequence achieves quadratic WTS. This is a conditional DE calculation.

**Hermite coefficients.** Let $z \sim \mathcal{N}(0,1)$ and let $\mathrm{He}_k$ be the probabilists' Hermite polynomials, so that the $\mathrm{He}_k/\sqrt{k!}$ are orthonormal for the law of $z$. Set

$$c_k = \frac{\mathbb{E}[\sigma(z)\mathrm{He}_k(z)]}{\sqrt{k!}}, \qquad \mathcal{K}_\sigma = \{k \geq 0 : \ c_k \neq 0\}.$$

Throughout, $\sigma$ is measurable with $|\sigma(r)| \leq Ce^{C|r|}$ for some $C$. Since the Hermite polynomials are complete, $\mathcal{K}_\sigma$ is infinite unless $\sigma$ is a polynomial.

**Lemma 7** (Harmonic and Hermite coefficients)**.**

(i) $\operatorname{tr}\Psi = \sum_{k\geq 0} N_k\sigma_k^2 = \mathbb{E}\big[\sigma(\sqrt{d}\,\langle u,x\rangle)^2\big] \to \mathbb{E}[\sigma(z)^2]$ as $d\to\infty$. In particular $\operatorname{tr}\Psi \lesssim 1$.

(ii) For each fixed $k$, $N_k\sigma_k^2 \to c_k^2$ as $d\to\infty$. In particular $\sigma_k^2 \asymp d^{-k}$ for each fixed $k \in \mathcal{K}_\sigma$.

(iii) If $\sigma$ is odd, then $\sigma_k = 0$ for every $d$ and every even $k$. If $\sigma$ is even, then $\sigma_k = 0$ for every $d$ and every odd $k$.

Item (ii) is standard (Ghorbani et al., 2021); we include the short proof.

*Proof.* The polynomials $h_k = \sqrt{N_k}\,P_k$ are orthonormal for the law of $t$, and by (F.1)

$$N_k\sigma_k^2 = \big(\mathbb{E}[\sigma(\sqrt{d}\,t)\,h_k(t)]\big)^2. \tag{F.2}$$

In words, $N_k\sigma_k^2$ is the squared $k$-th coefficient of $\sigma(\sqrt{d}\,\cdot)$ in the orthonormal polynomials of the law of $t$, while $c_k^2$ is the squared $k$-th coefficient of $\sigma$ in the orthonormal polynomials of the Gaussian.

*(i)* The first equality holds because $\Psi$ has blocks $\sigma_k^2 I_{N_k}$. The law of $t$ has compact support, so polynomials are dense in its $L^2$ space and $(h_k)_{k\geq 0}$ is an orthonormal basis. The second equality is therefore Parseval's identity, by (F.2). For the limit, $\sqrt{d}\,t$ has the density

$$\rho_d(z) = \kappa_d\Big(1-\frac{z^2}{d}\Big)^{\frac{d-3}{2}}, \qquad |z| < \sqrt{d},$$

where $\kappa_d$ is a normalizing constant. Since $\log(1-y) \leq -y$, we have $\rho_d(z) \leq Ce^{-z^2/4}$ for $d \geq 6$, and $\rho_d$ converges pointwise to the standard Gaussian density. By dominated convergence,

$$\mathbb{E}\big[g(\sqrt{d}\,t)\big] \ \longrightarrow\ \mathbb{E}[g(z)] \qquad \text{for every } g \text{ with } |g(r)| \leq Ce^{C|r|},$$

and in particular for $g = \sigma^2$.

*(ii)* In the variable $z = \sqrt{d}\,t$, the polynomial $h_k(z/\sqrt{d})$ is the degree-$k$ orthonormal polynomial of $\rho_d$ with positive leading coefficient. It is obtained from $1, z, \dots, z^k$ by Gram–Schmidt, so its coefficients are continuous functions of the moments of $\rho_d$ of order at most $2k$, as long as the moment matrix is nonsingular. By the dominated convergence in (i), these moments converge to those of the Gaussian, whose moment matrix is positive definite. Hence the coefficients of $h_k(z/\sqrt{d})$ converge to those of $\mathrm{He}_k(z)/\sqrt{k!}$, and a second application of dominated convergence gives

$$\mathbb{E}\big[\sigma(\sqrt{d}\,t)\,h_k(t)\big] \ \longrightarrow\ \frac{\mathbb{E}[\sigma(z)\mathrm{He}_k(z)]}{\sqrt{k!}} = c_k.$$

By (F.2), $N_k\sigma_k^2 \to c_k^2$. Since $N_k \asymp d^k$, this gives $\sigma_k^2 \asymp d^{-k}$ when $c_k \neq 0$.

*(iii)* The law of $t$ is symmetric, so $P_k(-t) = (-1)^k P_k(t)$. If $\sigma$ is odd and $k$ is even, the function $t \mapsto \sigma(\sqrt{d}\,t)P_k(t)$ is odd, and $\sigma_k = 0$ by (F.1). The case of $\sigma$ even and $k$ odd is the same. □

**Theorem 9** (Quadratic improvement for general activations)**.** Let $\sigma$ be as above, let $f^\star$ be a unit spherical harmonic of fixed degree $k_0 \in \mathcal{K}_\sigma$, $k_0 \geq 1$, and let $m = d^\alpha$ with fixed $\alpha > 0$. Assume that $c_{k_\star} \neq 0$ for some $k_\star > \alpha$; this holds for every $\alpha$ when $\sigma$ is not a polynomial. Then, with the stopping time $T$ of (4.2),

$$\overline{\mathcal{L}}_{\mathrm{TE}} \asymp d^{-(\alpha-k_0)_+}, \qquad \overline{\mathcal{L}}_{\mathrm{ST}}(T) \asymp \overline{\mathcal{L}}_{\mathrm{TE}}^2.$$

Together with Theorem 1, this gives optimal quadratic improvement for every such activation once $\alpha > k_0$. The powers of $d$ are those of Theorem 2 and do not depend on the activation. The only structural hypothesis is that the kernel sees the target degree, which is the hypothesis $k_0 \in \mathcal{K}$ of Theorem 2 with $\mathcal{K}$ replaced by $\mathcal{K}_\sigma$.

*Proof.* All constants below depend only on $\sigma$, $\alpha$, and $k_0$, and $d$ is taken large enough.

*Step 1: $\nu \asymp d^{-\alpha}$.* Since $\sigma_k^2/(\sigma_k^2+\nu) \le \sigma_k^2/\nu$, the fixed point (3.1) gives

$$d^\alpha = \sum_{k\ge 0} \frac{N_k \sigma_k^2}{\sigma_k^2+\nu} \le \frac{\operatorname{tr}\Psi}{\nu}, \qquad \text{so} \qquad \nu \lesssim d^{-\alpha}$$

by Lemma 7(i). For the lower bound we keep only the degree $k_\star$:

$$d^\alpha \;\ge\; \frac{N_{k_\star}\sigma_{k_\star}^2}{\sigma_{k_\star}^2+\nu}.$$

If $\nu \le \sigma_{k_\star}^2$, the right-hand side is at least $N_{k_\star}/2 \asymp d^{k_\star}$, which exceeds $d^\alpha$ because $k_\star > \alpha$. Hence $\nu > \sigma_{k_\star}^2$, and

$$\nu \;\ge\; \frac{N_{k_\star}\sigma_{k_\star}^2}{2\,d^\alpha} \;\gtrsim\; d^{-\alpha},$$

since $N_{k_\star}\sigma_{k_\star}^2 \to c_{k_\star}^2 > 0$ by Lemma 7(ii). This is the argument of Lemma 4.

*Step 2: the teacher.* Write $\lambda = \sigma_{k_0}^2$ and $L = \overline{\mathcal{L}}_{\mathrm{TE}}$. By Lemma 7(ii), $\lambda \asymp d^{-k_0}$. For a unit target of degree $k_0$, (3.2) and Step 1 give

$$L = \frac{\nu}{\lambda+\nu} \;\asymp\; \min\{1,\, d^{k_0}\nu\} \;\asymp\; d^{-(\alpha-k_0)_+}.$$

*Step 3: the student.* The lower bound $\overline{\mathcal{L}}_{\mathrm{ST}}(T) \ge L^2$ is Theorem 1. For the upper bound we bound the two terms of (3.3). The choice of $T$ gives $e^{-\lambda T} = L/m$, so the squared bias is

$$\overline{B}(T) = \Big(\frac{\nu+\lambda e^{-\lambda T}}{\lambda+\nu}\Big)^2 = L^2\Big(1+\frac{1-L}{m}\Big)^2 \le 4L^2.$$

In the variance, the prefactor is $\nu\lambda/(\lambda+\nu)^2 = L(1-L) \le L$, so

$$\overline{V}(T) \le L\,R(T), \qquad R(T) = \frac{\operatorname{tr}[\Psi(\Psi+\nu I)^{-2}(I-e^{-\Psi T})^2]}{\operatorname{tr}[\Psi(\Psi+\nu I)^{-2}]} \le 1.$$

If $\alpha \le k_0$, then $L \asymp 1$, and $\overline{\mathcal{L}}_{\mathrm{ST}}(T) \le 4L^2 + L \lesssim L^2$.

Let $\alpha > k_0$. It remains to show that $R(T) \lesssim L$. In the denominator we keep only the degree $k_\star$. Since $\sigma_{k_\star}^2 \asymp d^{-k_\star} \le \nu$,

$$\operatorname{tr}[\Psi(\Psi+\nu I)^{-2}] \;\ge\; \frac{N_{k_\star}\sigma_{k_\star}^2}{(\sigma_{k_\star}^2+\nu)^2} \;\ge\; \frac{N_{k_\star}\sigma_{k_\star}^2}{4\nu^2} \;\gtrsim\; \nu^{-2}.$$

In the numerator, we bound the degrees $k \le k_0$ using $(1-e^{-xT})^2 \le 1$ and $x/(x+\nu)^2 \le 1/(4\nu)$, and the degrees $k > k_0$ using $(1-e^{-xT})^2 \le x^2T^2$ and $x^2/(x+\nu)^2 \le 1$. This gives

$$\operatorname{tr}[\Psi(\Psi+\nu I)^{-2}(I-e^{-\Psi T})^2] \;\le\; \sum_{k\le k_0}\frac{N_k}{4\nu} + T^2\sum_{k>k_0} N_k\sigma_k^2 \;\lesssim\; \frac{d^{k_0}}{\nu} + d^{2k_0}(\log d)^2,$$

where we used $\operatorname{tr}\Psi \lesssim 1$ and $T = \lambda^{-1}\log(m/L) \lesssim d^{k_0}\log d$. Dividing, and using $\nu \asymp d^{-\alpha}$ and $k_0 < \alpha$,

$$R(T) \;\lesssim\; \nu^2\Big(\frac{d^{k_0}}{\nu} + d^{2k_0}(\log d)^2\Big) \;\asymp\; d^{-(\alpha-k_0)} + d^{-2(\alpha-k_0)}(\log d)^2 \;\lesssim\; d^{-(\alpha-k_0)} \;\asymp\; L.$$

Hence $\overline{V}(T) \lesssim L^2$, and $\overline{\mathcal{L}}_{\mathrm{ST}}(T) = \overline{B}(T) + \overline{V}(T) \lesssim L^2$. □

**Remark 4** (The hypotheses)**.** The hypothesis $k_0 \in \mathcal{K}_\sigma$ cannot be dropped. If $\sigma$ has the parity opposite to that of $k_0$, then $\sigma_{k_0} = 0$ for every $d$ by Lemma 7(iii). The target is then orthogonal to every feature, so the teacher's best fit is zero; the student, which starts at zero and is trained on the teacher, stays at zero; and both errors equal one at all times. The hypothesis $c_{k_\star} \neq 0$ for some $k_\star > \alpha$ fails only for a polynomial $\sigma$ of degree $p \leq \alpha$. Then $\sigma_k = 0$ for every $k > p$ and every $d$, so $\Psi$ has rank at most $\sum_{k \leq p} N_k \asymp d^p$. For $\alpha > p$ this is smaller than $m$, the fixed point (3.1) has no solution, and the teacher represents every target in the span of the features exactly.

The proof uses Lemma 7 only through $\operatorname{tr} \Psi \lesssim 1$ and $N_k \sigma_k^2 \asymp 1$ at the two degrees $k_0$ and $k_\star$. It uses neither the exact values of the $\sigma_k$, nor monotonicity of $\sigma_k^2$ in $k$, nor any rate of decay of the $c_k$. We expect the stopping-time phases of Theorem 4 and the mixed-target examples of Appendix E to carry over in the same way, with $\mathcal{K}$ replaced by $\mathcal{K}_\sigma$, since they use only the orders of $N_k$ and $\sigma_k^2$; we have not written this out.

**Example 1** (ReLU)**.** Let $\sigma(r) = r_+$, and let $\varphi$ be the standard Gaussian density. Two integrations by parts, using $\mathrm{He}_k \varphi = -(\mathrm{He}_{k-1}\varphi)'$, give

$$\mathbb{E}[z_+ \mathrm{He}_k(z)] = \varphi(0)\, \mathrm{He}_{k-2}(0), \qquad k \geq 2,$$

which vanishes exactly when $k$ is odd. Hence $\mathcal{K}_\sigma = \{0, 1\} \cup 2\mathbb{N} = \mathcal{K}$, with $c_0^2 = 1/(2\pi)$ and $c_1^2 = 1/4$, and Theorem 9 recovers Theorem 2. The coefficients $\sigma_k$ with $k \geq 3$ odd vanish exactly at every $d$, by Lemma 7(iii): $r_+ - r/2 = |r|/2$ is even, and the linear part $r/2$ contributes only to degree one.

**Example 2** (tanh)**.** Let $\sigma = \tanh$. We show that $\mathcal{K}_\sigma$ is the set of odd integers. Since $\tanh$ is odd, $c_k = 0$ for every even $k$, and $\sigma_k = 0$ exactly at every even degree by Lemma 7(iii). Let $k$ be odd. One Gaussian integration by parts gives

$$\mathbb{E}[\tanh(z)\mathrm{He}_k(z)] = \mathbb{E}[\operatorname{sech}^2(z)\mathrm{He}_{k-1}(z)].$$

The function $\operatorname{sech}^2$ has a positive, integrable Fourier transform,

$$\operatorname{sech}^2(r) = \frac{1}{2\pi} \int_{\mathbb{R}} \frac{\pi\xi}{\sinh(\pi\xi/2)}\, e^{i\xi r}\, d\xi,$$

and $\mathbb{E}[e^{i\xi z}\mathrm{He}_{k-1}(z)] = (i\xi)^{k-1} e^{-\xi^2/2}$. Exchanging the order of integration, and using that $k-1$ is even,

$$\mathbb{E}[\tanh(z)\mathrm{He}_k(z)] = \frac{(-1)^{(k-1)/2}}{2\pi} \int_{\mathbb{R}} \xi^{k-1} \frac{\pi\xi}{\sinh(\pi\xi/2)}\, e^{-\xi^2/2}\, d\xi.$$

The integrand is positive, so $c_k \neq 0$, with sign $(-1)^{(k-1)/2}$. Hence Theorem 9 applies to every odd target degree $k_0$ and every $\alpha > 0$:

$$\overline{\mathcal{L}}_{\mathrm{TE}} \asymp d^{-(\alpha - k_0)_+}, \qquad \overline{\mathcal{L}}_{\mathrm{ST}}(T) \asymp \overline{\mathcal{L}}_{\mathrm{TE}}^2,$$

exactly as for ReLU. For an even target degree nothing is learned, by Remark 4. The argument uses only that $\sigma$ is odd and that $\sigma'$ has a nonnegative Fourier transform that is not identically zero; it applies, for instance, to $\sigma = \operatorname{erf}$.

**Example 3** (Logistic sigmoid)**.** Let $s(r) = 1/(1 + e^{-r}) = \frac{1}{2} + \frac{1}{2}\tanh(r/2)$. The function $s - \frac{1}{2}$ is odd, and the constant $\frac{1}{2}$ contributes only to degree zero. Hence $c_0 = \frac{1}{2}$ and $c_k = 0$ for every even $k \geq 2$, and by Lemma 7(iii) also $\sigma_k = 0$ for every even $k \geq 2$ and every $d$. For odd $k$, the argument of Example 2 applies with $\operatorname{sech}^2$ replaced by $s'(r) = \frac{1}{4}\operatorname{sech}^2(r/2)$, whose Fourier transform $\pi\xi/\sinh(\pi\xi)$ is again positive. Hence $\mathcal{K}_s$ consists of 0 and the odd integers, and the conclusions of Example 2 hold unchanged.

A bias breaks the parity. For the activation $r \mapsto s(r + b)$, the same computation gives, for $k \geq 1$,

$$\mathbb{E}[s(z + b)\mathrm{He}_k(z)] = \frac{1}{2\pi} \int_{\mathbb{R}} \frac{\pi\xi}{\sinh(\pi\xi)}\, e^{i\xi b}\, (i\xi)^{k-1} e^{-\xi^2/2}\, d\xi.$$

As a function of $b$ this extends to an entire function. For odd $k$ it is nonzero at $b = 0$, and for even $k$ its derivative at $b = 0$ is nonzero by the same positivity. So it is not identically zero and has countably many real zeros. Therefore, for all $b$ outside a countable set, $c_k \neq 0$ for every $k$, and Theorem 9 applies to targets of every degree $k_0 \geq 1$.

# G NUMERICAL DETAILS FOR THE MAIN FIGURES

The main figures contain two kinds of curves. The lines are evaluations of the deterministic equivalents (3.1)–(3.3). The markers in Figure 1 are averages over random teachers of the population flow of Section 2, simulated as described next.

## G.1 SIMULATING THE POPULATION FLOW

A population-flow sample is one draw of the model of Section 2: the teacher's feature directions $u_1, \dots, u_m$ are drawn independently and uniformly on $\mathbb{S}^{d-1}$ with $m = d^3$, the teacher is fitted by minimizing its population loss, and the student is the population spectral flow (2.1) against that teacher. Every expectation over the data is computed in closed form, so the feature directions are the only randomness.

**Teacher.** Writing the teacher as $g(x) = \sum_{i=1}^{m} a_i \sigma(\langle u_i, x\rangle)$, its population least-squares weights solve $Ga = h$ with

$$G_{ij} = \mathbb{E}_x\big[\sigma(\langle u_i, x\rangle)\sigma(\langle u_j, x\rangle)\big], \qquad h_i = \mathbb{E}_x\big[\sigma(\langle u_i, x\rangle) f^\star(x)\big].$$

Both are available in closed form. For ReLU on the sphere, with $s = \langle u_i, u_j\rangle$, the kernel is $\big(\sqrt{1-s^2} + (\pi - \arccos s)\, s\big)/(2\pi d)$, and for a pure degree-$k_0$ harmonic target $h_i = \sigma_{k_0} f^\star(u_i)$. The simulations use the population-centered activation $\sigma(r) = r_+ - \mu_d$, which subtracts $\mu_d^2$ from every kernel entry and removes the degree-zero mode of the kernel; the targets have no degree-zero component. The $m \times m$ system is solved in double precision; its condition number is below $10^5$ for $d \le 28$, and the relative residual $\|Ga - h\|/\|h\|$ is checked to be at roundoff level for every teacher. The teacher error is $\mathcal{L}_{\mathrm{TE}} = 1 - 2\langle g, f^\star\rangle_{\mathcal{D}} + \|g\|_{\mathcal{D}}^2$.

**Student.** The population flow (2.1) acts on each harmonic degree $k$ by the scalar filter $1 - e^{-\sigma_k^2 t}$. Writing $\Pi_k$ for the projection onto degree-$k$ harmonics, the student error at every time is therefore determined by the harmonic energies of the teacher,

$$\mathcal{L}_{\mathrm{ST}}(t) = \sum_k \Big[\|\Pi_k f^\star\|_{\mathcal{D}}^2 - 2(1 - e^{-\sigma_k^2 t})\langle \Pi_k g, \Pi_k f^\star\rangle_{\mathcal{D}} + (1 - e^{-\sigma_k^2 t})^2 \|\Pi_k g\|_{\mathcal{D}}^2\Big]. \tag{G.1}$$

The energies $\|\Pi_k g\|_{\mathcal{D}}^2 = \sigma_k^2\, a^\top H_k a$ with $(H_k)_{ij} = N_k P_k(\langle u_i, u_j\rangle)$, where $P_k$ is the degree-$k$ Gegenbauer polynomial normalized by $P_k(1) = 1$, follow from the addition theorem; the matrices $H_k$ are generated by the three-term recurrence of the $P_k$ on the $m \times m$ Gram matrix of the directions, which works directly with the $m \times m$ Gram matrix. One teacher fit and one pass of the recurrence give the whole curve $t \mapsto \mathcal{L}_{\mathrm{ST}}(t)$.

**The harmonic cutoff.** The sum (G.1) has to be cut at a finite degree $K$; we use $K = 128$. The omitted part is $\sum_{k>K}(1 - e^{-\sigma_k^2 t})^2 \|\Pi_k g\|_{\mathcal{D}}^2$, and since the filters decrease in $k$ it is at most $(1 - e^{-\sigma_{K+2}^2 t})^2$ times the teacher energy above degree $K$, which is $\|g\|_{\mathcal{D}}^2 - \sum_{k \le K} \|\Pi_k g\|_{\mathcal{D}}^2$ and is known from the fit. This bound is evaluated for every teacher and every time. It is small on the displayed time window for a structural reason: the student learns degree $k$ on the time scale $1/\sigma_k^2 \asymp d^{k+1}$, so at times of order $T_2 \asymp d^3 \log d$ the degrees above the first few are still at their zero initialization and contribute nothing to (G.1). Figure 8 (bottom row) quantifies this: cutting at $K = 16$ instead of $K = 128$ changes the loss by less than $10^{-6}$ relative on the whole window, and the bound on the tail above $K = 128$ is below $4 \times 10^{-4}$ relative to the loss, attaining its maximum near $T_2$ where the loss itself is smallest.

## G.2 AVERAGING OVER TEACHERS

Each marker in Figure 1 is the mean of $\mathcal{L}_{\mathrm{ST}}(t)$ over independent teachers: 32 at $d = 8$ and $d = 16$, and 8 at $d = 28$, where $m = 21952$. Figure 8 shows what is being averaged. The top row plots every teacher's curve, their mean, a pointwise 95% $t$ interval for the mean, and the deterministic equivalent. The teacher-to-teacher spread is invisible before the stop, where the student is learning the target, and appears after it, where the student is copying the teacher's random residual: this is

the variance term $\overline{V}$ of (3.3). At the stop $T_2$ the half-width of the interval is 9%, 3%, and 5% of the mean for $d = 8, 16, 28$, and the deterministic equivalent lies below the mean by 15%, 4%, and 1%, a finite-$d$ discrepancy that closes as $d$ grows. In the normalized panel of Figure 1 the mean teacher error of the same teachers is used as denominator. At $d = 28$ the mean student loss at $T_2$ divided by the squared mean teacher loss is 1.067, with paired-bootstrap 95% interval $[1.065, 1.069]$ over teachers, against the deterministic-equivalent value 1.066.

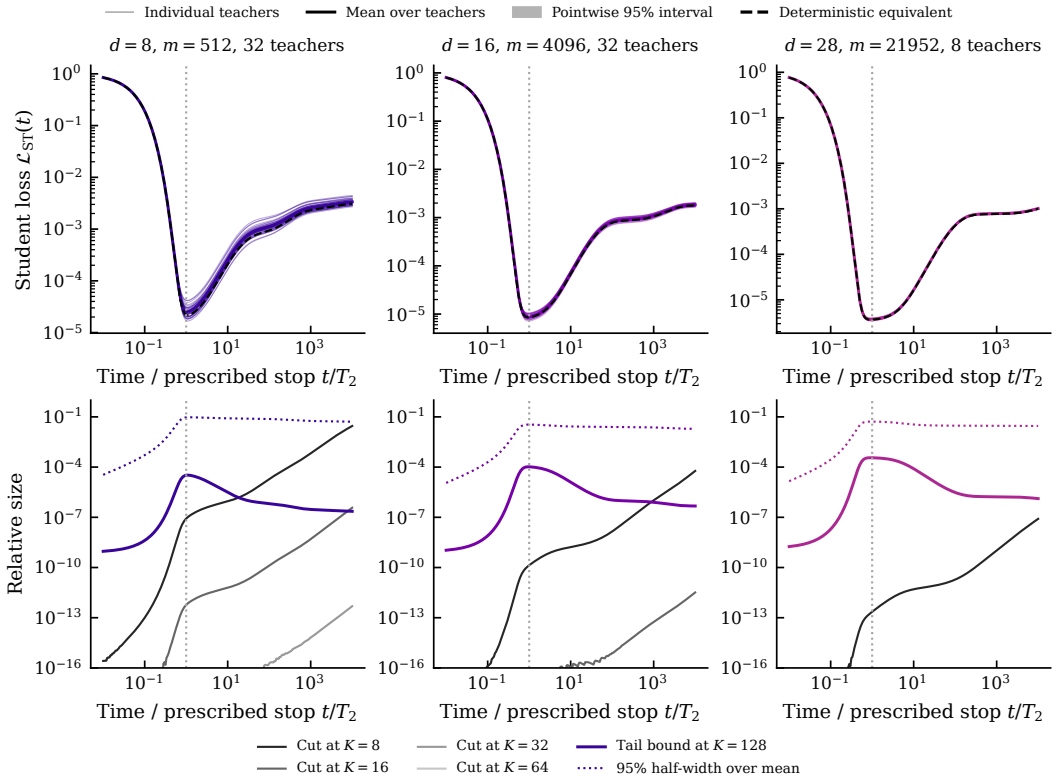



Figure 8: **Averaging over teachers and the harmonic cutoff.** Unit degree-two target, $m = d^3$, the three dimensions with population-flow samples in Figure 1. Top: individual teachers, their mean, a pointwise 95% interval for the mean, and the deterministic equivalent. Bottom, on the same time axis: the relative change of the loss when the harmonic sum (G.1) is cut at $K$ instead of $K = 128$ (worst teacher), the bound on the omitted tail above $K = 128$, and the relative half-width of the 95% interval. Cutting at $K \geq 16$ is indistinguishable from $K = 128$ on this window; the cutoff error is far below the statistical uncertainty everywhere.

### G.3 Deterministic-equivalent curves

The lines in the figures evaluate (3.1)–(3.3) with the exact $N_k$ and $\sigma_k^2$ of Appendix C.1, grouping every trace by harmonic degree. The scalar equation (3.1) for $\nu$ is solved by bisection. Multiplicities and eigenvalues are combined in the logarithmic domain, since $N_k$ overflows and $\sigma_k^2$ underflows long before their product does, and the filters $1 - e^{-x}$ are evaluated with a stable exponential-difference routine. The harmonic sums are cut at degree 2048; raising the cutoff from 1024 to 2048 changes every displayed curve by less than 0.04%, and the curves reproduce the exact endpoints, unit loss at $t = 0$ and $\overline{\mathcal{L}}_{\mathrm{TE}}$ at $t = \infty$. The mixed-target figures of Section 6 and Appendix E are such evaluations; the numerically attained minima marked there are minima over a grid in $\sigma_4^2 t$, and at the largest dimension shown they lie within 0.1% of the analytic lower bound (E.8), at a time within 0.2% of $\log 2$.

### G.4 THE DEGREE-ZERO MODE

All figures were generated without the degree-zero mode of the kernel (4.1). The deterministic-equivalent curves use the spectrum over positive degrees, and the population-flow samples use the centered activation $\sigma(r) = r_+ - \mu_d$, whose kernel has no constant component. The main text keeps the mode, so we record its effect. For the deterministic equivalent, adding the term $N_0 = 1$, $\sigma_0^2 = \mu_d^2$ to the spectrum changes $\nu$ and $\overline{\mathcal{L}}_{\mathrm{TE}}$ by a relative amount close to $1/m$, and the student curve of Figure 1 by at most twice that: $0.6\%$ at $d = 8$, $0.07\%$ at $d = 16$, $0.01\%$ at $d = 28$, and below $10^{-5}$ for $d \geq 64$. For the population flow, refitting the archived teachers of Figure 8 with the uncentered kernel changes the teacher error by at most $1.9\%$, $0.2\%$, and $0.03\%$ at $d = 8, 16, 28$ (on average $0.4\%$, $0.05\%$, and $0.01\%$), and the student curve at any time by at most $5.8\%$, $0.4\%$, and $0.06\%$. The constant component of the uncentered teacher itself carries less than $10^{-6}$ of its error at every dimension; the change comes through the fitted weights. Both effects are below the teacher-to-teacher variability shown in Figure 8.