Weak-to-Strong Generalization Even in Random Feature Networks, Provably

Marko Medvedev [*1] **Kaifeng Lyu** [*2] **Dingli Yu** [3] **Sanjeev Arora** [4] **Zhiyuan Li** [5] **Nathan Srebro** [5]

Abstract

Weak-to-Strong Generalization (Burns et al., 2024) is the phenomenon whereby a strong student, say GPT-4, learns a task from a weak teacher, say GPT-2, and ends up significantly outperforming the teacher. We show that this phenomenon does not require a complex and pre-trained learner like GPT-4, can arise even in simple non-pretrained models, simply due to the size advantage of the student. But, we also show that there are inherint limits to the extent of such weak to strong generalization. We consider students and teachers that are random feature models, described by two-layer networks with a random and fixed bottom layer and trained top layer. A 'weak' teacher, with a small number of units (i.e. random features), is trained on the population, and a 'strong' student, with a much larger number of units (i.e. random features), is trained only on labels generated by the weak teacher. We demonstrate, prove, and understand how the student can outperform the teacher, even though trained only on data labeled by the teacher. We also explain how such weak-to-strong generalization is enabled by early stopping. We then show the quantitative limits of weak-to-strong generalization in this model, and in fact in a much broader class of models, for arbitrary teacher and student feature spaces and a broad class of learning rules, including when the student features are pre-trained or otherwise more informative. In particular, we show that in such models the student's error can only approach zero if the teacher's error approaches zero, and a strong student cannot "boost" a slightly-better-then-chance teacher to obtain a small error.

[*]Equal contribution [1]Department of Mathematics, University of Chicago, Chicago, IL, US [2]Simons Institute, University of California, Berkeley, San Francisco, CA, US [3]Microsoft Research, Seattle, WA, US [4]Princeton University, Princeton, NJ, US [5]TTIC, Chicago, IL, US. Correspondence to: Marko Medvedev <medvedev@uchicago.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

1. Introduction

(Samuel, 1959), in the first ever paper introducing the term "machine learning", emphasized how his learned system outperformed its human teacher. More broadly, the idea of a good student outperforming their teacher is much older, as epitomized by the quote "*Poor is the pupil who does not surpass his master*", widely attributed to Leonardo de Vinci.[*] Recently, Burns et al. (2024) discussed the concept in the context of "superalignment" (weak humans controlling strong machines), and suggested studying (strong) student machines outperforming their (weaker) machine teachers as a surrogate model. Burns et al. demonstrated the phenomenon by fine-tuning a "weak" teacher model, namely a pre-trained GPT-2, on task-specific data with true labels, and then using this "weak" teacher model to annotate data for the same task, and fine-tuning a "strong" student model, namely a pre-trained GPT-4, only using labels generated by the "weak" teacher (see Figure 1). In these experiments, the "strong" students indeed outperformed their "weak" teachers, and Burns et al. asked for an understanding of this *weak-to-strong generalization* phenomenon.

There are several mechanisms by which a student can outperform their teacher. A student can do a more thorough test-time search, e.g. looking ahead more moves or thinking of more ways to solve a problem. Taking a step further, on problems where such a search can be helpful, a student can also learn to internalize the search, learning a better value function or next-action (or token) predictor through self-play or training-time search (as in Samuel's checker learner, and modern RL-based training methods for LLMs). A student might also be able to correct a teacher by having better inductive bias, perhaps obtained through pre-training on similar or related tasks. For example, a Chinese-speaking student can outperform their non-Chinese-speaking teacher when being taught to pronounce names of Chinese colleagues.

But can a student outperform their teacher simply by being "bigger", having more units and a larger representation, without more specialized or specific inductive bias, nor more or better pre-training? Does this require sophisticated models with emergent capabilities, such as GPT-4 scale deep transformers, or can it also arise in much simpler models?

[*]The quote does not appear in de Vinci's surviving writing and its source is unclear to us

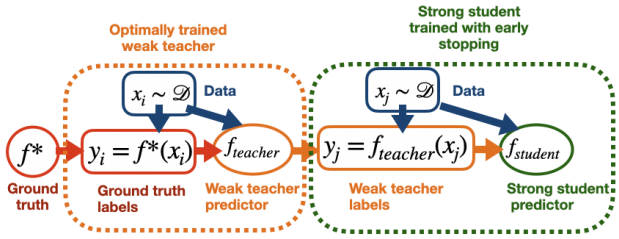


Figure 1. Illustration of our setup for weak-to-strong generalization. The teacher model of smaller size is trained on ground truth labels, while the student model of larger size is trained with early stopping on labels produced by the teacher model.

If the model sizes are further scaled up, will weak-to-strong generalization be enhanced or diminished?

In the first part of this paper, we analyze and prove how weak-to-strong generalization (i.e. a student outperforming their student) can happen even in models as simple as "random feature models", i.e., two-layer networks with a random bottom layer and only the top layer trained. More specifically, we study the setup where the teacher and student models are two-layer networks with M_{TE} and M_{ST} (hidden, random) units, respectively. The student model is significantly stronger than the teacher, in the sense that it has many more units than the teacher, $M_{\mathrm{ST}} \gg M_{\mathrm{TE}}$. In both models, we sample the bottom layer weights from the same spherical distribution, and then train only the top layer weights. As in Burns et al. (2024) and Figure 1, the teacher model is trained with true labels, and the student model is trained only with labels generated by the teacher. We further assume that the teacher model is trained so well that it attains the minimum possible population loss. Thus, the only remaining source of error is the approximation error due to the finite number of units.

We consider two different random feature models, corresponding to two-layer networks with different activation functions and input distributions: (a) ReLU activations with an isotropic input distribution; and (b) linear activations with a simple anisotropic input distribution. Denoting the teacher and student predictive mean squared errors by $\mathcal{L}_{\mathrm{TE}}$ and $\mathcal{L}_{\mathrm{ST}}$ respectively (normalized such that the null loss is 1), in both cases we show that **the student can infinitely outperform its teacher, in the sense that the ratio $\mathcal{L}_{\mathrm{ST}}/\mathcal{L}_{\mathrm{TE}}$ can be arbitrarily small.** Thus, **the *Performance Gap Recovered (PGR)*** introduced by Burns et al. (2024, see definition in Section 2 Equation (7)) **can be arbitrarily close to** 1. Furthermore, we show an even stronger gap, in the sense that the student learns polynomially better: **for ReLU networks we show that we can have $\mathcal{L}_{\mathrm{ST}} = \tilde{O}\left(\mathcal{L}_{\mathrm{TE}}^{1.49}\right)$** (Theorem 3.1, relying on the Gassian Universality Ansatz), and for **linear networks with anisotropic inputs we show that we can have $\mathcal{L}_{\mathrm{ST}} = \tilde{O}\left(\mathcal{L}_{\mathrm{TE}}^{2}\right)$** (Theorem 3.2).

As in Burns et al. (2024), we use *early stopping* when training the student—it is easy to see how in our setting early stopping is essential: Since the student is more expressive than the teacher, without early stopping they would just replicate the teacher's mistakes. In Section 5 we explain how such early stopping allows the weak-to-strong generalization described above, shedding light on its empirically observed role.

Complementing these results, we also study the limits of weak-to-strong generalization. In Section 4 we show that the quadratic gap described above is the largest possible, and in any random feature model we would always have $\mathcal{L}_{\mathrm{ST}} = \Omega(\mathcal{L}_{\mathrm{TE}}^2)$. In fact, this limit extends well beyond random feature models, to models with arbitrary teacher and student feature spaces (even if the student features are highly specialized or pre-trained), with early stopped gradient descent (starting from zero initialization) or with other convex regularizers. In particular, this limit implies that in such settings, we can have vanishing student error $\mathcal{L}_{\mathrm{ST}} \to 0$ only when the teacher error also vanishes $\mathcal{L}_{\mathrm{TE}} \to 0$, and that if the teacher error is only slightly better than null, the student error will also be close to null—we cannot "boost" a teacher with a tiny edge to a near-perfect student.

Asymptotic Notation. We use Θ_c, $\tilde{O}_c$, and Ω_c, where c is one or more variables, to indicate a multiplicative factor that depends on c. E.g., for any two quantities exp_1, exp_2, that could depend on various variables including d and k, writing $exp_1 = \tilde{O}_{d,k}(exp_2)$ means that there exists some function $f(d,k)$ and exponent $r \geq 0$ s.t. it always holds that $exp_1 \leq f(d,k) \cdot exp_2 \cdot \log^r(exp_2)$.

2. Two Layer Networks, Random Features, and the Weak-to-Strong Setup

We consdier learning a predictor $f : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}$, aiming for small population loss $\mathcal{L}(f) := \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[(f(\boldsymbol{x}) - f^*(\boldsymbol{x}))^2]$, where $\mathcal{D}$ is some input distribution and $f^* : \mathcal{X} \to \mathcal{Y}$ our target function. Learning is done using two-layer networks with m units of the form

$$f_{\boldsymbol{w},\boldsymbol{u}}(\boldsymbol{x}) = \sum_{i=1}^{m} w_i \sigma(\langle \boldsymbol{u}_i, \boldsymbol{x} \rangle), \tag{1}$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is some fixed activation function (e.g. ReLU or linear), the **bottom layer weights $\boldsymbol{u}_i \in \mathbb{R}^d$ are randomly drawn**

$$\boldsymbol{u}_i \overset{\text{i.i.d.}}{\sim} \mathcal{U} \tag{2}$$

from some fixed distribution $\mathcal{U}$ over $\mathbb{R}^d$ (e.g. spherical Gaussian or similar) and then **fixed during training**, and **only the top layer weights $w_i \in \mathbb{R}$ are trained**. Such models,

which amount to learning a linear predictor over the *random features* defined by the outputs of the hidden units, have been extensively studied as finite approximations of kernel methods (Arriaga & Vempala, 2006; Rahimi & Recht, 2007), as models and limits of deep learning in certain regimes (Mei et al., 2022) and recently also for deriving scaling laws and optimal tradeoffs (Lin et al., 2024; Paquette et al., 2024).

Weak-to-Strong Setup. We train two networks, a 'weak' teacher network f_{teacher} with M_{TE} units, and a 'strong' student network f_{student} with $M_{\text{ST}} \gg M_{\text{TE}}$ units. First, we train the teacher network $f_{\text{teacher}} := f_{\boldsymbol{w}_{\text{t}},\boldsymbol{u}_{\text{t}}}$ on the true target labels $(x, f^*(x))$ using very large amount of data. We will model this by training the teacher on population. That is, to train the teacher we draw the bottom layer weights $\{\boldsymbol{u}_{\text{t},i}\}_{i=1}^{M_{\text{TE}}}$ at random as in (2), and for most results in this paper we assume that the top layer weights achieves the population loss minimization (Condition 2.1):

Condition 2.1. The teacher model attains the minimum loss among all possible models with the given bottom layers:

$$\boldsymbol{w}_{\text{t}} = \operatorname*{arg\,min}_{\boldsymbol{w}\in\mathbb{R}^{M_{\text{TE}}}} \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\left(f_{\boldsymbol{w},\boldsymbol{u}_{\text{t}}}(\boldsymbol{x}) - f^*(\boldsymbol{x})\right)^2\right]. \quad (3)$$

We denote the teacher's population loss (which is the minimal possible loss using its M_{TE} random units) by $\mathcal{L}_{\text{TE}} := \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\left(f_{\text{teacher}}(\boldsymbol{x}) - f^*(\boldsymbol{x})\right)^2\right]$. We then train the student on inputs $x \sim \mathcal{D}$ but only using labels $f_{\text{teacher}}(x)$ generated by the teacher. To train the student, we draw the bottom layer at random according to (2), initialize the top layer weights to zero, $\forall_i \boldsymbol{w}_{\text{s},i}(0) = 0$, and train these weights using SGD on i.i.d. samples $\boldsymbol{x}(n) \sim \mathcal{D}$ labeled with by the teacher:

$$\begin{aligned} w_{\text{s},i}(n+1) &= w_{\text{s},i}(n) \\ &- \eta \cdot \frac{\partial}{\partial w_{\text{s},i}} \left(f_{\boldsymbol{w}_{\text{s}}(n),\boldsymbol{u}_{\text{s}}}(\boldsymbol{x}(n)) - f_{\text{teacher}}(\boldsymbol{x}(n))\right)^2. \end{aligned} \quad (4)$$

Gradient Flow Limit. We assume that the student has unrestricted access to the examples labeled by the teacher, and so can take arbitrarily small learning rate η and an appropriately large number of steps N. Accordingly, we study the behavior with infinitesimally small learning rate η where the student's dynamics (4) converge to the gradient flow trajectory:

$$\frac{\mathrm{d}w_{\text{s},i}(t)}{\mathrm{d}t} = -\frac{1}{M_{\text{ST}}}\frac{\partial}{\partial w_{\text{s},i}} \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\left(f_{\boldsymbol{w}_{\text{s}}(t),\boldsymbol{u}_{\text{s}}}(\boldsymbol{x}) - f_{\text{teacher}}(\boldsymbol{x})\right)^2\right] \quad (5)$$

where we replaced the step number n with training time $t = n\eta M_{\text{ST}}$. We add an additional factor of M_{ST} here to ensure the gradient flow has a proper limit when $M_{\text{ST}} \to \infty$, which will be useful in the analysis of the rest of the paper. Gradient flow has been show to approximate the trajectory of gradient descent well, both for 2-layer networks (Ali et al., 2019) and for deep networsk (Elkabetz & Cohen, 2021). In our analysis, we consider a student trained with the gradient flow dynamics (5) with stopping time T that needs to be specified and denote the resulting student predictor $f_{\text{student}} = f_{\boldsymbol{w}_{\text{s}}(T),\boldsymbol{u}_{\text{s}}}$.

Infinitely Large Student. Furthermore, for simplicity we consider an infinitely large student $M_{\text{ST}} \to \infty$ (see formal mathematical definitions in Section 6). All of our infinite-width results can be extended to the finite-width case for sufficiently large width.

Generalization Gap. We evaluate the student on the ground truth labels (which they do not have access to!), that is

$$\mathcal{L}_{\text{ST}} := \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[f_{\text{student}}(\boldsymbol{x}) - f^*(\boldsymbol{x}))^2]. \quad (6)$$

We ask how the student error $\mathcal{L}_{\text{ST}}$ compares with the teacher error $\mathcal{L}_{\text{TE}}$ and, in particular, whether and how it can be smaller even though it only sees examples labeled by the teacher.

To quantify the extent to which the student can best its teacher, Burns et al. (2024) suggested the *Performance Gap Recovered* (PGR), which is defined as[†]:

$$\text{PGR} = \frac{\mathcal{L}_{\text{TE}} - \mathcal{L}_{\text{ST}}}{\mathcal{L}_{\text{TE}} - \mathcal{L}_{\text{ST}}^{\text{ceil}}} \geq \frac{\mathcal{L}_{\text{TE}} - \mathcal{L}_{\text{ST}}}{\mathcal{L}_{\text{TE}}} \quad (7)$$

where $\mathcal{L}_{\text{ST}}^{\text{ceil}} \geq 0$ is the "ceiling performance"(Burns et al., 2024) of the student (the loss the student model could potentially have attained with direct access to the real labels). We do not carefully define or analyze this ceiling performance, since we provide lower bounds on the right-hand side in (7), and hence on the PGR regardless of the ceiling.

Specific Activation and Data Models. We study two different variants of the random feature model (aka 2-layer network with random and fixed bottom layer) introduced above, which differ in the choice of activation function and input distribution:

Model 2.2 (2-layer ReLU Network)**.** The activation functions $\sigma(z) = \max(z, 0)$ is a standard ReLU function, the bottom layer weights $\boldsymbol{u}$ are uniform over the sphere, i.e. $\mathcal{U} = \text{Unif}(\mathbb{S}^{d-1})$ in (2), and the inputs $\boldsymbol{x}$ are also uniformly distributed on the sphere, i.e. $\mathcal{D} = \text{Unif}(\mathbb{S}^{d-1})$. In this model, we will consider target function $f^*(x)$ which are sums of linear functions and even polynomials.

Model 2.3 (Linear Network)**.** The activation functions $\sigma(z) = z$ is linear (i.e. this is a linear neural net), and the bottom layer weights $\boldsymbol{u}$ are standard Gaussian, i.e. $\mathcal{U} = N(0, \boldsymbol{I}_d)$ in (2). This time the inputs $\boldsymbol{x}$ are non-isotropic

[†]This definition in terms of loss is equal to Burns et al.'s definition in terms of accuracy.

and are drawn from a Gaussian $\mathcal{D} = N(0, \Psi)$ with diagonal covariance $\Psi = \text{diag}(\psi_1, \psi_2, \ldots, \psi_d)$ with ψ_i decreasing. Here, we will consider linear targets supported on the first top few coordinates.

Model 2.3 is fairly general and was studied by e.g. Lin et al. (2024) as a model for studying scaling laws. It can also be thought of as capturing a model where each unit outputs $\langle \boldsymbol{u}, \phi(\boldsymbol{x}) \rangle$, where $\phi(\cdot)$ captures what happens in additional lower layers, which creates a non-isotropic representation at the top layer.

3. Quantifying Weak-to-Strong Generalization

We are now ready to establish weak-to-strong generalization, starting with the ReLU 2.2:

Theorem 3.1 (Weak-to-Strong generalization with 2-layer ReLU Network)**.** *Consider the ReLU Model 2.2 and the weak-to-strong setup of Section 2, for any dimension d and teacher size M_{TE}. Consider a target f^* that is an even polynomial[‡] of degree at most k normalized s.t. $\mathbb{E}[f^{*2}] = 1$, with any even degree k. If $M_{\text{TE}} \geq \Theta(d^{2k})$ and the student is trained with some stopping time $T = \Theta_{d,k}(\log M_{\text{TE}})$, then we have that with probability at least $1 - \frac{3}{M_{\text{TE}}}$,*

$$\mathcal{L}_{\text{ST}} \leq O_{d,k}\left(\frac{\log^2 M_{\text{TE}} + \log^2 \mathcal{L}_{\text{TE}}}{\sqrt{M_{\text{TE}}}} \mathcal{L}_{\text{TE}}\right),$$

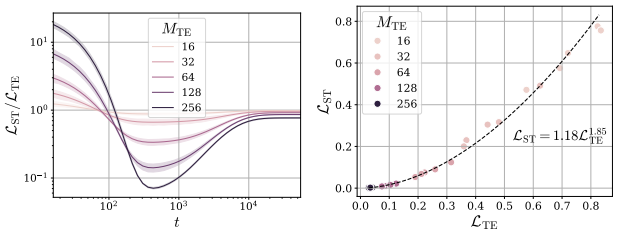
and so $\mathcal{L}_{\text{ST}}/\mathcal{L}_{\text{TE}} \to 0$ and $\text{PGR} \to 1$ as M_{TE} increases.

Furthermore, under the Gaussian Universality Ansatz[§], for[¶] $d > 200$, $\mathcal{L}_{\text{TE}} = \Omega_{d,k}(\frac{1}{M_{\text{TE}}^{1.02}})$, and so as $M_{\text{TE}} \to \infty$ we have that

$$\mathcal{L}_{\text{ST}} = \tilde{O}_{d,k}\left(\mathcal{L}_{\text{TE}}^{1.49}\right).$$

For a more detailed statement spelling out the dependnce on the degree k and stopping time T see Corollary E.3 and for the proof see Appendix F.

Theorem 3.1 already establishes that in Model 2.2, weak-to-strong generalization provably happens for a large set of targets, and for large teachers, we can get a performance gap recovered (PGR) arbitrarily close to 1. We see also that the PGR is not a fine enough measure of weak-to-strong generalization, since the separation is obtained when the teacher error is small, and the PGR does not tell us how small the student error $\mathcal{L}_{\text{ST}}$ can be as a function of the teacher error $\mathcal{L}_{\text{TE}}$: having $PGR \to 1$ only says that the student error $\mathcal{L}_{\text{ST}}$ goes to zero superlinearly faster than the teacher error $\mathcal{L}_{\text{TE}}$, but not how much faster. Indeed, we can see that even in the ReLU Model 2.2, the student error is *polynomially* smaller, by an *exponent* of at least 1.49. This last result relies on the widely verified Gaussian Universality Ansatz, and is also verified in our experiments: in Figure 2(a), we show that when the target f^* is linear, with proper early stopping time, the loss ratio $\mathcal{L}_{\text{ST}}/\mathcal{L}_{\text{TE}}$ decreases when M_{TE} grows. Furthermore, in Figure 2(b), we observe the student loss $\mathcal{L}_{\text{ST}}$ is polynomially smaller than the teacher loss $\mathcal{L}_{\text{TE}}$ with an estimated exponent even above our bound. Indeed, we expect our bound on the behavior of ReLU Random Feaure Networks is loose, and it might be possible to establish an even stronger separation.



(a) $\mathcal{L}_{\text{ST}}/\mathcal{L}_{\text{TE}}$ v.s. t (b) $\mathcal{L}_{\text{ST}}$ v.s. $\mathcal{L}_{\text{TE}}$

Figure 2. Weak-to-strong generalization happens in ReLU random feature networks (Model 2.2) with input dimension $d = 32$, student size $M_{\text{ST}} = 16384$, and teacher size $M_{\text{TE}} \in \{16, \ldots, 256\}$. We consider a linear target function $f^*(\boldsymbol{x}) = \langle \beta, \boldsymbol{x} \rangle$ for unit norm some β. Figure 2(a) plots the ratio between student loss $\mathcal{L}_{\text{ST}}$ and teacher loss $\mathcal{L}_{\text{TE}}$, with varying teacher size M_{TE} and gradient flow training time t. With appropriate stopping time, we see a significant weak-to-strong generalization gain. This gain diminishes with overtraining and running gradient flow to convergence, the student mimics the teacher, has the same error, and does not excite weak-to-strong generalization. In Figure 2(b), we fit the minimal student loss $\mathcal{L}_{\text{ST}}$ (at the optimal optimal stopping time for each teacher size) as a power law function of the student loss $\mathcal{L}_{\text{TE}}$, confirming Theorem 3.1. See Appendix J for simulation details.

In order to demonstrate a stronger separation theoretically, and also avoid the Gaussian Universality Ansatz, we turn to the Linear Network Model 2.3:

Theorem 3.2 (Weak-to-Strong Generalization with Linear Network)**.** *Consider the linear network Model 2.3, with $\Psi_d = (1, \ldots, 1, \psi_*, \ldots, \psi_*)$, where 1 is repeated k times and $\psi_* = (d-k)^{-2/3}$ is repeated $d-k$ times, and a linear target $f^*(\boldsymbol{x}) = \langle \beta, \boldsymbol{x} \rangle$ supported by the first k coordinates,*

[‡]Or more generally, a sum of a linear function and an even polynomial.

[§]The Gaussian Universality Ansatz states that when sampling $\boldsymbol{x}$ the Gaussian universality holds for the eigenfunctions in the sense that the expected risk remains unchanged if we replace them with Gaussian with appropriate parameters (Simon et al., 2023). In our setting, this can be differently stated as the expected risk for Model 2.2 will remain unchanged if we replace it with infinitely dimensional Model 2.3 with the appropriate covariance. Formally, the lower bound on $\mathcal{L}_{\text{TE}}$ and an upper bound on T are conditional on the eigenframework. This assumption has been empirically verified in a number of settings (Simon et al., 2023; Canatar et al., 2021; Wei et al., 2022; Misiakiewicz & Saeed, 2024).

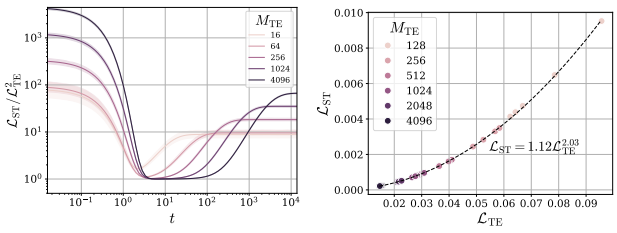
[¶]This choice of dimension is arbitrary. We can get close to $\mathcal{L}_{\text{ST}} = \tilde{O}_{d,k}(\mathcal{L}_{\text{TE}}^{1.5})$ with a larger d.

i.e. $\beta_i = 0$ *for* $i > k$ *that is normalized* $\mathbb{E}[f^{*2}] = 1$. *In the weak-to-strong setting of Section 2, for any* k *and* $d > k + Ck^3$*(where* C *is an absolute constant), and a teacher of size* $M_{\mathrm{TE}} = (d-k)^{2/3}$*, if the student is trained until time* $T = \Theta(\log M_{\mathrm{TE}})$*, we have that with probability at least* 0.99 *for all* M_{TE}*,*

$$\mathcal{L}_{\mathrm{ST}} \leq \tilde{O}(k\mathcal{L}_{\mathrm{TE}}^2).$$

In particular, for any fixed k*, as* $d, M_{\mathrm{TE}} \to \infty$*, we have that* $\mathrm{PGR} \to 1$.

A more general statement for Model 2.3 that holds for any covariance matrix Ψ and target is given in Theorem E.4. For the proof, see Appendix D. We indeed observe the behavior described by the Theorem in simulation experiments, as presented in Figure 3, where we can see the student loss $\mathcal{L}_{\mathrm{ST}}$ is indeed quadratically smaller than the teacher loss $\mathcal{L}_{\mathrm{TE}}$, already for moderately sized networks.



(a) $\mathcal{L}_{\mathrm{ST}}/\mathcal{L}_{\mathrm{TE}}^2$ v.s. t (b) $\mathcal{L}_{\mathrm{ST}}$ v.s. $\mathcal{L}_{\mathrm{TE}}$

Figure 3. Weak-to-Strong generalization happens in random linear feature networks (Model 2.3). Here we used an input distribution as in Theorem 3.2, with $k = 1$ and a target function $f^* = \langle e_1, \boldsymbol{x} \rangle$ where e_1 is the first standard basis vector. Figure 3(a) plots the ratio between the student loss $\mathcal{L}_{\mathrm{ST}}$ and squared teacher loss $\mathcal{L}_{\mathrm{TE}}^2$, with varying teacher size M_{TE}, and where the dimensionality $d = M_{\mathrm{TE}}^{3/2}$ as set as in the scaling of Theorem 3.2, as a function of the gradient flow time t. With proper early stopping time $\mathcal{L}_{\mathrm{ST}}/\mathcal{L}_{\mathrm{TE}}^2$ converges to approximately 1 as M_{TE} grows, confirming that for large M_{TE} we have $\mathcal{L}_{\mathrm{ST}} \propto \mathcal{L}_{\mathrm{TE}}^2$ as in Theorem 3.2. This is also confirmed in Figure 3(b), where we fit the student loss $\mathcal{L}_{\mathrm{ST}}$ as a power law function of teacher loss $\mathcal{L}_{\mathrm{TE}}$, and recover an excellent fit with an exponent very close to 2. We again see overtraining diminishes weak-to-strong generalization. See Appendix J for simulation details.

4. The Limit of Weak to Strong Generalization

In the previous Section we saw how a student can significantly ourperform its teacher, even in random feature models. In particular, we showed that a quadratic reductoin in error is possible. One might ask if in some cases an even greater improvement is possible. Perhaps a cubic or higher order polynomial improvement? In this Section, we show that the quartatic improvement of Theorem 3.2 is the largest possible in *any* random feature model. No matter the the feature distribution and target, and how related they are, if the teacher has error $\mathcal{L}_{\mathrm{TE}}$, a student, even with many more features, cannot obtain error better that $\mathcal{L}_{\mathrm{ST}} = \Omega(\mathcal{L}_{\mathrm{TE}})$.

In fact, we show that this limitation holds much more broadly, not only in *random* feature models, but for any teacher optimal on some set of features and for any student trained with early stopped gradient flow on any other set of features. The Random Feature models are a special case where the teacher and student models are small and large random subsets from the same feature distribution, but the limitation holds even if the student features are pre-trained or otherwise cleverly selected.

Furthermore, the limitation applies to a broad class of training methods, which go beyond just early stopped gradient flow. All we require is that the learned predictor is *shrinkage optimal* in the following sense:

Definition 4.1 (Shrinking Optimality)**.** Given two functions $\hat{f}, f_{\text{train}} : \mathcal{X} \to \mathbb{R}$, we say $\hat{f}$ is *shrinnking-optimal* with respect to f_{train} if for any $0 \leq \alpha \leq 1$, $\mathbb{E}_{\boldsymbol{x}}((\hat{f}(\boldsymbol{x}) - f_{\text{train}}(\boldsymbol{x}))^2) \leq \mathbb{E}_{\boldsymbol{x}}((\alpha\hat{f}(\boldsymbol{x}) - f_{\text{train}}(\boldsymbol{x}))^2)$.

That is, all we require is that the training method returns a predictor $\hat{f}$ which cannot improve its training objective, simply by scaling down, or shrinking toward the origin. Shrink optimality is satisfied by loss minimization on a linear subspace (i.e. our learning rule for the teacher), or more generally any convex class that includes the origin. It is also satisfied by early stopped gradient flow over any feature space or with any kernel, and with any stopping time (including $T = \infty$, which corresponds to just loss minimization):

Lemma 4.2 (Shrinkage Optimality of Gradient Flow Solutions)**.** *For any feature map* $\phi(x)$ *and any target* $f_{train}(x)$ *consider gradient flow* $\dot{\boldsymbol{w}} = -\nabla_{\boldsymbol{w}}\mathbb{E}_{\boldsymbol{x}}\left[(\langle \boldsymbol{w}, \phi(\boldsymbol{x}) - f_{train}(\boldsymbol{x}))^2\right]$ *initialized at* $\boldsymbol{w}(0) = 0$*. Then at any time* $0 \leq T \leq \infty$*, the predictor* $f_T(\boldsymbol{x}) = \langle \boldsymbol{w}(T), \phi(\boldsymbol{x}) \rangle$ *is shrink optimal w.r.t.* f_{train}.

Lemma 4.2 shows that in the Random Feature setup considered in Section 2, the teacher and the student both satisfy the shrinkage optimality Definition 4.1, with respect to target and the teacher predictor respectively.

With these definitions in place we are ready to state our main lower bonud on weak-to-strong generalization:

Theorem 4.3 (Genereal Limitation of Weak-to-Strong Generalization)**.** *For any target* f^* *normalized s.t.* $\mathbb{E}_{\boldsymbol{x}}\left[f^*(\boldsymbol{x})^2\right] \leq 1$*, for any teacher* $f_{teacher}$ *that is shrinking-optimal to* f^* *and any student* $f_{student}$ *that is shrinking-optimal to* $f_{teacher}$*, it holds that*

$$\mathcal{L}_{\mathrm{ST}} \geq \frac{\left(\sqrt{1+3\mathcal{L}_{\mathrm{TE}}} - \sqrt{1-\mathcal{L}_{\mathrm{TE}}}\right)^2}{4} \geq \frac{3}{4}\mathcal{L}_{\mathrm{TE}}^2.$$

Corollary 4.4 (Limit of Random Feature Weak to Strong Generalization)**.** *For any Random Feature Network Weak-to-Strong training as in Section 2, any normalized target* $\mathbb{E}[f^{*2}] = 1$*, and any student stopping time* T*, we have that*

$$\mathcal{L}_{\mathrm{ST}} \geq \frac{\left(\sqrt{1+3\mathcal{L}_{\mathrm{TE}}} - \sqrt{1-\mathcal{L}_{\mathrm{TE}}}\right)^2}{4} \geq \frac{3}{4}\mathcal{L}_{\mathrm{TE}}^2.$$

For the proof of Lemma 4.2, Theorem 4.3, and Corollary 4.4, see Appendix H.

Theorem 4.3 and Corollary 4.4 establish significant limits on the extent of weak-to-strong generalization. We see that although the PGR can be arbitrarily close to one, and the student error can be quadratically smaller than the teacher error, it cannot be any smaller than that. We cannot have arbitrarily high teacher error and arbitrarily low student error. In particular, we cannot have a situation in which the teacher error is very close to the null risk $\mathbb{E}[f^{*2}]$, i.e. the teacher is barely learning, while the student error is low (if $\mathcal{L}_{\mathrm{TE}}$ is closer to one, then $\mathcal{L}_{\mathrm{ST}}$ must also be close to one). That is, weak-to-strong learning *cannot* significantly "boost" a teacher that is only slightly better than chance. And, in order for the student error to go to zero, *the teacher error must also go to zero*.

These limitations hold not only in random feature models, but based on Theorem 4.3, also much more broadly. First of all, they hold with *any* student and teacher feature space—these need not be random. We saw in Section 3 that a quadratic weak-to-strong improvement is possible even with random features, with the same student and teacher feature distributions. But what Theorem 4.3 tells us is that even if the student features are much more specialized than the teacher features, even if they are more aligned with the taregt, or even if they are pre-trained, we cannot get any larger improvement. Furthermore, this holds for a very general class of learning rules. Beyond early stopped gradient flow, using any type of convex regularizer (minimized at the origin) or constraining to any convex hypothesis class (which includes the origion) also ensures shrink-optimality. Any such learning rule is still subject to the limitation of Theorem 4.3.

The main caveat here is that the limitation applies only with gradient flow *initialized at the origin*, with convex hypothesis classes *containing the origin* or with convex regularizers *minimized at the origin*. Concretly, consider a student that uses a pre-trained feature space, but trains the top layer weights from scratch (starting from the origin). This student is subject to the limitation of Theorem 4.3. In contrast, consider a student that fine tunes a pre-trained top layer, starting from the pre-trained initialization (or alternatively, explicitly regularizes to be close to the pre-trained model). Such a student is *not* shrink optimal and is *not* subject to the limitation of Theorem 4.3. Indeed, if the pre-trained model is already very good (at an extreme—if the pre-trained model already happens to be better on the task than the teacher), the student can be arbitrarily better than its teacher (similar to the Chinese-speaking student outperforming its non-Chinese-speaking teacher in Chinese). In the setting of (Charikar et al., 2024), the difference is whether the teacher and student classes contain the origin (in which case shrink optimality holds and Theorem 4.3 applies), or do not contain the origin. One can think of this distinction in terms of whether the inductive bias specified by the feature space, learning rule, initialization, and hyopthesis class is *generic*, i.e. extends from the origin, prefering certain directions but not a particular point in predictor space, or *specialized*, i.e. prefering not only particular directions of features, but a very specific bias.

Note that the quadratic lower bound in terms of teacher error for student error holds even in a bootstrapping setting with multiple students, see Appendix G for details.

5. How Does Weak-to-Strong Generalization Happen?

Linear Network. To understand how the strong student can correct their weak teacher and obtain a lower error, let us first consider linear networks as in Model 2.3. In this case, both weak and strong models, $f_{\text{teacher}}(\boldsymbol{x})$ and $f_{\text{student}}(\boldsymbol{x})$, are linear functions of $\boldsymbol{x}$, and thus can be written as $f_{\boldsymbol{w},\boldsymbol{u}}(\boldsymbol{x}) = \langle \beta, \boldsymbol{x} \rangle$ where $\beta = \boldsymbol{U}^\top \boldsymbol{w}$. First, let us understand the student's learning rule. Let $f_{\text{student}}(\boldsymbol{x}) = f_{\boldsymbol{w},\boldsymbol{u}_s}(\boldsymbol{x})$ be the student predictor. Learning a predictor f with an early-stopped gradient descent shrinks $f - f_{\boldsymbol{w},\boldsymbol{u}_{\mathrm{s}}}$ in the direction of the student's random features $\sigma(\langle \boldsymbol{u}_{\mathrm{s},i}, \boldsymbol{x} \rangle) = (\boldsymbol{U}_{\mathrm{ST}}\boldsymbol{x})_i$ proportional to the variance of each of the features ψ_i. For a very wide student, i.e. with $M_{\mathrm{ST}} \to \infty$, the student's features will effectively be the coordinate directions x_i, so the shrinkage of $f - f_{\boldsymbol{w},\boldsymbol{u}_{\mathrm{s}}}$ will be proportional to the variance ψ_i of x_i in directions x_i. This is because as $M_{\mathrm{ST}} \to \infty$, we have that $\boldsymbol{U}_{\mathrm{ST}}^\top \boldsymbol{U}_{\mathrm{ST}} \to M_{\mathrm{ST}} I$. This makes learning directions x_i with larger variance ψ_i (i.e. for smaller i) significantly faster than the ones with smaller variance ψ_i (i.e. for larger i). If there is a sudden drop in variance ψ_i at index K, i.e. $\psi_K \gg \psi_{K+1}$, the time to learn any direction x_i for $i \leq K$ will be much smaller than the time to learn any of the directions x_i for $i > K$. So, in this case, we can choose a stopping time T to be such that all large variance directions x_i are learned, that is $(f - f_{\boldsymbol{w},\boldsymbol{u}_s})_i$ is very small or in other words $(f_{\boldsymbol{w},\boldsymbol{u}_s})_i \approx f_i$, but all the small variance directions x_i of $f_{\boldsymbol{w},\boldsymbol{u}_s}$ are close to initialization, i.e. zero. In this way, the student predictor $f_{\text{student}}(\boldsymbol{x}) = f_{\boldsymbol{w},\boldsymbol{u}_s}$ effectively learned the high variance (i.e. small index i) directions x_i of the teacher predictor f_{teacher} and zeroed out the small variance (i.e. large index i) directions x_i of the teacher predictor f_{teacher}. All of this can be explicitly seen in the closed form solution of

gradient flow in function space, Equation (5). Let us now turn to the teacher. Although the signal f^* (i.e. the target of GF) is entirely in the first few high variance coordinates of $\boldsymbol{x}$, say $\{x_i\}_{i\leq K}$, if we have a small number of teacher nodes M_{TE}, they will not be directly represented in the hidden layer, i.e. in $\boldsymbol{U}_{\mathrm{TE}}\boldsymbol{x}$. Instead, the teacher can only learn $\boldsymbol{\beta}_{\mathrm{TE}} \in \mathrm{span}(\boldsymbol{U}_{\mathrm{TE}})$, and so learns $\boldsymbol{\beta}_{\mathrm{TE}}$ to be the projection of f^* to $\mathrm{span}(\boldsymbol{U}_{\mathrm{TE}})$. This means that $f_{\text{teacher}}(\boldsymbol{x})$ has some energy (non-zero coefficients) in all the coordinates. The student shrinkage will reduce the noise (i.e. teacher signal uncorrelated to the ground truth) in low variance coordinates by zeroing out the coefficients along those directions. Therefore, the weak-to-strong improvement comes exactly from improvement along the noise directions (i.e. directions uncorrelated with the ground truth). Note that if we train for too long, the effect will be the same on all directions, so there will be no weak-to-strong generalization. So early stopping is crucial in this setup.

As student training progresses we fit more of the teacher's signal (i.e. the target of GF), but the shrinkage effect reduces. This can be seen in Figures 2 and 3, where overtraining hurts weak-to-strong generalization. At the limit, as training time goes to infinity, the student would converge to exactly mimicking the teacher, and thus converge to the teacher error, without any weak-to-strong benefit, as can be clearly seen in Figure 2.

The difficulty of applying this reasoning is that we need to establish that a significant portion of the teacher error is actually present in the noise direction. In Theorem 3.1 and Theorem 3.2, we not only show that a significant portion of the teacher's error $\mathcal{L}_{\mathrm{TE}}$ will be along the low energy directions but also that when the number of teacher's units M_{TE} is large, most of the teacher's error lies along the low energy, i.e. noise directions. Our main result, Theorem 6.5, handles this difficulty. Proving that most of the teacher error $\mathcal{L}_{\mathrm{TE}}$ is in the low energy directions allows us to conclude that the PGR goes to 1 as the number of teacher units M_{TE} increases, instead of just being positive, and establish a quantitative relationship between the student and teacher errors.

2-layer ReLU. For ReLU, we have a similar situation except that the relevant directions for a wide student will the spherical harmonics. Learning a predictor $f_{\boldsymbol{w},\boldsymbol{u}_s}$ using f as a target with early stopped gradient descent will shrink $f - f_{\boldsymbol{w},\boldsymbol{u}_s}$ in the direction of the student's random features $\sigma(\langle \boldsymbol{u}_{s,i}, \boldsymbol{x}\rangle) = \mathrm{ReLU}(\langle \boldsymbol{u}_{s,i}, \boldsymbol{x}\rangle)$, similarly to the case of linear networks. The relevant basis for a wide teacher, i.e. in the $M_{\mathrm{TE}} \to \infty$ case, is not the coordinates of the input $\boldsymbol{x}_i$ but rather spherical harmonics $\phi_{i,k}(\boldsymbol{x})$. The difference comes from the fact that our probability distribution is now $\mathcal{D} = \mathrm{Unif}(\mathbb{S}^{d-1})$, and the spherical harmonics $\{\{\phi_{i,k}\}_{i=1}^{N_k}\}_{k=1}^{\infty}$ are the relevant basis here. In this basis, the teacher predictor is given by $f_{\text{teacher}}(\boldsymbol{x}) = \sum_{i,k} \beta_{i,k}\phi_{i,k}(\boldsymbol{x})$.

The student will learn the teacher's coefficients $\beta_{i,k}$ for $\phi_{i,k}$ faster for the spherical harmonics of lower order k. Since the covariance of the features is given by the square of the coefficients in the decomposition of the ReLU function in the basis $\phi_{i,k}$, σ_k^2, they will naturally have jumps for certain indices k. Similarly, for any one of these jumps at, say, $k = K$, i.e. $\sigma_K \gg \sigma_{K+1}$, the gradient descent with a wide student will learn the directions $\phi_{i,k}(\boldsymbol{x})$ for $k \leq K$ significantly faster than $\phi_{i,k}(\boldsymbol{x})$ for $k > K$. So, again, we can choose an appropriate stopping time T such that the student predictor $f_{\boldsymbol{w},\boldsymbol{u}_s}$ learns the large variance directions, $\phi_{i,k}$ for $k \leq K$, of the teacher predictor f_{teacher} almost completely, but is very close to the initialization, i.e. zero, for small variance directions, $\phi_{i,k}$ for $k > K$. This means that the student can denoise the direction of $\phi_{i,k}$ for $k > K$. So, in particular, if all of the signal f^* is aligned with the directions of the first k_0 order of spherical harmonics, then everything the teacher learns in the directions $\phi_{i,k}$ for $k > k_0$ is the incorrectly learned noise. Zeroing out this incorrectly learned noise enables weak-to-strong generalization. To understand the inner workings of weak-to-strong generalization in this setup further, we turn to the dynamics of GF (Equation (5)) in Section 6. This will be simplified if we consider the student's network as a kernel.

6. General Multiplicative Weak-to-Strong Improvement

In Section 6, we show a general result that describes weak-to-strong improvement of a student described by a kernel trained with gradient flow and an optimally trained teacer with some fixed set of features, which is Theorem 6.5.

The application of Theorem 6.5 to the particular cases of ReLU and linear random feature networks is delayed to Appendix E.

Gradient Flow Dynamics Recap. We can rewrite Equation (5) in function space from parameter space in order to analyze the infinite-width limit $M_{\mathrm{ST}} \to \infty$ in the same space. To do that, it will be convenient to consider the student to be described by its kernel

$$\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') := \mathbb{E}_{\boldsymbol{u}\sim\mathrm{unif}(\mathbb{S}^{d-1})}\left[\sigma(\langle \boldsymbol{u}, \boldsymbol{x}\rangle)\sigma(\langle \boldsymbol{u}, \boldsymbol{x}'\rangle)\right]. \quad (8)$$

The kernel $\mathcal{K}$ can be decomposed as $\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{k\geq 1} \lambda_k e_k(\boldsymbol{x}) e_k(\boldsymbol{x}')$, where $\{\lambda_i\}_{i=1}^{\infty}$ are the eigenvalues of the associated kernel operator in descending order, and $\{e_i(\boldsymbol{x})\}_{i=1}^{\infty}$ are orthonormal eigenfunctions in the inner product with respect to $\mathcal{D}$, which we denote $\langle\cdot,\cdot\rangle_{\mathcal{D}}$, i.e. $\langle e_i, e_j\rangle_{\mathcal{D}} = \delta_{ij}$. The inner product is defined as $\langle f, g\rangle_{\mathcal{D}} := \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[f(\boldsymbol{x})g(\boldsymbol{x})]$. We can derive (see Appendix A) the following closed-form expression for the

gradient flow dynamics of the student (Equation (5)) at time t in the $M_{\text{ST}} \to \infty$ limit:

$$f_t = \sum_{k \geq 1} \left(1 - e^{-\lambda_k t}\right) \langle f_{\text{teacher}}, e_k \rangle_{\mathcal{D}} e_k \tag{9}$$

We note that given data distribution $\mathcal{D}$ (and thus the inner-product $\langle \cdot, \cdot \rangle_{\mathcal{D}}$), the above solution of gradient flow at time t is only a function of target f_{teacher}, time t, and the kernel $\mathcal{K}$, which we denote by $f_t = \mathcal{T}_t^{\mathcal{K}}(f_{\text{teacher}})$, where $\mathcal{T}_t^{\mathcal{K}} \triangleq \texttt{id} - e^{-t\mathcal{K}}$, and $\texttt{id}$ is the identity mapping. In other words, if different two-layer networks have different bottom layers but with the same induced kernel, they lead to same training behavior under gradient flow. For an outline of the derivation of gradient flow dynamics, see Appendix A.

Student-Teacher Improvement Bound for Early Stopped Gradient Flow. First, in Lemma 6.2, we will derive a bound on the error of a student trained with an early stopped gradient flow in the weak-to-strong setup. Consider a generalized 2-stage learning process of weak-to-strong setup from Section 2, where the student is trained with gradient flow for time t with kernel $\mathcal{K}$ on population given by the teacher predictor f_{teacher}, i.e. as given by Equation (9) and denoted with $f_t = \mathcal{T}_t^{\mathcal{K}}(f_{\text{teacher}})$, but the teacher predictor can be an arbitrary square integrable function in the eigenspace of $\mathcal{K}$. Furthermore, we require that the student's kernel $\mathcal{K}$ correctly captures the prior of the target function f^* in the sense that f^* is supported on the top-K eigenspace of $\mathcal{K}$ with nonzero eigenvalues:

Condition 6.1 (Ground Truth)**.** f^* is supported on the top-K eigenspace of $\mathcal{K}$ corresponding to nonzero eigenvalues, i.e., if $\lambda_k = 0$ for $k \leq K$ then $\langle f^*, e_k \rangle_{\mathcal{D}} = 0$ and $\forall k > K : \langle f^*, e_k \rangle_{\mathcal{D}} = 0$.

Then, we can bound the students error at time T with respect to the ground truth f^* in terms of the spectrum of the kernel λ_i.

Lemma 6.2. *Under Condition 6.1 for any kernel $\mathcal{K}$, any function $f_{teacher}$ in the eigenspace of $\mathcal{K}$, all stopping times $T > 0$, and any index $S \geq K$, we have that for the predictor $f_t = \mathcal{T}_t^{\mathcal{K}}(f_{teacher})$, then $\mathcal{L}(f_T) \leq \mathcal{L}(f_{teacher}) + \frac{e^{-\lambda_K T}}{2 - e^{-\lambda_K T}} \|f^*\|_{\mathcal{D}}^2 - (1 - \lambda_{S+1}^2 T^2) \sum_{k \geq S+1} \langle f_{teacher}, e_k \rangle_{\mathcal{D}}^2$.*

Lemma 6.2 is an intermediate lemma that shows that the error of an early stopped GF solution with respect to the grond truth f^* can be bounded in terms of the target f_{teacher} and the projection of the target f_{teacher} into the eigenspace of order $\geq S + 1$, $\sum_{k \geq S+1} \langle f_{\text{teacher}}, e_k \rangle_{\mathcal{D}}^2$. So, to get the final weak-to-strong upper bound (i.e. Theorem 6.5), we show in Lemma 6.4 that for the optimally trained teacher, the proportion of error in eigendirections of order $\geq S+1$ is lower bounded by a quantity we call *teacher-student feature alignment*.

Teacher-Student Feature Alignment. The key quantity in our characterization of weak-to-strong generalization gap is *teacher-student feature alignment*, κ_S. Let $M_{\text{TE}} = m$ and let $\{e_i\}_{i=1}^{\infty}$ be the eigenbasis of the student's kernel $\mathcal{K}$. Let $\boldsymbol{w} \in \mathbb{R}^m$ be the teacher's weights, and let the teacher be parametrized as $f_{\text{teacher}} = f_w = \sum_{i=1}^m w_i g_i$, where $\{g_i\}_{i=1}^m$ are teacher's units, i.e. features. Let the student's predictor at time T be f_T. For $S \geq K$, let $\boldsymbol{A}_S, \boldsymbol{B}_S \in \mathbb{R}^{m \times m}$ be the Gram matrices of teacher's features projected onto the top-S eigenspace of $\mathcal{K}$, *i.e.*, $A_{S,i,j} := \sum_{k=1}^S \langle g_i, e_k \rangle_{\mathcal{D}} \langle g_j, e_k \rangle_{\mathcal{D}}$, $B_{S,i,j} := \sum_{k \geq S+1} \langle g_i, e_k \rangle_{\mathcal{D}} \langle g_j, e_k \rangle_{\mathcal{D}}$. Then we define the teacher-student feature alignment κ_S as

$$\kappa_S := \frac{1}{1 + \lambda_1 \left((\sqrt{\boldsymbol{A}_S})^+ \boldsymbol{B}_S (\sqrt{\boldsymbol{A}_S})^+ \right)}. \tag{10}$$

κ_S measures the alignment of the teacher's features with the top-S eigenspace of the student kernel and is a completely deterministic quantity. As explained in Section 5, this misalignment is crucial for enabling weak-to-strong generalization in this setup.

We will refer to $\hat{f}_w$ the optimally trained teacher. We will require that no two units in the first layer of the teacher will be linearly dependent among the first K directions $\{e_i\}_{i=1}^K$.

Condition 6.3. $\text{rank}(\boldsymbol{A}_K) = K$.

Lemma 6.4 (Error Ratio of Weak-to-Strong Predictor in High Order Eigendirections)**.** *In the random feature model in Equation (1), under Conditions 6.1 and 6.3, if $\mathcal{L}(\hat{f}_{\boldsymbol{w}}) = \min_{\boldsymbol{w}' \in \mathbb{R}^m} \mathcal{L}(\hat{f}_{\boldsymbol{w}'})$, then $\sum_{k \geq S+1} \langle \hat{f}_{\boldsymbol{w}}, e_k \rangle_{\mathcal{D}}^2 \geq \kappa_S \cdot \mathcal{L}(\hat{f}_{\boldsymbol{w}})$.*

Lemma 6.4 shows that if the signal is spanned by first S eigendirections then at least κ_S portion of error of the random feature predictor in Equation (1) will be in the eigendirections of order $\geq S + 1$.

Using Equation (9), Lemma 6.2, and Lemma 6.4 we can understand how weak-to-strong happens from Section 5 more clearly. In this notation, in Section 5, $\langle f_{\text{teacher}}, e_k \rangle_{\mathcal{D}}$ is what we call β_i and $e_i(\boldsymbol{x})$ are x_i in Model 2.3 and spherical harmonics $\phi_{i,k}(\boldsymbol{x})$ in Model 2.2. Note that since λ_k are decreasing, $1 - e^{-\lambda_k t}$ is closer to 1 for larger eigenvalues λ_i, i.e. smaller indices i. In the special case that there is K for which λ_{K+1} is significantly smaller than λ_K, then for stopping time $T \in (\frac{1}{\lambda_K}, \frac{1}{\lambda_{K+1}})$, all $e^{-\lambda_i T}$ all going to be close to 0 for $i \leq K$, but $e^{-\lambda_i T}$ for $i \geq K + 1$ will be close to 1, effectively zeroing out f_{teacher} in directions $e_{\geq K+1}$. This means that in Lemma 6.2, the second term $\frac{e^{-\lambda_K T}}{2 - e^{-\lambda_K T}} \|f^*\|_{\mathcal{D}}^2$ will be small. Taking $S = K + 1$, the thirm term will be close to κ_{K+1} portion of the error by Lemma 6.4, which Lemma 6.2 shows that we zero out. This is the intuition behind Theorem 6.5.

Multiplicative Error Improvement in Weak-to-Strong Generalization. Our main result establishes multiplicative error improvement in the weak-to-strong setup for a

2-layer network teacher with any features, *even non-random*, and a student described by a kernel $\mathcal{K}$.

Theorem 6.5 (Weak-to-Strong Multiplicative Error Improvement)**.** *For any groundtruth f^*, p.s.d. kernel $\mathcal{K}$ and positive integer K satisfying Condition 6.1 and any teacher model $f_{teacher}$ satisfying Condition 6.3 and Condition 2.1 ($f_{teacher}$ attaining minimum loss), then for any time $T > 0$, then the test loss of student $f_{student} \triangleq \mathcal{T}_T^{\mathcal{K}}(f_{teacher})$ has the following upper bound:*

$$\mathcal{L}_{\mathrm{ST}} \leq \inf_{S \geq K} \left\{ \left(1 - (1 - \lambda_{S+1}^2 T^2)\kappa_S\right) \mathcal{L}_{\mathrm{TE}} + \frac{e^{-\lambda_K T}}{2 - e^{-\lambda_K T}} \|f^*\|_{\mathcal{D}}^2 \right\}.$$

Theorem 6.5 shows that if there is a large enough eigengap between λ_i so that κ_S is close to 1, then we can select the early stopping time T so that the student is only left with $(1 - \kappa_S)$ fraction of teacher's loss. Note that Theorem 6.5 is completely deterministic and holds both or random and non-random teachers. For the proof of Lemma 6.2, Lemma 6.4, and Theorem 6.5, and the explanation for when equality holds, see Appendix C.

7. Related work

Theoretical Understanding of Weak-to-Strong Generalization. Lang et al. (2024) propose a theoretical framework that establishes weak-to-strong generalization for classification when the strong student is assumed to learn functions that are robust to mislabeled data among many correctly labeled neighbors. Shin et al. (2025) use this framework to formulate a mechanism for weak-to-strong generalization for data with easy and hard patterns. Charikar et al. (2024) study a general regression setup with squared loss, where a key example is training only the last linear layer of the teacher and student networks. They show that if the strong student's capacity is not large enough to express the weak teacher's function, then the student can outperform the teacher by an amount quantified by how much the student does not fit the teacher's labels. Mulgund & Pabbaraju (2025) further generalize this result to a broader class of loss functions, including the cross entropy loss. Wu & Sahai (2025) demonstrated that weak-to-strong generalization can happen in an overparametrized spiked covariance model for classification. (Ildiz et al., 2025) show that in a two-stage learning process in high-dimensional ridgeless regression, a student trained on teacher labels can be better than a student trained on real labels (which is an effect of "distillation" rather than "weak-to-strong").

Charikar et al. (2024) and Wu & Sahai (2025) consider linear models that are most closely related to our setup. Charikar et al. (2024) consider a general regression setup with square loss and (Wu & Sahai, 2025) consider classification with a minimum norm interpolator. In the setup of Charikar et al. (2024), if the teacher and student classes contain the origin, our lower bound applies to their setting.

Empirical Work. Following up on Burns et al. (2024), many works explored ways to leverage weak models to improve the performance of strong models. Shin et al. (2025); Li et al. (2024) propose data selection methods to improve W2S generalization. Somerstep et al. (2025) propose to use in-context learning to refine weak teacher labels. Yang et al. (2024); Bansal et al. (2025) show that small models can be effectively used to generate reasoning data with Chain-of-Thought (CoT) for supervising large models - Bansal et al. (2025) point out that this can be more compute-optimal as compared to generating data from large models. Sun et al. (2024) showed that a weak model can be trained as a reward model on easy tasks and then used to guide the training of a strong model on harder tasks. Ji et al. (2024) proposed a plug-and-play module for alignment, where a small model learns to redistribute the output of a large model. Tao & Li (2025) systematically evaluated the performance of aligning a large model with feedback from a small model instead of human feedback. Yang et al. (2025) highlighted that the strong model can deceive the weak model by being well aligned only in the areas familiar to the weak model.

8. Conclusion

We prove that weak-to-strong generalization can happen even in a simple model of two-layer neural networks with random features. We show that in this case, the PGR (Performance Gap Recovered) can converge to 1 when teacher loss goes to 0. Additionally, we show an even stronger separation between the teacher and the student, namely that the student error is polynomially smaller than the teacher error. For the ReLU network, we show that the exponent in this polynomial dependence is at least 1.49 while for the linear networks, we show that it is 2. Further, we explain that in his setup, early stopping is crucial for weak-to-strong generalization to occur. We also explain that in this setup, the weak-to-strong improvement comes from the student zeroing out the teacher signal in the high-order eigendirections (which in our setup are noise directions under the condition that the ground truth is spanned by the first few eigendirections). Most interestingly, we show that there is an inherent limitation to weak-to-strong generalization in random feature models trained with gradient descent or gradient flow, and even more generally. Namely, we show that if the target is normalized, the student error can be at most quadratically smaller than the teacher error. Our quadratic lower bound asserts that the student loss cannot be smaller than the square of teacher loss, regardless of the choice of student feature, the early stopping time.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. The phenomenon of weak to strong generalization was introduced as a route to scalable oversight, where the weak model is the human overseer who is trying to align the strong model, which is a super-AGI. The current paper develops a fundamental understanding of this phenomenon in the context of the random features model. Currently, there are no societal implications, but in the future, this work could lead to a better understanding of the uses of W2SG in super alignment. But that is currently hypothetical.

Acknowledgements

ZL and SA are supported by OpenAI superalignment grant. SA is also supported by NSF, DARPA, and ONR. MM would like to thank Theodor Misiakiewicz for providing the reference (Defilippis et al., 2024) and for a useful discussion about it.

References

Ali, A., Kolter, J. Z., and Tibshirani, R. J. A continuous-time view of early stopping for least squares. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan*, 2019.

Arriaga, R. I. and Vempala, S. An algorithmic theory of learning: Robust concepts and random projection. *Machine Learning*, 63-2:161–182, 2006.

Bansal, H., Hosseini, A., Agarwal, R., Tran, V. Q., and Kazemi, M. Smaller, weaker, yet better: Training llm reasoners via compute-optimal sampling. *The Thirteenth International Conference on Learning Representations*, 2025.

Burns, C., Izmailov, P., Kirchner, J. H., Baker, B., Gao, L., Aschenbrenner, L., Chen, Y., Ecoffet, A., Joglekar, M., Leike, J., Sutskever, I., and Wu, J. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=ghNRg2mEgN.

Canatar, A., Bordelon, B., and Pehlevan, C. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature Communications*, 12(1):1–12, 2021.

Charikar, M., Pabbaraju, C., and Shiragur, K. Quantifying the gain in weak-to-strong generalization. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 126474–126499. Curran Associates, Inc., 2024.

Defilippis, L., Loureiro, B., and Misiakiewicz, T. Dimension-free deterministic equivalents and scaling laws for random feature regression. *38th Conference on Neural Information Processing Systems (NeurIPS 2024).*, 2024.

Elkabetz, O. and Cohen, N. Continuous vs. discrete optimization of deep neural networks. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Ildiz, M. E., Gozeten, H. A., Taga, E. O., Mondelli, M., and Oymak, S. High-dimensional analysis of knowledge distillation: Weak-to-strong generalization and scaling laws. *The Thirteenth International Conference on Learning Representations*, 2025.

Ji, J., Chen, B., Lou, H., Hong, D., Zhang, B., Pan, X., Qiu, T. A., Dai, J., and Yang, Y. Aligner: Efficient alignment by learning to correct. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 90853–90890. Curran Associates, Inc., 2024.

Lang, H., Sontag, D., and Vijayaraghavan, A. Theoretical analysis of weak-to-strong generalization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Li, M., Zhang, Y., He, S., Li, Z., Zhao, H., Wang, J., Cheng, N., and Zhou, T. Superfiltering: Weak-to-strong data filtering for fast instruction-tuning. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14255–14273, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.769.

Lin, L., Wu, J., Kakade, S. M., Bartlett, P. L., and Lee, J. D. Scaling laws in linear regression: Compute, parameters, and data. *38th Conference on Neural Information Processing Systems (NeurIPS 2024)*, 2024.

Mei, S., Misiakiewicz, T., and Montanari, A. Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, 2022. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2021.12.003. URL https://www.sciencedirect.com/

science/article/pii/S1063520321001044. Special Issue on Harmonic Analysis and Machine Learning.

Misiakiewicz, T. and Saeed, B. A non-asymptotic theory of kernel ridge regression: deterministic equivalents, test error, and GCV estimator. *CoRR*, abs/2403.08938, 2024. doi: 10.48550/ARXIV.2403.08938. URL https://doi.org/10.48550/arXiv.2403.08938.

Mulgund, A. and Pabbaraju, C. Relating misfit to gain in weak-to-strong generalization beyond the squared loss. *arXiv preprint arXiv:2501.19105*, 2025.

Paquette, E., Paquette, C., Xiao, L., and Pennington, J. 4+3 phases of compute-optimal neural scaling laws. *38th Conference on Neural Information Processing Systems (NeurIPS 2024)*, 2024.

Petrini, L., Cagnetta, F., Vanden-Eijnden, E., and Wyart, M. Learning sparse features can lead to overfitting in neural networks. *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, 2022.

Pinelis, I. Optimum bounds for the distributions of martingales in banach spaces. *Ann. Probab. 22(4)*, pp. 1679–1706, 1994.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. *Advances in Neural Information Processing Systems 20 (NIPS 2007)*, 2007.

Samuel, A. L. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):210–229, 1959. doi: 10.1147/rd.33.0210.

Shin, C., Cooper, J., and Sala, F. Weak-to-strong generalization through the data-centric lens. In *The Thirteenth International Conference on Learning Representations*, 2025.

Simon, J. B., Dickens, M., Karkada, D., and DeWeese, M. R. The eigenlearning framework: A conservation law perspective on kernel regression and wide neural networks. *Transactions on Machine Learning Research*, 2023.

Somerstep, S., Polo, F. M., Banerjee, M., Ritov, Y., Yurochkin, M., and Sun, Y. A transfer learning framework for weak to strong generalization. In *The Thirteenth International Conference on Learning Representations*, 2025.

Sun, Z., Yu, L., Shen, Y., Liu, W., Yang, Y., Welleck, S., and Gan, C. Easy-to-hard generalization: Scalable alignment beyond human supervision. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 51118–51168. Curran Associates, Inc., 2024.

Tao, L. and Li, Y. Your weak LLM is secretly a strong teacher for alignment. In *The Thirteenth International Conference on Learning Representations*, 2025.

Vershynin, R. *Introduction to the non-asymptotic analysis of random matrices*, pp. 210–268. Cambridge University Press, 2012.

Wei, A., Hu, W., and Steinhardt, J. More than a toy: Random matrix models predict how real-world neural representations generalize. In *International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2022.

Wu, D. X. and Sahai, A. Provable weak-to-strong generalization via benign overfitting. In *The Thirteenth International Conference on Learning Representations*, 2025.

Yang, W., Shen, S., Shen, G., Yao, W., Liu, Y., Zhi, G., Lin, Y., and Wen, J.-R. Super(ficial)-alignment: Strong models may deceive weak models in weak-to-strong generalization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=HxKSzulSD1.

Yang, Y., Ma, Y., and Liu, P. Weak-to-strong reasoning. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8350–8367, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.490.

Appendix

A. Gradient Flow Dynamics

We first introduce the function space view of gradient flow over the top layer of the student's network Equation (5). The main benefit of working in function space from parameter space is allowing us to analyze the infinite-width limit in the same space. We first recall we parametrize the student as:

$$f_{\text{student}} = f_{\boldsymbol{w}_{\text{s}},\boldsymbol{u}_{\text{s}}}(\boldsymbol{x}) = \sum_{i=1}^{M_{\text{ST}}} w_{\text{s},i}\sigma(\langle \boldsymbol{u}_{\text{s},i}, \boldsymbol{x}\rangle).$$

Let $\sigma_{\text{s},i}(\boldsymbol{x}) = \sigma(\langle \boldsymbol{u}_{\text{s},i}, \boldsymbol{x}\rangle)$. Then, Equation (5) in function space is

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} f_{\boldsymbol{w}_{\text{s}}(t),\boldsymbol{u}_{\text{s}}}(\boldsymbol{x}) &= -\frac{1}{M_{\text{ST}}}\sum_{i=1}^{M_{\text{ST}}} \mathbb{E}_{\boldsymbol{y}\sim\mathcal{D}}\langle f_{\boldsymbol{w}_{\text{s}}(t),\boldsymbol{u}_{\text{s}}}(\boldsymbol{y}) - f_{\text{teacher}}(\boldsymbol{y}), \sigma_{\text{s},i}(\boldsymbol{y})\rangle\, \sigma_{\text{s},i}(\boldsymbol{x}) \\ &= -\mathbb{E}_{\boldsymbol{y}\sim\mathcal{D}}\left[K_{M_{\text{ST}}}(\boldsymbol{x},\boldsymbol{y})(f_{\boldsymbol{w}_{\text{s}}(t),\boldsymbol{u}_{\text{s}}}(\boldsymbol{y}) - f_{\text{teacher}}(\boldsymbol{y}))\right]\end{aligned} \tag{11}$$

where $K_{M_{\text{ST}}}(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{M_{\text{ST}}}\sum_{i=1}^{M_{\text{ST}}} \sigma_{\text{s},i}(\boldsymbol{x})\sigma_{\text{s},i}(\boldsymbol{y})$ is the empirical kernel.

Taking $M_{\text{ST}} \to \infty$, the solution of Equation (11) converges to the solution of the following equation:

$$\frac{\mathrm{d}}{\mathrm{d}t} f_t(\boldsymbol{x}) = -\mathbb{E}_{\boldsymbol{y}\sim\mathcal{D}}\left[\mathcal{K}(\boldsymbol{x},\boldsymbol{y})\left(f_t(\boldsymbol{y}) - f_{\text{teacher}}(\boldsymbol{y})\right)\right], \tag{12}$$

where

$$\mathcal{K}(\boldsymbol{x},\boldsymbol{x}') := \mathbb{E}_{\boldsymbol{u}\sim\text{unif}(\mathbb{S}^{d-1})}\left[\sigma(\langle \boldsymbol{u},\boldsymbol{x}\rangle)\sigma(\langle \boldsymbol{u},\boldsymbol{x}'\rangle)\right]. \tag{13}$$

The kernel $\mathcal{K}$ can be decomposed as $\mathcal{K}(\boldsymbol{x},\boldsymbol{x}') = \sum_{k\geq 1}\lambda_k e_k(\boldsymbol{x})e_k(\boldsymbol{x}')$, where $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots$ are the eigenvalues of the associated kernel operator in descending order, and $e_1(\boldsymbol{x}), e_2(\boldsymbol{x}), e_3(\boldsymbol{x}), \ldots$ are orthonormal eigenfunctions, that is, $\langle e_i, e_j\rangle_{\mathcal{D}} = \delta_{ij}$. This allows us to decompose the gradient flow dynamics (Equation (11)) into simple ODEs for each eigendirection, which can be easily solved. With this decomposition, we can derive the following closed-form expression for the gradient flow dynamics of the student (Equation (5)) at time t:

$$f_t = \sum_{k\geq 1}\left(1 - e^{-\lambda_k t}\right)\langle f_{\text{teacher}}, e_k\rangle_{\mathcal{D}} e_k \tag{14}$$

Theorem A.1 (Closed Form Solution to Gradient Flow Dynamics)**.** *Let $\mathcal{K} : \mathcal{X}\times\mathcal{X} \to \mathcal{R}$ be a p.s.d. kernel with eigendecomposition $\mathcal{K}(\boldsymbol{x},\boldsymbol{x}') = \sum_{i=1}^{\infty}\lambda_i e_i(\boldsymbol{x})e_i(\boldsymbol{t})$, where $\langle e_i, e_j\rangle_{\mathcal{D}} = \delta_{ij}$. For any $f^* : \mathcal{X}\to\mathbb{R}$, the equation of gradient flow for a predictor f_t*

$$\frac{\mathrm{d}}{\mathrm{d}t} f_t(\boldsymbol{x}) = -\mathbb{E}_{\boldsymbol{y}\sim\mathcal{D}}\left[\mathcal{K}(\boldsymbol{x},\boldsymbol{y})\left(f_t(\boldsymbol{y}) - f^*(\boldsymbol{y})\right)\right], \tag{15}$$

with initial condition given by $f_0 = 0$ has a closed form solution f_t

$$f_t = \sum_{k\geq 1}\left(1 - e^{-\lambda_k t}\right)\langle f^*, e_k\rangle_{\mathcal{D}} e_k.$$

Proof of Theorem A.1. To solve the gradient flow equation in (15), we project it onto the eigenbasis of the kernel operator $\mathcal{K}$. Let $\beta_t^i = \langle f_t, e_i\rangle_{\mathcal{D}}$ be the coefficients of f_t.

Taking the inner product of both sides of (15) with e_j:

$$\begin{aligned}\frac{d}{dt}\beta_t^j &= -\mathbb{E}_{y\sim\mathcal{D}}\left[\langle \mathcal{K}(\cdot,y), e_j\rangle_{\mathcal{D}}\left(f_t(y) - f^*(y)\right)\right] \\ &= -\lambda_j\langle e_j, f_t - f^*\rangle_{\mathcal{D}} \\ &= -\lambda_j(\beta_t^j - \langle f^*, e_j\rangle_{\mathcal{D}})\end{aligned}$$

Solving this ODE with initial condition $\beta_0^j = 0$:

$$\beta_t^j = (1 - e^{-\lambda_j t})\langle f^*, e_j\rangle_{\mathcal{D}}.$$

Therefore,

$$f_t = T_t^{\mathcal{K}} f^* = \sum_{j=1}^{\infty}(1 - e^{-\lambda_j t})\langle f^*, e_j\rangle_{\mathcal{D}} e_j,$$

which concludes the proof. □

B. Different Model Types

We will first introduce a model that is slightly more general than Model 2.3.

Model B.1 (Diagonal Covariate Features)**.** Let $\{e_i\}_{i=1}^{\infty}$ be a fixed orthonormal basis of $L_2^{\mathcal{D}}(\mathcal{X})$. Take the random features to be given by

$$g = \sum_{i=1}^{\infty}\langle g, e_i\rangle_{\mathcal{D}} e_i$$

where $\{\langle g, e_i\rangle_{\mathcal{D}}\}_{i=1}^{\infty} \sim N(0, \Lambda)$ where $\Lambda = (\lambda_1, \lambda_2, \dots)$ in decreasing order.

Our results will hold for Model B.1 and they are more general than Model 2.3.

Proposition B.2 (Linear network as Diagonal Coavariate Features)**.** *To get Model 2.3 from Model B.1, we take* $\Lambda = (\psi_1, \dots, \psi_d, 0, 0, \dots)$, $\langle x \mapsto x^\top u, x \mapsto x^\top v\rangle_{\mathcal{D}} = \sum_{i=1}^{d}\psi_i u_i v_i$, *and* $e_i(x) = x_i/\sqrt{\psi_i}$. *Thus if we set* g_i *in Model B.1 as* $g(x) = \boldsymbol{x}^\top \boldsymbol{u}_i$, *where* $\boldsymbol{U} \sim N(0, I_d)$, *then* $\{\langle g, e_i\rangle_{\mathcal{D}}\}_{i=1}^{\infty} \sim N(0, \Lambda)$.

Proposition B.3 (ReLU Network as Diagonal Covariate Features)**.** *If in Model B.1 we take* $\mathcal{D} = Unif(\mathbb{S}^{d-1})$, $\{e_i\}_{i=1}^{\infty}$ *to be the spherical harmonics* $\{\{\phi_{i,k}\}_{i=1}^{N_k}\}_{k=1}^{\infty}$, *and choose* $\Lambda = (\sigma_1, \dots, \sigma_1, \sigma_2, \dots, \sigma_2, \dots, \sigma_k, \dots, \sigma_k, \dots)$, *where* σ_k *repeats* N_k *times, we recover the same feature covariance as in Model 2.2. Under Gaussian Universality assumptions, for this instance of Model B.1, the teacher and the student risk behave the same as in Model 2.2.*

Proof of Proposition B.3. In Model 2.2, the kernel induce by the random features is given by

$$\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') := \mathbb{E}_{\boldsymbol{u}\sim \text{unif}(\mathbb{S}^{d-1})}\left[\sigma(\langle \boldsymbol{u}, \boldsymbol{x}\rangle)\sigma(\langle \boldsymbol{u}, \boldsymbol{x}'\rangle)\right].$$

A random feature in the ReLU case Model 2.2 is given by

$$\sigma(\langle \boldsymbol{w}, \boldsymbol{x}\rangle) = \sum_{k\geq 0} N_k \sigma_k P_{d,k}(\langle \boldsymbol{w}, \boldsymbol{x}\rangle) = \sum_{k\geq 0}\sigma_k\langle \boldsymbol{\phi}_k(\boldsymbol{w}), \boldsymbol{\phi}_k(\boldsymbol{x})\rangle = \sum_{k=0}^{\infty}\sum_{i=1}^{N_k}\sigma_k\phi_{i,k}(\boldsymbol{w})\phi_{i,k}(\boldsymbol{x}).$$

A random feature in the Diagonal Feature Covariance case Model B.1 is given by

$$g(\boldsymbol{x}) = \sum_{k=0}^{\infty}\sum_{i=1}^{N_k} g_{i,k}\phi_{i,k}(\boldsymbol{x}).$$

$$\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}_{\boldsymbol{u}}[\sigma(\langle \boldsymbol{u}, \boldsymbol{x}\rangle)\sigma(\langle \boldsymbol{u}, \boldsymbol{x}'\rangle)] = \mathbb{E}_{\boldsymbol{u}}\left[\sum_{k\geq 0}\sigma_k\langle \boldsymbol{\phi}_k(\boldsymbol{u}), \boldsymbol{\phi}_k(\boldsymbol{x})\rangle \cdot \sigma_k\langle \boldsymbol{\phi}_k(\boldsymbol{u}), \boldsymbol{\phi}_k(\boldsymbol{x}')\rangle\right] = \sum_{k\geq 0}\sigma_k^2\langle \boldsymbol{\phi}_k(\boldsymbol{x}), \boldsymbol{\phi}_k(\boldsymbol{x}')\rangle.$$

On the other hand, we have that for Model B.1 the kernel is

$$\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}_g\left[g(\boldsymbol{x})g(\boldsymbol{x}')\right] = \mathbb{E}_g\left[\sum_{i,k}\sum_{j,l} g_{i,k}\phi_{i,k}(\boldsymbol{x})g_{j,l}\phi_{j,l}(\boldsymbol{x})\right] = \sum_{k=0}^{\infty}\sigma_k^2\langle \boldsymbol{\phi}_k(\boldsymbol{x}), \boldsymbol{\phi}_k(\boldsymbol{x}')\rangle.$$

□

B.1. Diagonalizability of the Feature Covariance in Model 2.3

In this section we establish the generality of Model 2.3. Namely, we show that one can assume that $\boldsymbol{u}$ is isotropic Gaussuian distribution and $\boldsymbol{x}$ has diagonal covariance if and only if their covariance matrices are codiagonalizable.

Proposition B.4 (Diagonalizability of the Linear Network Covariance)**.** *A linear network $\sigma(z) = z$ with bottom layer weights $\boldsymbol{u}$ distributed accoding to a Gaussian distribution $\mathcal{U} = N(0, \Psi_{\boldsymbol{u}})$ and the inputs $\boldsymbol{x}$ distributed according to $\mathcal{D} = N(0, \Psi_{\boldsymbol{x}})$ is equivalent to Model 2.3 if and only if $\Psi_{\boldsymbol{x}}$ and $\Psi_{\boldsymbol{u}}$ are codiagonalizable, with $\Psi = \boldsymbol{V}^T \boldsymbol{S}_{\boldsymbol{u}}^{-1/2} \boldsymbol{V}^T \boldsymbol{S}_{\boldsymbol{x}} \boldsymbol{V} \boldsymbol{S}_{\boldsymbol{u}}^{-1/2} \boldsymbol{V}$ where $\boldsymbol{V}$ is the shared basis.*

Proof of Proposition B.4. Note that the distribution in the first case is given by $z \mid \boldsymbol{x} \sim \langle \boldsymbol{u}, \boldsymbol{x} \rangle \mid \boldsymbol{x} \sim N(0, \boldsymbol{x} \Sigma_u \boldsymbol{x}^T)$. Note further that for Model 2.3 with Ψ we have that $\tilde{z} \mid \tilde{\boldsymbol{x}} \sim \langle \tilde{\boldsymbol{u}}, \tilde{\boldsymbol{x}} \rangle \mid \tilde{\boldsymbol{x}} \sim N(0, \Psi)$. Now if $\tilde{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x}$ for some $\boldsymbol{A}$, then we have that $\Psi = cov(\tilde{\boldsymbol{x}}) = cov(\boldsymbol{A}\boldsymbol{x}) = \boldsymbol{A}^T \Sigma_x \boldsymbol{A} = \boldsymbol{A}^T \boldsymbol{V}_x \boldsymbol{S}_x \boldsymbol{V}_x \boldsymbol{A}$, where $\boldsymbol{V}_x$ is the basis of Σ_x, i.e. $\Sigma_x = \boldsymbol{V}_x^T \boldsymbol{S}_x \boldsymbol{V}_x$ for $\boldsymbol{S}_x$ diagonal. Since Ψ is diagonal. So $\boldsymbol{A}^T \boldsymbol{V}_x \boldsymbol{S}_x \boldsymbol{V}_x \boldsymbol{A}$ is diagonal if and only if $\boldsymbol{A} = \boldsymbol{D}\boldsymbol{V}_x$ for $\boldsymbol{V}_x$ a basis of Σ_x and some diagonal matrix $\boldsymbol{D}$. So the distributions of z and $\tilde{z}$ are the same if and only if $\boldsymbol{A} = \boldsymbol{D}\boldsymbol{V}_x$ and $\boldsymbol{x} \Sigma_{\boldsymbol{u}} \boldsymbol{x}^T = \boldsymbol{x} \boldsymbol{V}_x \boldsymbol{D}^2 \boldsymbol{V}_x \boldsymbol{x}^T$, so $\Sigma_{\boldsymbol{u}} = \boldsymbol{V}_x \boldsymbol{D}^2 \boldsymbol{V}_x$. □

C. Proof of Weak-to-Strong Multiplicative Error Improvement

Here we present the proof of Theorem 6.5. We will introduce the following notation. Let $M_{\mathrm{TE}} = m$ and let $\{e_i\}_{i=1}^{\infty}$ be the eigenbasis of the student's kernel $\mathcal{K}$. Let the units, i.e. features, of the student be functions g drawn from some distribution $\mathcal{G}$. For example, in Model 2.2, we gave that $\mathcal{G}$ is given by $g(\boldsymbol{x}) = \sigma(\langle \boldsymbol{u}, \boldsymbol{x} \rangle)$ where $\boldsymbol{u} \sim \mathrm{unif}(\mathbb{S}^{d-1})$. Let $\boldsymbol{w} \in \mathbb{R}^m$ be the teacher's weights, and let the teacher be parametrized as $f_{\text{teacher}} = f_w = \sum_{i=1}^m w_i g_i$, where $\{g_i\}_{i=1}^m$ are teacher's units, i.e. features. We will refer to $\hat{f}_w$ the optimally trained teacher. Let the student's predictor at time T be f_T. Let $\mathcal{H}$ be the Hilbert space of square-integrable functions with respect to $\mathcal{D}$. Let $\boldsymbol{\Phi} \in \mathbb{R}^{m \times m}$ be the Gram matrix of random units, *i.e.*, $\Phi_{ij} := \langle \sigma_i, \sigma_j \rangle_{\mathcal{D}} = \sum_{k \geq 1} \langle \sigma_i, e_k \rangle_{\mathcal{D}} \langle \sigma_j, e_k \rangle_{\mathcal{D}}$.

C.1. Preliminary Lemmas

Lemma C.1. *For all $k, k' \geq 1$,*

$$\mathbb{E}_{g \sim \mathcal{G}}[\langle g, e_k \rangle_{\mathcal{D}} \langle g, e_{k'} \rangle_{\mathcal{D}}] = \begin{cases} \lambda_k & \text{if } k = k', \\ 0 & \text{otherwise.} \end{cases}$$

Lemma C.2. *For $c \in [0, 1)$,*

$$((1-c)(b+\Delta) - b)^2 \leq \Delta^2 + \frac{c}{2-c} b^2.$$

Proof of Lemma C.2.

$$\begin{aligned} ((1-c)(b+\Delta) - b)^2 = ((1-c)\Delta - cb)^2 &= \left((1-c)^2 \cdot \frac{1}{1-c} \Delta + \left(2c - c^2\right) \cdot \frac{1}{2c - c^2} (-cb) \right)^2 \\ &\leq (1-c)^2 \cdot \frac{1}{(1-c)^2} \Delta^2 + \left(2c - c^2\right) \cdot \frac{b^2}{(2-c)^2} \\ &= \Delta^2 + \frac{c}{2-c} b^2, \end{aligned}$$

where the inequality follows from Jensen's inequality. □

The following is a standard result on the generalized Rayleigh quotient.

Lemma C.3. *If $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{m \times m}$ are symmetric PSD matrices, then*

$$\sup_{\boldsymbol{u} \in \mathrm{colspan}(\boldsymbol{X})} \frac{\boldsymbol{u}^\top \boldsymbol{Y} \boldsymbol{u}}{\boldsymbol{u}^\top \boldsymbol{X} \boldsymbol{u}} = \lambda_1 \left((\sqrt{\boldsymbol{X}})^+ \boldsymbol{Y} (\sqrt{\boldsymbol{X}})^+ \right).$$

Proof. Let $\boldsymbol{u} \in \mathrm{colspan}(\boldsymbol{X}) \setminus \{\boldsymbol{0}\}$. Then $\boldsymbol{u} = \sqrt{\boldsymbol{X}}\boldsymbol{z}$ for some $\boldsymbol{z} \in \mathbb{R}^m \setminus \{\boldsymbol{0}\}$. Thus,

$$\frac{\boldsymbol{u}^\top \boldsymbol{Y} \boldsymbol{u}}{\boldsymbol{u}^\top \boldsymbol{X} \boldsymbol{u}} = \frac{\boldsymbol{z}^\top (\sqrt{\boldsymbol{X}})^+ \boldsymbol{Y} (\sqrt{\boldsymbol{X}})^+ \boldsymbol{z}}{\boldsymbol{z}^\top \boldsymbol{z}} \le \lambda_1\Big((\sqrt{\boldsymbol{X}})^+ \boldsymbol{Y} (\sqrt{\boldsymbol{X}})^+\Big).$$

The equality is achieved when $\boldsymbol{z}$ is the top eigenvector of $(\sqrt{\boldsymbol{X}})^+ \boldsymbol{Y} (\sqrt{\boldsymbol{X}})^+$. □

C.2. Proof of Theorem 6.5

Lemma C.4. *Under Condition 6.1, for all $T > 0, S \ge K$ and $\boldsymbol{w} \in \mathbb{R}^m$, we have*

$$\mathcal{L}(f_T) \le \mathcal{L}(\hat{f}_{\boldsymbol{w}}) + \frac{e^{-\lambda_K T}}{2 - e^{-\lambda_K T}} \|f^*\|_{\mathcal{D}}^2 - (1 - \lambda_{S+1}^2 T^2) \sum_{k \ge S+1} \langle \hat{f}_{\boldsymbol{w}}, e_k \rangle_{\mathcal{D}}^2.$$

Proof of Lemma C.4. By (9),

$$\mathcal{L}(f_T) = \sum_{k \ge 1} \underbrace{\Big(\big(1 - e^{-\lambda_k T}\big) \langle \hat{f}_{\boldsymbol{w}}, e_k \rangle_{\mathcal{D}} - \langle f^*, e_k \rangle_{\mathcal{D}} \Big)^2}_{=: \ell_k}.$$

We provide upper bounds for ℓ_k in three cases. First, for all $1 \le k \le K$, by Lemma C.2 we have

$$\begin{aligned}
\ell_k = \Big(\big(1 - e^{-\lambda_k T}\big) \langle \hat{f}_{\boldsymbol{w}}, e_k \rangle_{\mathcal{D}} - \langle f^*, e_k \rangle_{\mathcal{D}} \Big)^2 &\le \Big(\langle \hat{f}_{\boldsymbol{w}}, e_k \rangle_{\mathcal{D}}^2 - \langle f^*, e_k \rangle_{\mathcal{D}} \Big)^2 + \frac{e^{-\lambda_k T}}{2 - e^{-\lambda_k T}} \langle f^*, e_k \rangle_{\mathcal{D}}^2 \\
&\le \Big(\langle \hat{f}_{\boldsymbol{w}}, e_k \rangle_{\mathcal{D}}^2 - \langle f^*, e_k \rangle_{\mathcal{D}} \Big)^2 + \frac{e^{-\lambda_K T}}{2 - e^{-\lambda_K T}} \langle f^*, e_k \rangle_{\mathcal{D}}^2,
\end{aligned}$$

where we used the fact that $e^{-\lambda_k T} \le e^{-\lambda_K T}$ for $1 \le k \le K$.

Then, for $K + 1 \le k \le S$, we have

$$\ell_k = \Big(\big(1 - e^{-\lambda_k T}\big) \langle \hat{f}_{\boldsymbol{w}}, e_k \rangle_{\mathcal{D}} - \langle f^*, e_k \rangle_{\mathcal{D}} \Big)^2 = \big(1 - e^{-\lambda_k T}\big)^2 \langle \hat{f}_{\boldsymbol{w}}, e_k \rangle_{\mathcal{D}}^2 \le \Big(\langle \hat{f}_{\boldsymbol{w}}, e_k \rangle_{\mathcal{D}} - \langle f^*, e_k \rangle_{\mathcal{D}} \Big)^2.$$

Finally, for $k \ge S + 1$, we have $e^{-\lambda_k T} \ge e^{-\lambda_{S+1} T} \ge 1 - \lambda_{S+1} T$, and thus $\big(1 - e^{-\lambda_k T}\big)^2 \le \lambda_{S+1}^2 T^2$. This implies the following bound for ℓ_k:

$$\ell_k = \Big(\big(1 - e^{-\lambda_k T}\big) \langle \hat{f}_{\boldsymbol{w}}, e_k \rangle_{\mathcal{D}} - \langle f^*, e_k \rangle_{\mathcal{D}} \Big)^2 = (1 - e^{-\lambda_k T})^2 \langle \hat{f}_{\boldsymbol{w}}, e_k \rangle_{\mathcal{D}}^2 \le \lambda_{S+1}^2 T^2 \langle \hat{f}_{\boldsymbol{w}}, e_k \rangle_{\mathcal{D}}^2..$$

Putting all these together proves the following:

$$\begin{aligned}
\mathcal{L}(f_T) &\le \sum_{k=1}^{S} \Big(\langle \hat{f}_{\boldsymbol{w}}, e_k \rangle_{\mathcal{D}}^2 - \langle f^*, e_k \rangle_{\mathcal{D}} \Big)^2 + \frac{e^{-\lambda_K T}}{2 - e^{-\lambda_K T}} \sum_{k=1}^{K} \langle f^*, e_k \rangle_{\mathcal{D}}^2 + \lambda_{S+1}^2 T^2 \sum_{k \ge S+1} \langle \hat{f}_{\boldsymbol{w}}, e_k \rangle_{\mathcal{D}}^2 \\
&= \sum_{k \ge 1} \Big(\langle \hat{f}_{\boldsymbol{w}}, e_k \rangle_{\mathcal{D}}^2 - \langle f^*, e_k \rangle_{\mathcal{D}} \Big)^2 + \frac{e^{-\lambda_K T}}{2 - e^{-\lambda_K T}} \sum_{k=1}^{K} \langle f^*, e_k \rangle_{\mathcal{D}}^2 - (1 - \lambda_{S+1}^2 T^2) \sum_{k \ge S+1} \langle \hat{f}_{\boldsymbol{w}}, e_k \rangle_{\mathcal{D}}^2 \\
&= \mathcal{L}(\hat{f}_{\boldsymbol{w}}) + \frac{e^{-\lambda_K T}}{2 - e^{-\lambda_K T}} \|f^*\|_{\mathcal{D}}^2 - (1 - \lambda_{S+1}^2 T^2) \sum_{k \ge S+1} \langle \hat{f}_{\boldsymbol{w}}, e_k \rangle_{\mathcal{D}}^2,
\end{aligned}$$

which completes the proof. □

It remains to bound $\sum_{k \ge S+1} \langle \hat{f}_{\boldsymbol{w}}, e_k \rangle_{\mathcal{D}}^2$. Let $G : \mathbb{R}^m \to \mathcal{H}$ be the linear operator that maps $\boldsymbol{u} \in \mathbb{R}^m$ to $\sum_{i=1}^m u_i g_i \in \mathcal{H}$. Let $Pf := \sum_{k=1}^S \langle f, e_k \rangle e_k$ be the projection operator onto the top-K eigenspace of $\mathcal{K}$, and Qf be the projection onto the span of $\{g_1, \ldots, g_m\}$. Then $G^* G = \boldsymbol{\Phi}$, $G^* P G = \boldsymbol{A}$ and $Q = G^* \boldsymbol{\Phi}^+ G$.

In the following, we give a characterization of κ.

Lemma C.5. *If $\boldsymbol{A} \neq \mathbf{0}$, then $R := QPQ \neq 0$ and κ equals the smallest non-zero eigenvalue of R.*

Proof. If $\boldsymbol{A} \neq \mathbf{0}$, then there exist $1 \leq i < j \leq m$ such that $\langle Pg_i, Pg_j \rangle \neq 0$. This implies $R \neq 0$ because $\langle g_i, Rg_j \rangle = \langle PQg_i, PQg_j \rangle \neq 0$.

Let $\rho_{\min}$ be the smallest non-zero eigenvalue of R. To see why $\kappa = \rho_{\min}$, we first express $\rho_{\min}$ as the minimum of a Rayleigh quotient:

$$\rho_{\min} = \inf_{\alpha \in \mathrm{range}(R)} \frac{\langle \alpha, R\alpha \rangle}{\langle \alpha, \alpha \rangle}. \tag{16}$$

Then we have

$$\begin{aligned}\mathrm{range}(R) = \mathrm{range}(QPQ) = \{QPG\boldsymbol{v} \mid \boldsymbol{v} \in \mathbb{R}^m\} &= \{GG^*QPG\boldsymbol{v} \mid \boldsymbol{v} \in \mathbb{R}^m\} \\ &= \{GG^*PG\boldsymbol{v} \mid \boldsymbol{v} \in \mathbb{R}^m\}.\end{aligned}$$

Since $\boldsymbol{A} = G^*PG$, we further have

$$\mathrm{range}(R) = \{G\boldsymbol{A}\boldsymbol{v} \mid \boldsymbol{v} \in \mathbb{R}^m\} = \{G\boldsymbol{u} \mid \boldsymbol{u} \in \mathrm{colspan}(\boldsymbol{A})\}.$$

Plugging this into (16), we have

$$\begin{aligned}\rho_{\min} = \inf_{\boldsymbol{u} \in \mathrm{colspan}(\boldsymbol{A})} \frac{\langle G\boldsymbol{u}, RG\boldsymbol{u} \rangle}{\langle G\boldsymbol{u}, G\boldsymbol{u} \rangle} = \inf_{\boldsymbol{u} \in \mathrm{colspan}(\boldsymbol{A})} \frac{\langle \boldsymbol{u}, G^*PG\boldsymbol{u} \rangle}{\langle \boldsymbol{u}, G^*G\boldsymbol{u} \rangle} &= \inf_{\boldsymbol{u} \in \mathrm{colspan}(\boldsymbol{A})} \frac{\boldsymbol{u}^\top \boldsymbol{A}\boldsymbol{u}}{\boldsymbol{u}^\top \boldsymbol{\Phi}\boldsymbol{u}} \\ &= \left(\sup_{\boldsymbol{u} \in \mathrm{colspan}(\boldsymbol{A})} \frac{\boldsymbol{u}^\top \boldsymbol{\Phi}\boldsymbol{u}}{\boldsymbol{u}^\top \boldsymbol{A}\boldsymbol{u}} \right)^{-1}.\end{aligned}$$

Since $\boldsymbol{\Phi} = \boldsymbol{A} + \boldsymbol{B}$, we further have

$$\rho_{\min} = \left(\sup_{\boldsymbol{u} \in \mathrm{colspan}(\boldsymbol{A})} \frac{\boldsymbol{u}^\top (\boldsymbol{A} + \boldsymbol{B})\boldsymbol{u}}{\boldsymbol{u}^\top \boldsymbol{A}\boldsymbol{u}} \right)^{-1} = \left(1 + \sup_{\boldsymbol{u} \in \mathrm{colspan}(\boldsymbol{A})} \frac{\boldsymbol{u}^\top \boldsymbol{B}\boldsymbol{u}}{\boldsymbol{u}^\top \boldsymbol{A}\boldsymbol{u}} \right)^{-1}.$$

By Lemma C.3, the supermum here equals $\lambda_1\left((\sqrt{\boldsymbol{A}})^+ \boldsymbol{B} (\sqrt{\boldsymbol{A}})^+\right)$. Therefore,

$$\rho_{\min} = \frac{1}{1 + \lambda_1\left((\sqrt{\boldsymbol{A}})^+ \boldsymbol{B} (\sqrt{\boldsymbol{A}})^+\right)} = \kappa,$$

which completes the proof. □

Now we give a lower bound for $\sum_{k \geq S+1} \langle \hat{f}_{\boldsymbol{w}}, e_k \rangle_{\mathcal{D}}^2$ in terms of $\mathcal{L}(\hat{f}_{\boldsymbol{w}})$.

Lemma C.6. *Under Conditions 6.1 and 6.3, if $\mathcal{L}(\hat{f}_{\boldsymbol{w}}) = \min_{\boldsymbol{w}' \in \mathbb{R}^m} \mathcal{L}(\hat{f}_{\boldsymbol{w}'})$, then*

$$\sum_{k \geq S+1} \langle \hat{f}_{\boldsymbol{w}}, e_k \rangle_{\mathcal{D}}^2 \geq \kappa \cdot \mathcal{L}(\hat{f}_{\boldsymbol{w}}), \tag{17}$$

where the equality is attained iff $f^* = Ph^*$ for some $h^* \in \mathcal{H}$ that is in the span of the eigenvectors of $R := QPQ$ corresponding to the eigenvalue 1 (if exists) and the smallest non-zero eigenvalue, which is κ.

Proof of Lemma C.6. By Condition 6.3, $f^* \in \mathrm{span}\{Pg_1, \ldots, Pg_m\} = \mathrm{range}(PG) = \mathrm{range}(PQ)$, which implies that there exists $h \in \mathcal{H}$ such that $f^* = PQh$. As $\boldsymbol{w}$ minimizes $\mathcal{L}(\hat{f}_{\boldsymbol{w}})$, $\hat{f}_{\boldsymbol{w}}$ must be the least square solution and $\hat{f}_{\boldsymbol{w}} = Qf^* = QPQh = Rh$. So we have the following identities:

$$\begin{aligned}\sum_{k \geq S+1} \langle \hat{f}_{\boldsymbol{w}}, e_k \rangle_{\mathcal{D}}^2 = \|(I - P)\hat{f}_{\boldsymbol{w}}\|_{\mathcal{D}}^2 = \|(I - P)QPQh\|_{\mathcal{D}}^2 &= \langle h, QPQ(I - P)QPQh \rangle_{\mathcal{D}}, \\ \mathcal{L}(\hat{f}_{\boldsymbol{w}}) = \|f^* - \hat{f}_{\boldsymbol{w}}\|_{\mathcal{D}}^2 = \|PQh - QPQh\|^2 = \|(I - Q)PQh\|^2 &= \langle h, QP(I - Q)PQh \rangle_{\mathcal{D}}.\end{aligned}$$

We can express the linear operators $QPQ(I-P)QPQ$ and $QP(I-Q)PQ$ in terms of R as follows.

$$QPQ(I-P)QPQ = (QPQ)(QPQ) - (QPQ)(QPQ)(QPQ) = R(R-R^2),$$
$$QP(I-Q)PQ = QPQ - (QPQ)(QPQ) = R - R^2.$$

Note that $0 \preceq R \preceq I$ since $\langle f, QPQf\rangle_{\mathcal{D}} = \langle Qf, P(Qf)\rangle_{\mathcal{D}}$ and $0 \le \langle Qf, P(Qf)\rangle_{\mathcal{D}} \le \langle Qf, Qf\rangle_{\mathcal{D}} = \langle f, f\rangle_{\mathcal{D}}$. Let $Rf = \sum_{i=1}^{r} \rho_i \langle v_i, f\rangle v_i$ be an eigendecomposition of R, where $1 \ge \rho_1 \ge \rho_2 \ge \cdots \ge \rho_r > 0$ and $\{v_1, \dots, v_r\}$ are the corresponding eigenvectors.

By Condition 6.3, $\boldsymbol{A} \neq \mathbf{0}$. So by Lemma C.5, $R \neq 0$, $r \ge 1$ and $\rho_r = \kappa$. Then we have

$$\sum_{k \ge S+1} \langle \hat{f}_{\boldsymbol{w}}, e_k\rangle_{\mathcal{D}}^2 = \big\langle h, (R(R-R^2))h\big\rangle_{\mathcal{D}} = \sum_{i=1}^{r} \rho_i \cdot (\rho_i - \rho_i^2)\langle h, v_i\rangle_{\mathcal{D}}^2$$
$$\ge \sum_{i=1}^{r} \kappa \cdot (\rho_i - \rho_i^2)\langle h, v_i\rangle_{\mathcal{D}}^2 = \kappa \cdot \big\langle h, (R-R^2)h\big\rangle_{\mathcal{D}} = \kappa \cdot \mathcal{L}(\hat{f}_{\boldsymbol{w}}),$$

which proves (17).

To understand when the equality is attained, we note that the equality holds iff the inequality in the second line is an equality. This means that for all $1 \le i \le r$, either $\rho_i \in \{1, \kappa\}$ or $\langle h, v_i\rangle_{\mathcal{D}} = 0$. Equivalently, h can be expressed as $h = \mu_1 h_1 + \mu_2 h_2$ for some $\mu_1, \mu_2 \in \mathbb{R}$, where h_1 is an eigenvector of R corresponding to eigenvalue 1 and h_2 is an eigenvector of R corresponding to eigenvalue κ.

In this case, $f^* = PQh = \mu_1 PQh_1 + \mu_2 PQh_2$. This can be further simplified to $f^* = \mu_1 Ph_1 + \mu_2 Ph_2 = Ph$ since $h_1, h_2 \in \mathrm{range}(R) \subseteq \mathrm{range}(Q)$. □

Proof of Theorem 6.5. Putting Lemma C.4, Lemma C.6 and Lemma C.5 together, we have

$$\mathcal{L}(f_T) \le \mathcal{L}(\hat{f}_{\boldsymbol{w}}) + \frac{e^{-\lambda_K T}}{2 - e^{-\lambda_K T}} \|f^*\|_{\mathcal{D}}^2 - (1 - \lambda_{S+1}^2 T^2)\kappa \cdot \mathcal{L}(\hat{f}_{\boldsymbol{w}})$$
$$= \big(1 - (1 - \lambda_{S+1}^2 T^2)\kappa\big)\, \mathcal{L}(\hat{f}_{\boldsymbol{w}}) + \frac{e^{-\lambda_K T}}{2 - e^{-\lambda_K T}} \|f^*\|_{\mathcal{D}}^2,$$

which completes the proof. □

C.3. General Lower Bound on κ_S

We will bound

$$\kappa_S := \frac{1}{1 + \lambda_1\big((\sqrt{\boldsymbol{A}_S})^+ \boldsymbol{B}_S (\sqrt{\boldsymbol{A}_S})^+\big)}.$$

in the general case.

Let F_A and F_B be defined as follows:

$$F_A : \mathrm{span}(e_1, \dots, e_J) \to \mathbb{R}^m$$
$$F_A : h \mapsto \begin{pmatrix} \vdots \\ \langle h, g_j\rangle_{\mathcal{D}} \\ \vdots \end{pmatrix}$$
$$F_B : \mathrm{span}(e_J, e_{J+1}, \dots) \to \mathbb{R}^m$$
$$F_B : h \mapsto \begin{pmatrix} \vdots \\ \langle h, g_j\rangle_{\mathcal{D}} \\ \vdots \end{pmatrix}.$$

We will denote $F_A^\dagger$ and $F_B^\dagger$ their adjoint operators. We can rewrite the largest eigenvalue of $(\sqrt{\boldsymbol{A}_S})^+\boldsymbol{B}_S(\sqrt{\boldsymbol{A}_S})^+$ as

$$\lambda_1\left((\sqrt{\boldsymbol{A}_S})^+\boldsymbol{B}_S(\sqrt{\boldsymbol{A}_S})^+\right) = \sup_{v\in\text{span}(e_1,\dots,e_J)} \frac{v^T F_A^\dagger F_B F_B^\dagger F_A v}{v^T F_A^\dagger F_A F_A^\dagger F_A v}$$

Let $\mathcal{H}$ be a Hilbert space defined as

$$\mathcal{H} = \{\text{linear mappings from span}(e_1,\dots,e_J) \text{ to span}(e_J, e_{J+1},\dots)\}$$

with the inner product defined as

$$\langle f, f'\rangle_{\mathcal{H}} = \sum_{j=1}^{J}\langle f(e_i), f'(e_i)\rangle_{\mathcal{D}}.$$

We will show the following lower bound first.

Proposition C.7. *For $v \in$ span$(e_1,\dots,e_J)$ it holds that*

$$v^T F_A^\dagger F_A F_A^\dagger F_A v \geq \lambda_J(\boldsymbol{A})^2\|v\|_2^2.$$

Proof. Let $\sum_{i=1}^J u_i e_i = F_A^\dagger F_A v$. Then we have that

$$v^T F_A^\dagger F_A F_A^\dagger F_A v \geq \langle\sum_{i=1}^J u_i e_i, \sum_{i=1}^J u_i e_i\rangle_{\mathcal{D}} \geq \|u\|_2^2 \geq \lambda_J(\boldsymbol{A})^2\|v\|_2^2.$$

□

For the upper bound, we will use the concentration of $\|F_B^\dagger F_A v\|_{\mathcal{D}}^2$.

Proposition C.8. *For $v = \sum_{j=1}^J v_i e_i$ it holds that*

$$\|F_B^\dagger F_A v\|_{\mathcal{D}}^2 \leq \left(\sum_{j=1}^J \|F_B^\dagger F_A e_i\|_{\mathcal{D}}^2\right)\left(\sum_{j=1}^J v_i^2\right).$$

Proof. This follows immediately from Cauchy-Schwarz inequality. □

Proposition C.9. *Let $Y \in \mathcal{H}$ be such that $\langle e_j, Y(v)\rangle = \langle e_j, g\rangle_{\mathcal{D}}\langle g, v\rangle_{\mathcal{D}}$*

$$\|Y\|_H^2 = \sum_{j=1}^J\sum_{k=J+1}^\infty \langle e_j, g\rangle_{\mathcal{D}}^2\langle g, e_k\rangle_{\mathcal{D}}^2$$

so $\|Y\|_H^2 = \|P_{\leq J}g\|^2\|P_{\geq J+1}g\|^2$, that is $\|Y\|_H = \|P_{\leq J}g\|_{\mathcal{D}}\|P_{\geq J+1}g\|_{\mathcal{D}}$.

Furthermore, let Y_i be the r.v. corresponding to the draw of the i-th feature g_i. Then

$$F_B^\dagger F_A = \sum_{i=1}^m Y_i.$$

We have the following concentration bound in a Hilbert space.

Theorem C.10 ((Pinelis, 1994))**.** *Let Y_i be independent $\mathcal{H}$ valued random variable with $\mathbb{E}(Y_i) = 0$, where $\mathcal{H}$ is a seperable Hilbert space. Assume that for every $q \geq 2, q \in \mathbb{N}$ we have that $\mathbb{E}\left(\|Y_i\|_{\mathcal{H}}^q\right) \leq \frac{1}{2}q!B^2L^{q-2}$. Then, for all $n \in \mathbb{N}$ and $\epsilon > 0$ we have*

$$P\left(\left\|\frac{1}{n}\sum_{i=1}^n Y_i\right\|_{\mathcal{H}} > \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{B^2 + L\epsilon + B\sqrt{B^2+2L\epsilon}}\right).$$

Putting these together in the bound for $\lambda_1\left((\sqrt{\boldsymbol{A}_S})^+\boldsymbol{B}_S(\sqrt{\boldsymbol{A}_S})^+\right)$, we get the following.

Proposition C.11. *The largest eigenvalue of $(\sqrt{\boldsymbol{A}_S})^+\boldsymbol{B}_S(\sqrt{\boldsymbol{A}_S})^+$ is bounded by*

$$\lambda_1\left((\sqrt{\boldsymbol{A}_S})^+\boldsymbol{B}_S(\sqrt{\boldsymbol{A}_S})^+\right) \leq \frac{\sum_{j=1}^J \|F_B^+F_Ae_j\|_{\mathcal{D}}^2}{\lambda_J(\boldsymbol{A})^2}$$

Proof of Proposition C.11. This follows directly from Proposition C.8 and Proposition C.7. □

We can now use this to bound the norm of $F_B^\dagger F_A$.

Proposition C.12 (Hilbert Space Concentration)**.** *With probability at least $1-2\delta$*

$$\|F_B^\dagger F_A\|_{\mathcal{H}}^2 \leq mB\left(\log\frac{1}{\delta}\right)^2$$

where B is as defined in Theorem C.10.

Proof of Proposition C.12. In Theorem C.10, taking $m^2\epsilon^2 = mB^2\log^2\frac{1}{\delta}$, we get the desired result. □

We can lower bound the eigenvalues of $\boldsymbol{A}$ using matrix concentration results.

Theorem C.13. *Let $\boldsymbol{a} = \begin{pmatrix}\vdots\\ \langle g, e_k\rangle \\ \vdots\end{pmatrix}_{k=1}^J$ as $g\sim\mathcal{G}$ be a row of $\boldsymbol{A}$. Depending on whether the rows of $\boldsymbol{A}$ are subgaussian or bounded a.s. we have the following two bounds:*

1. *Under Model B.1, i.e. with $D_g = N(0,\Lambda)$ where $\Lambda = diag(\lambda_1,\lambda_2,\dots)$, then with probability $1-2\exp(-ct_A^2)$ it holds that*

$$\lambda_J(\boldsymbol{A}) \geq \lambda_J\left(\sqrt{m} - C\sqrt{J} - t_A\right)^2,$$

   if $\sqrt{m}\geq C\sqrt{J}+t_A$, where $C=\Theta(L_g^2)$ and $c=\Theta(1/L_g^4)$ where L_g is the subgaussian norm of the vector $\boldsymbol{a}$. In this case $L_g\leq C_g\lambda_1$, where C_g is an absolute constant.

2. *Under Model 2.2, i.e. if $g=\sigma(\langle w,x\rangle)$ with $w\sim unif(S^{d-1})$ and $N_0+N_1+\cdots+N_{s-1}+1\leq J\leq N_0+N_1+\cdots+N_s$ and s is 1 or even, we have that with probability $1-2J\exp(-ct_A^2)$*

$$\lambda_J(\boldsymbol{A}) \geq \sigma_s^2\left(\sqrt{m} - t_A\sqrt{J}\right)^2,$$

   if $\sqrt{m}\geq t_A\sqrt{J}$, where c is an absolute constant.

Under Model E.1, with probability $1-2J\exp(-ct_A^2)$ we have that

$$\lambda_J(\boldsymbol{A}) \geq \sigma_s^2\left(\sqrt{m} - t_A\sqrt{J}\right)^2,$$

This computation allows us to lower bound κ_S in the cases of Model 2.2 in Appendix F and Model 2.3 in Model B.1.

Proof of Theorem C.13. We will use Theorems 5.39 and 5.41 in Vershynin (2012). These theorems say that if the rows of a matrix are either bounded a.s. or sub-gaussian, then the singular values of this matrix are upper and lower bounded. Note first that in both cases, $\boldsymbol{A}$ is a PSD matrix, so it's sufficient to consider the singular values of $\boldsymbol{A}$. In both cases, note that in order to apply the theorems from Vershynin (2012) to a matrix $\boldsymbol{X}$, we need to have $\Sigma(\boldsymbol{X}) = \boldsymbol{I}$, i.e. the rows of $\boldsymbol{X}$ have to be isotropic random variables. Note that in both cases

$$\Sigma(\boldsymbol{A})_{ij} = \mathbb{E}_{\mathcal{G}}\left(\langle g,e_i\rangle\langle g,e_j\rangle\right).$$

If we let $\Sigma = \Sigma(\boldsymbol{A})$, we can apply these theorems to $\Sigma^{-\frac{1}{2}}\boldsymbol{G}$, where $\boldsymbol{G}$ is such that $\boldsymbol{A} = \boldsymbol{G}^T\boldsymbol{G}$. In that case, the theorems say that with corresponding probabilities that

$$s_J(\Sigma^{-\frac{1}{2}}\boldsymbol{G}) \geq (\sqrt{m} - c\sqrt{J} - t_A)^2$$
$$s_J(\Sigma^{-\frac{1}{2}}\boldsymbol{G}) \geq (\sqrt{m} - t_A\sqrt{J})^2,$$

respectively, where s_J is the J-th singular value of a matrix. Note that we have that

$$s_J(\boldsymbol{G}) \geq s_{\min}(\Sigma^{\frac{1}{2}})s_J(\Sigma^{-\frac{1}{2}}\boldsymbol{G}).$$

Therefore, we get that with corresponding probabilities

$$\lambda_J(\boldsymbol{A}) \geq \lambda_J(\Sigma)\left(s_J(\Sigma^{-\frac{1}{2}}\boldsymbol{G})\right)^2.$$

Here, we need lower bounds on m so that we can square the inequality and go from singular values to eigenvalues.

In the case of Model 2.3, we have that

$$\Sigma(\boldsymbol{A})_{ij} = \Lambda.$$

Therefore, we have that $\lambda_J(\Sigma) = \lambda_J$. Note that we are only left with verifying the conditions of the theorem. First is that the rows of $\Sigma^{-\frac{1}{2}}\boldsymbol{A}$ are isotropic which is true by design. Second, we need that the rows of $\Sigma^{-\frac{1}{2}}\boldsymbol{A}$ are subgaussian with a subgaussian constant L_g. Note note that if $X \sim N(0, \sigma)$ then $\|X\|_{\psi_2} \leq C\sigma$ for some absolute constant C. Note also that $\|\boldsymbol{a}\|_{\psi_2} = \sup_{x\in S^{J-1}} \|x^T\boldsymbol{a}\|_{\psi_2} \leq \sqrt{C_1\sum_{i=1}^J x_i^2\|\boldsymbol{a}_i^2\|_{\psi_2}^2} \leq \sqrt{C_1C\sum_{i=1}^J x_i^2\lambda_i^2} \leq \sqrt{C_1C\lambda_1^2\sum_{i=1}^J x_i^2} = \sqrt{C_1C\lambda_1^2} = C_g\lambda_1$. Note that therefore under Model 2.3, the conditions of the Theorem 5.39 in (Vershynin, 2012) are satisfied, so we get that with probability $1 - 2\exp(-ct_A^2)$

$$\lambda_J(\boldsymbol{A}) \geq \lambda_J\left(\sqrt{m} - C\sqrt{J} - t_A\right)^2,$$

as long as $\sqrt{m} \geq c\sqrt{J} + t_A$. Here $C = \Theta(L_g^2)$ and $c = \Theta(\frac{1}{L_g^4})$ in the sense that both C and c are upper and lower bounded by an absolute constant times the appropriate expression of L_g.

In the case of Model 2.2, note that since the target Condition 6.1 will not have anything in the directions with zero σ_k, we can just ignore and reindex the basis functions. So, for now we will just write as if they are nonzero, but we will keep in mind that we skip the zero σ_k. Therefore, we have that

$$\Sigma(\boldsymbol{A})_{i,j} = \text{diag}(\sigma_1^2, \sigma_1^2, \dots, \sigma_2^2, \dots, \sigma_4^2, \dots),$$

where each σ_i repeats with the multiplicity N_i for $i = 1$ or i even (otherwise it's 0). Furthermore, note that $\|\boldsymbol{a}\|_2^2 = J$, so we can indeed apply Theorem 5.41 from (Vershynin, 2012). Therefore, we have that with probability at least $1 - 2J\exp(-ct_A^2)$

$$\lambda_J(\boldsymbol{A}) \geq \sigma_s^2(\sqrt{m} - t_A\sqrt{J})^2$$

where c is an absolute constant. □

D. Weak-to-Strong Bound Applied to the Case of Linear Networks

Proof of Theorem 3.2. Follows from Theorem D.5. □

Proof of Theorem E.4. This follows directly from Proposition B.2 and the general bound Theorem D.1. □

Theorem D.1 (Weak-to-Strong Generalization with Gaussian Feature Distribution)**.** *Under Condition 6.1, if Model B.1 holds for some K, and if the teacher attains minimum loss $\mathcal{L}_{\text{TE}} = \hat{\mathcal{L}}_{\min}$, then for the student trained until time T we have with probability at least $1 - \frac{4}{m}$ that*

$$\mathcal{L}_{\text{ST}} \leq \inf_{S\geq K}\left\{\frac{\left(\sum_{i=1}^S \lambda_i\right)\left(\sum_{i=S+1}^\infty \lambda_i\right)}{\lambda_S^2}\frac{(\log m)^2}{m}\mathcal{L}_{\text{TE}} + \lambda_{S+1}T^2\mathcal{L}_{\text{TE}}\right\} + \frac{e^{-\lambda_K T}}{2 - e^{-\lambda_K T}}\left\|f^*\right\|_{\mathcal{D}}^2.$$

Proof of Theorem D.1. This follows from Theorem 6.5 using Theorem D.4 and noting that we can replace $\lambda_{S+1}T^2(1 - \kappa_S) \leq \lambda_{S+1}T^2$. □

D.1. Proofs Related to The Case of Linear Networks

We first aim to show that the k-th moments of $\|Y\|_{\mathcal{H}} = \|P_{\leq J} g\|_{\mathcal{D}} \|P_{\geq J+1} g\|_{\mathcal{D}}$ are bounded.

Proposition D.2. *Let $X = \sum_{i=1}^n a_i^2 X_i^2$ where X_i are i.i.d. standard Gaussian. Then for $k \geq 2$*

$$\mathbb{E}\left(X^k\right) \leq k! \left(\sum_{i=1}^m a_i^2\right)^k.$$

Proof. Expanding the expression for $X^k = \left(\sum_{i=1}^n a_i^2 X_i^2\right)^k$ and using linearity of expectation, we have that

$$\mathbb{E}\left(X^k\right) = \sum_{|\alpha|=k} \frac{k!}{\alpha!} \prod_{i=1}^m \left((a_i)^2\right)^{\alpha_i} \mathbb{E}\left(\left(X_i^2\right)^{\alpha_i}\right).$$

Note that $\mathbb{E}(X_i^{2\alpha_i}) = (2\alpha_i - 1)!! \leq 2^{\alpha_i}(\alpha_i)!$, so we have that

$$\begin{aligned}
\mathbb{E}\left(X^k\right) &= \sum_{|\alpha|=k} \frac{k!}{\alpha!} \prod_{i=1}^m \left((a_i)^2\right)^{\alpha_i} \mathbb{E}\left(\left(X_i^2\right)^{\alpha_i}\right) \leq \sum_{|\alpha|=k} \frac{k!}{\alpha!} \prod_{i=1}^m \left((a_i)^2\right)^{\alpha_i} 2^{\alpha_i}(\alpha_i)! \\
&= \sum_{|\alpha|=k} \frac{k!}{\alpha!} \alpha! \prod_{i=1}^m \left(2(a_i)^2\right)^{\alpha_i} = k! \sum_{|\alpha|=k} \frac{\alpha!}{\alpha!} \prod_{i=1}^m \left(2(a_i)^2\right)^{\alpha_i} \\
&\leq k! \sum_{|\alpha|=k} \frac{k!}{\alpha!} \prod_{i=1}^m \left(2(a_i)^2\right)^{\alpha_i}
\end{aligned}$$

where the last inequality is true because $\alpha! \leq k!$. Note that $\sum_{|\alpha|=k} \frac{k!}{\alpha!} \prod_{i=1}^m \left(2(a_i)^2\right)^{\alpha_i} = \left(\sum_{i=1}^m 2a_i^2\right)^k$, so the final bound is

$$\mathbb{E}\left(X^k\right) \leq k! \sum_{|\alpha|=k} \frac{k!}{\alpha!} \prod_{i=1}^m \left(2(a_i)^2\right)^{\alpha_i} = k! \left(\sum_{i=1}^m 2a_i^2\right)^k.$$

□

This implies that $\|Y\|_{\mathcal{H}}^k$ is bounded for all $k \geq 2$ with an expression of the required form.

Proposition D.3. *Let $g \sim \mathcal{G}$ with $\langle g, e_i \rangle \sim N(0, \Lambda)$ with $\Lambda = diag(\lambda_1, \lambda_2, \dots)$. Then for Y it holds that for all $k \geq 2$*

$$\mathbb{E}\left(\|Y\|_{\mathcal{H}}^k\right) \leq k! \left(\sqrt{\left(2\sum_{i=1}^J \lambda_i\right)\left(2\sum_{i=J+1}^\infty \lambda_i\right)}\right)^k.$$

Therefore, in this case we can take $B = \sqrt{2\left(2\sum_{i=1}^J \lambda_i\right)\left(2\sum_{i=J+1}^\infty \lambda_i\right)}$ and $L = \sqrt{\left(2\sum_{i=1}^J \lambda_i\right)\left(2\sum_{i=J+1}^\infty \lambda_i\right)}$.

Note that Proposition C.12 applies in this case as well. Additionally, Theorem C.13 applies in this case.

Theorem D.4. *Let $g \sim \mathcal{G}$ with $\langle g, e_i \rangle \sim N(0, \Lambda)$ with $\Lambda = diag(\lambda_1, \lambda_2, \dots)$. Then with probability at least $1 - 2\delta - 2\exp(-c_1 \frac{1}{\lambda_1^4} t_A^2)$ we have that*

$$\lambda_1\left((\sqrt{\boldsymbol{A}_S})^+ \boldsymbol{B}_S (\sqrt{\boldsymbol{A}_S})^+\right) \leq \frac{\|F_B^\dagger F_A\|_{\mathcal{H}}^2}{\lambda_S(\boldsymbol{A})^2} \leq \frac{8m\left(\sum_{i=1}^S \lambda_i\right)\left(\sum_{i=S+1}^\infty \lambda_i\right)\left(\log\frac{1}{\delta}\right)^2}{\lambda_S^2\left(\sqrt{m} - C\sqrt{S} - t_A\right)^4}$$

where c_1 and C are absolute constants. Setting $t_A = \lambda_1^2 \log \frac{1}{\delta_1}$, $\delta = \frac{1}{m}, \delta_1 = \frac{1}{m}$, we get that with probability at least $1 - \frac{4}{m}$ we have

$$\lambda_1\left((\sqrt{\boldsymbol{A}_S})^+ \boldsymbol{B}_S (\sqrt{\boldsymbol{A}_S})^+\right) \le \frac{8m\left(\sum_{i=1}^S \lambda_i\right)\left(\sum_{i=S+1}^\infty \lambda_i\right)(\log m)^2}{\lambda_S^2\left(\sqrt{m} - C\sqrt{S} - \lambda_1^2 \log m\right)^4}.$$

If additionally $m = \omega(S)$ and $m = \omega((\lambda_1 \log m)^2)$ then with probaility $1 - \frac{4}{m}$

$$\lambda_1\left((\sqrt{\boldsymbol{A}_S})^+ \boldsymbol{B}_S (\sqrt{\boldsymbol{A}_S})^+\right) \le 8\frac{\left(\sum_{i=1}^S \lambda_i\right)\left(\sum_{i=S+1}^\infty \lambda_i\right)}{\lambda_S^2}\frac{(\log m)^2}{m}.$$

Proof of Theorem D.4. This theorems follows directly by taking the approproate lower bound from Theorem C.13 and combining that with the Hilbert concentration bound Proposition C.12. □

D.2. Additional Proofs for Linear Networks

Proof of Theorem I.4. This follows from Theorem D.6 by taking $\epsilon_d = 0.01$, $\epsilon_m = 0.01$, $m \ge \Theta(\alpha^{\frac{2}{1-\epsilon_d}})$ and $\delta = 0.01$. Note that since $m = \sqrt{\frac{d-1}{\alpha}}$ we can turn this inequality into an inequality with α and d, i.e $\alpha \ge \Theta(\frac{1}{d^{0.49}})$. The last claim also follows from Theorem D.6, since as $m \to \infty$ the lower bound converges to $0.99\frac{\alpha}{1+\alpha}$ and the probability to 1. □

Theorem D.5 (Weak-to-Strong Generalization with Linear Network)**.** *Consider Model 2.3 when f^* is supported by the first k coordinates, i.e. $f^* = \beta \boldsymbol{x}$, where $\beta_i = 0$ for $i > k$ and take f^* to be norm 1. Take $\Psi_m = (1, \dots, 1, \frac{1}{m}, \dots, \frac{1}{m})$, where 1 is repeated k times and $\frac{1}{m}$ is repeated $N_1(m) = \alpha m^2$ times, with $\alpha = m^{-\eta}$ for any $\eta \in [0, \frac{1}{2}]$ (so $d = k + N_1(m)$). Then, for all $m > C\sqrt{k}$ [‖], if the student is trained until time $T = \log m + \log \frac{1}{\mathcal{L}_{\text{TE}}}$ with probability at least 0.99,*

$$\mathcal{L}_{\text{TE}} \ge \frac{\frac{1}{k}m^{-\eta}}{1 + \frac{1}{k}m^{-\eta}}$$

$$\mathcal{L}_{\text{ST}} \le k\frac{\log^2 m}{m^\eta}\mathcal{L}_{\text{TE}} + \frac{1}{m}(\log m + \log\frac{1}{\mathcal{L}_{\text{TE}}})^2\mathcal{L}_{\text{TE}} + \frac{1}{m}\mathcal{L}_{\text{TE}}$$

So, for all m we have with probability at least 0.99,

$$\mathcal{L}_{\text{ST}} \le \tilde{O}(k\mathcal{L}_{\text{TE}}^2).$$

In particular $\text{PGR} \to 1$ *as* $M_{\text{TE}} \to \infty$.

Proof of Theorem D.5. For $k = 1$ this follows from Theorem D.6. Consider k copies of Theorem D.6 with $N(m) = \frac{\alpha}{k}m^2$ each, $\Psi_{m,1}, \dots, \Psi_{m,k}$. For example, $\Psi_{m,1} = (1, 0, \dots, 0, \frac{1}{m}, \dots, \frac{1}{m}, 0, \dots, 0)$, were we have $N(m) = \frac{N_1(m)}{k}$ repeating $\frac{1}{m}$. Note that a set of weights $\boldsymbol{w} \in \mathbb{R}^m$ that the teacher chooses to optimize over all k problems is worse than choosing a separate set of weights for each of the problems, so the loss will be greater than the sum of the losses in each of the subproblems. So if $\mathcal{L}_{\text{TE}}^i$ is the loss in the i-th instance, we have that $\mathcal{L}_{\text{TE}} \ge \sum_i \mathcal{L}_{\text{TE}}^i$. Let B_j^2 be the norm of f^* in the directions of the basis functions corrresponding to the j-th subset of Ψ_m, i.e. $B_j^2 = \sum_{i \in \Psi_{m,j}} \beta_i^2$. Note that by the normalization of f^*, $\sum_{j=1}^k B_j^2 = 1$. Furthermore, note that our setup is scale homogeneous, i.e. if we scale $\|f^*\|_{\mathcal{D}}^2$ by c^2 then the loss will scale by c^2 as well. Therefore, we can apply Theorem D.6 to each of the k subprolems to get with probability at least $1 - 2\exp(-\frac{1}{4}(\alpha m^2)^{\epsilon_d}) - 2\exp(-\frac{1}{4}m^{\epsilon_m})$ that

$$\mathcal{L}_{\text{TE}}^j \ge B_j^2\left(0.99\frac{\alpha}{k}\frac{1}{\frac{\alpha}{k}+1} - 2m^{-1/2+\epsilon_m} - (\alpha m^2)^{-1/2+\epsilon_d}\alpha\right) \ge 0.98B_j^2\frac{\frac{m^{-\eta}}{k}}{\frac{m^{-\eta}}{k}+1}.$$

Therefore, we have that $\mathcal{L}_{\text{TE}} \ge \sum_{j=1}^k 0.98B_j^2\frac{\frac{m^{-\eta}}{k}}{\frac{m^{-\eta}}{k}+1} = 0.98\frac{\frac{m^{-\eta}}{k}}{\frac{m^{-\eta}}{k}+1}$.

[‖] here C is an absolute constant given by Theorem D.4

Now consider Theorem D.1 in this case. Set $S = k$. Then we get for $\delta_T = \frac{1}{m}\mathcal{L}_{\text{TE}}$ the upper bound since the sums of ψ_i are $\sum_{i=1}^{k}\psi_i = k$ and $\sum_{i=k+1}^{d}\psi_i = \alpha m$ respectively with probability at least $1 - \frac{1}{100}\frac{1}{m^4} - 2\exp(-c_1 m^{1/4})$

$$\mathcal{L}_{\text{ST}} \leq 40k\alpha \log^2 m \mathcal{L}_{\text{TE}} + \frac{1}{m}(\log m + \log\frac{1}{\mathcal{L}_{\text{TE}}})^2\mathcal{L}_{\text{TE}} + \frac{1}{m}\mathcal{L}_{\text{TE}}$$

where the constant 40 comes from the constant 8 from the upper bound and the constant from $\log\frac{1}{\delta}$ by setting $\delta = 100m^4$. Note now that since $\mathcal{L}_{\text{TE}}$ is lower bounded by $\frac{m^{-\eta}}{2k}$, we can combine $\log\frac{1}{\mathcal{L}_{\text{TE}}}$ and $\log m$ terms into

$$\mathcal{L}_{\text{ST}} \leq 50k\frac{\log^2 m}{m^\eta}\mathcal{L}_{\text{TE}}.$$

Here we bounded $(\log m + \log 2k)^2 \leq 2\log m$. This exactly gives with probability at least $1 - \frac{1}{10}\frac{1}{m^4} - 2\exp(-c_1 m^{1/4}) - 2\exp(-\frac{1}{4}(\alpha m^2)^{\epsilon_d}) - 2\exp(-\frac{1}{4}m^{\epsilon_m})$

$$\mathcal{L}_{\text{ST}} \leq \tilde{O}(k\mathcal{L}_{\text{TE}}^2).$$

By setting $\epsilon_d < \frac{1}{8}$ and $\epsilon_m < \frac{1}{8}$, we get that by union bound over all m, with probability at least 0.99 for all m

$$\mathcal{L}_{\text{ST}} \leq \tilde{O}(k\mathcal{L}_{\text{TE}}^2).$$

□

In Theorem D.6 we will establish a case where we can provably lower bound the teacher loss.

Theorem D.6 (Lower bound for Teacher Loss)**.** *In Condition 6.1 take $f^* = e_1$ and in Model 2.3 take $\langle g, e_i\rangle \sim N(0, \Lambda_m)$, where $\Lambda_m = diag(1, \frac{1}{m}, \frac{1}{m}, \ldots, \frac{1}{m})$, where $\frac{1}{m}$ is repeated $N(m)$ times. If $N(m) = \alpha m^2$ then for all $\epsilon_d, \epsilon_m \in [0, \frac{1}{2}]$ with probability $1 - 2\exp(-\frac{1}{4}(\alpha m)^{\epsilon_d}) - 2\exp(-\frac{1}{4}m^{\epsilon_m})$ we have $\mathcal{L}_{\text{TE}} \geq 0.99\frac{\alpha}{1+\alpha} - 2m^{-\frac{1}{2}+\epsilon_m} - (\alpha m)^{-\frac{1}{2}+\epsilon_d}\alpha$. Furthermore, as long as $\alpha m\left(\log\frac{1}{\delta}\right)^2 = \Omega(m^{-2+\epsilon_s})$ for some $\epsilon_s > 0$, $1 - \delta - 2\exp(-c_1 m^{\frac{1}{4}})$, we have that $\mathcal{L}_{\text{ST}} \leq 8\alpha m(\log(\frac{1}{\delta}))^2\mathcal{L}_{\text{TE}}\,(1 + O(m^{-\epsilon_s}))$.*

Lemma D.7. *Let $\boldsymbol{G}_{ij} = \sum_{i=2}^{\infty}\langle g_j, e_i\rangle\langle g_l, e_i\rangle$, let $\boldsymbol{H}_{i,k} = \langle g_i, e_k\rangle \in \mathbb{R}^{m\times d}$, and let $\boldsymbol{g} = \begin{pmatrix}\vdots\\ \langle g_j, e_1\rangle \\ \vdots\end{pmatrix}_{j=1}^{m} \sim N(0, \psi\boldsymbol{I}_m)$ be the rows of $\boldsymbol{H}$, i.e. $\Lambda = \psi\boldsymbol{I}_m$. Then with probability at least $1 - 2\exp(-\frac{d\delta^2}{4})$ for all $\boldsymbol{v} \in \mathbb{R}^m$*

$$\boldsymbol{v}^T\boldsymbol{G}\boldsymbol{v} \geq (1-\delta)tr(\Lambda)\boldsymbol{v}^T\boldsymbol{v}.$$

Proof of Lemma D.7. Note that $\boldsymbol{G} = \boldsymbol{H}\boldsymbol{H}^T$ so we have that $\boldsymbol{v}^T\boldsymbol{G}\boldsymbol{v} = \|\boldsymbol{H}^T\boldsymbol{v}\|_2^2$. Note further that

$$\boldsymbol{H}^T\boldsymbol{v} = \begin{pmatrix}\vdots\\ \sum_{i=1}^{m} v_i\langle g_i, e_j\rangle \\ \vdots\end{pmatrix}_{j=1}^{d} \sim N(0, \psi\boldsymbol{I}_d\|v\|_2^2),$$

since the entries $\sum_{i=1}^{m} v_i\langle g_i, e_j\rangle$ are independent and each has variance $\psi\|v\|_2^2$.

Note now that

$$\|\boldsymbol{H}^T\boldsymbol{v}\|_2^2 = \sum_{j=1}^{d}\left(\sum_{i=1}^{m} v_i\langle g_i, e_j\rangle\right)^2 \sim \psi\|\boldsymbol{v}\|_2^2\left(Z_1^2 + \cdots + Z_d^2\right),$$

where $Z_i \sim N(0,1)$ are iid, i.e. $\|\boldsymbol{H}^T\boldsymbol{v}\|_2^2 \sim \psi\|\boldsymbol{v}\|_2^2 X$ where $X \sim \chi_d^2$. By concentation of χ_d^2 distribution, we have that with probability at least $1 - 2\exp(-\frac{d\delta^2}{4})$ that

$$\boldsymbol{v}^T\boldsymbol{G}\boldsymbol{v} = \|\boldsymbol{H}^T\boldsymbol{v}\|_2^2 \geq (1-\delta)\psi d\boldsymbol{v}^T\boldsymbol{v} = (1-\delta)tr(\Lambda)\boldsymbol{v}^T\boldsymbol{v}.$$

□

Proof of Theorem D.6. Let $\hat{f}_{\hat{\boldsymbol{w}}}$ be the teacher predictor. Note that we can rewrite this as

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\left(f^* - \hat{f}_{\boldsymbol{w}}\right)^2 &= \mathbb{E}_{\mathcal{D}}\left(e_1(x)^2\left(1 - \sum_{j=1}^m w_j\langle g_j, e_1\rangle\right)^2\right) + \mathbb{E}_{\mathcal{D}}\left(\left(\sum_{j=1}^m w_j \sum_{i=2}^\infty \langle g_j, e_i\rangle e_i(x)\right)^2\right) \\
&= 1 - 2\sum_{j=1}^m w_j\langle g_j, e_1\rangle + \left(\sum_{j=1}^m w_j\langle g_j, e_1\rangle\right)^2 + \mathbb{E}_{\mathcal{D}}\left(\left(\sum_{i=1}^\infty \sum_{j=1}^m w_j\langle g_j, e_i\rangle e_i(x)\right)^2\right) \\
&= 1 - 2\sum_{j=1}^m w_j\langle g_j, e_1\rangle + \left(\sum_{j=1}^m w_j\langle g_j, e_1\rangle\right)^2 + \sum_{i=2}^\infty \mathbb{E}_{\mathcal{D}}\left(\left(\sum_{j=1}^m w_j\langle g_j, e_i\rangle\right)^2\right) \\
&= 1 - 2\sum_{j=1}^m w_j\langle g_j, e_1\rangle + \left(\sum_{j=1}^m w_j\langle g_j, e_1\rangle\right)^2 + \sum_{i=2}^\infty \left(\sum_{j=1}^m w_j\langle g_j, e_i\rangle\right)^2
\end{aligned}$$

Let $[\boldsymbol{G}]_{jl} = \sum_{i=2}^\infty \langle g_j, e_i\rangle\langle g_l, e_i\rangle$ and $\boldsymbol{g} = \begin{pmatrix} \vdots \\ \langle g_j, e_1\rangle \\ \vdots \end{pmatrix}_{j=1}^m$.

Now, we can rewrite the above equation as follows.

$$\begin{aligned}
\mathcal{L}(\hat{f}_{\boldsymbol{w}}) &= 1 - 2\sum_{j=1}^m w_j\langle g_j, e_1\rangle + \left(\sum_{j=1}^m w_j\langle g_j, e_1\rangle\right)^2 + \sum_{i=2}^\infty \left(\sum_{j=1}^m w_j\langle g_j, e_i\rangle\right)^2 \\
&= 1 - 2\boldsymbol{w}^T\boldsymbol{g} + (\boldsymbol{w}^T\boldsymbol{g})^2 + \boldsymbol{w}^T\boldsymbol{G}\boldsymbol{w}.
\end{aligned}$$

By Lemma D.7, we have that with probability at least $1 - 2\exp(-\frac{d\delta^2}{4})$ that

$$\boldsymbol{w}^T\boldsymbol{G}\boldsymbol{w} \geq (1-\delta)tr(\Lambda)\boldsymbol{w}^T\boldsymbol{w},$$

where $d = N(m)$. Therefore, we have that

$$\mathcal{L}(\hat{f}_{\boldsymbol{w}}) = 1 - 2\boldsymbol{w}^T\boldsymbol{g} + (\boldsymbol{w}^T\boldsymbol{g})^2 + \boldsymbol{w}^T\boldsymbol{G}\boldsymbol{w} \geq 1 - 2\boldsymbol{w}^T\boldsymbol{g} + (\boldsymbol{w}^T\boldsymbol{g})^2 + (1-\delta)tr(\Lambda)\boldsymbol{w}^T\boldsymbol{w}.$$

The last expression is minimized for $\boldsymbol{w} = \frac{1}{\|\boldsymbol{g}\|_2^2 + (1-\delta)tr(\Lambda)}\boldsymbol{g}$. Note that we have $tr(\Lambda) = \alpha m$. Note further that $\|\boldsymbol{g}\|_2^2 \sim \chi^2_m$, so it concetrates around m, i.e. with probability at least $1 - 2\exp(-\frac{\delta_m^2 m}{4})$ we have

$$(1+\delta_m)m \geq \|\boldsymbol{g}\|_2^2 \geq (1-\delta_m)m.$$

So overall, we have that with probability at least $1 - 2\exp(-\frac{d\delta^2}{4}) - 2\exp(-\frac{m\delta_m^2}{4})$

$$\begin{aligned}
\mathcal{L}(\hat{f}_{\boldsymbol{w}}) &\geq \left(1 - \frac{\|\boldsymbol{g}\|_2^2}{\|\boldsymbol{g}\|_2^2 + (1-\delta)tr(\Lambda)}\right)^2 + (1-\delta)tr(\Lambda)\left(\frac{1}{\|\boldsymbol{g}\|_2^2 + (1-\delta)tr(\Lambda)}\right)^2\|\boldsymbol{g}\|_2^2 \\
&\geq \left(1 - \frac{m(1+\delta_m)}{m(1+\delta_m) + (1-\delta)\alpha m}\right)^2 + (1-\delta)\alpha\left(\frac{1}{(1+\delta_m) + (1-\delta)\alpha}\right)^2(1-\delta_m) \\
&= \frac{(1-\delta)\alpha}{1+\delta_m + (1-\delta)\alpha} - \frac{2\delta_m}{1+\delta_m + (1-\delta)\alpha}.
\end{aligned}$$

Theorem D.4 in this case says that for $J = 1$ we have with probability $1 - \delta - 2\exp(-c_1 m^{\frac{1}{4}})$

$$1 - \kappa \leq 8\frac{1 \cdot \frac{1}{m}N(m)}{m}(\log(\frac{1}{\delta}))^2 = 8\alpha m(\log(\frac{1}{\delta}))^2.$$

Therefore Theorem 6.5 implies that with probability at least $1 - \delta - 2\exp(-c_1 m^{\frac{1}{4}})$ with $T = \log(\frac{1}{\delta_T})$

$$\mathcal{L}_{\text{ST}} \leq 8\alpha m(\log(\frac{1}{\delta}))^2\mathcal{L}_{\text{TE}} + \frac{1}{m^2}(\log(\frac{1}{\delta_T}))^2\mathcal{L}_{\text{TE}} + \frac{\delta_T}{2-\delta_T}\|f^*\|_2^2.$$

For $\delta_T = \frac{1}{m^2}\mathcal{L}_{\text{TE}}$, i.e. for stopping time $T = 2\log m + \log\frac{1}{\mathcal{L}_{\text{TE}}}$

$$\mathcal{L}_{\text{ST}} \leq 8\alpha m(\log(\frac{1}{\delta}))^2\mathcal{L}_{\text{TE}}\left(1 + \frac{2\log^2 m + \log^2\frac{1}{\mathcal{L}_{\text{TE}}}}{m^2 \cdot \alpha m(\left(\log\frac{1}{\delta}\right)^2}\right).$$

As long as $\alpha m\left(\log\frac{1}{\delta}\right)^2 = \Omega(m^{-2+\epsilon_s})$ for some $\epsilon > 0$ this can be simplified to

$$\mathcal{L}_{\text{ST}} \leq 8\alpha m(\log(\frac{1}{\delta}))^2\mathcal{L}_{\text{TE}}\left(1 + O(m^{-\epsilon_s})\right).$$

Note that we can simplify the lower bound for $\mathcal{L}_{\text{TE}}$, i.e. we have that with probabilit at least $1-2\exp(-\frac{d\delta^2}{4})-2\exp(-\frac{m\delta_m^2}{4})$

$$\mathcal{L}_{\text{TE}} \geq \frac{\alpha}{1+\alpha+\delta_m-\delta\alpha} - 2\delta_m - \delta\alpha.$$

As long as $\frac{1}{100} \geq \delta_m - \delta\alpha$, this can be further simplified to

$$\mathcal{L}_{\text{TE}} \geq 0.99\frac{\alpha}{1+\alpha} - 2\delta_m - \delta\alpha.$$

Therefore, taking $\delta = \frac{1}{d^{\frac{1}{2}-\epsilon_d}}$ and $\delta_m = \frac{1}{m^{\frac{1}{2}-\epsilon_m}}$, we have that with probability at least $1 - 2\exp(-\frac{1}{4}(\alpha m^2)^{\epsilon_d}) - 2\exp(-\frac{1}{4}m^{\epsilon_m})$ that

$$\mathcal{L}_{\text{TE}} \geq 0.99\frac{\alpha}{1+\alpha} - 2m^{-\frac{1}{2}+\epsilon_m} - (\alpha m^2)^{-\frac{1}{2}+\epsilon_d}\alpha.$$

□

E. Weak-to-Strong Improvement in Random Feature Networks.

In this section, we provide detailed versions of Theorems 3.1 and 3.2, namely Corollary E.3 and Theorem E.4, respectively. Further, we show that the result for ReLU networks can be generalized to any activation function.

2-layer Network with Arbitrary Activation Function Further, we show that the result for ReLU Model 2.2 generalizes to a 2-layer network with features drawn uniformly and any activation function, which we define in Model E.1.

Model E.1 (2-layer Network)**.** Consider a 2-layer network with an activation function σ that is bounded on $[-1, 1]$. Let the distribution $\mathcal{D}$ be $\boldsymbol{x} \sim \text{Unif}(\mathbb{S}^{d-1})$. The first layer parameters $\boldsymbol{u}_i$ are initialized randomly isotropically, i.e. $\boldsymbol{u}_i \sim \text{Unif}(\mathbb{S}^{d-1})$. Let $\{\sigma_i\}_{i=1}^{\infty}$ be the nonzero coefficients of the activation function $\sigma(\langle \boldsymbol{x}, \boldsymbol{t}\rangle)$ in the basis of spherical harmonics, i.e. $\sigma(\langle \boldsymbol{x}, \boldsymbol{t}\rangle) = \sum_{i,k}\sigma_i\phi_{i,k}(\boldsymbol{x})\phi_{i,k}(\boldsymbol{t})$, where the sum goes over only the nonzero σ_i. Here $\phi_{i,k}$ is the i-th spherical harmonic of order k and is also the eigenbasis of the induced kernel (Equation (8))[**].

Theorem E.2 (Weak-to-Strong Generalization with 2-layer NN)**.** *Under Model E.1, let Condition 6.1 be satisfied for groundturth f^*, population kernel $\mathcal{K}$ (Equation (8)) and positive integer K. If $M_{\text{TE}} \geq K$ and the teacher attains the minimum loss (Condition 2.1), then for any $\delta_T \in (0, 1)$ and the student trained to stopping time $T = \frac{1}{\sigma_K^2}\log\frac{1}{\delta_T}$, with probability at least $1 - \frac{2}{M_{\text{TE}}} - K\exp(-\frac{c}{2}\frac{M_{\text{TE}}}{K})$ over the randomness of teacher features, it holds that*

$$\mathcal{L}_{\text{ST}} \leq \inf_{S\geq K}\left\{\frac{4(\sum_{k=1}^{S}\sigma_k^2)(\sum_{k=S+1}^{\infty}\sigma_k^2)\log^2 M_{\text{TE}}}{\sigma_S^4 M_{\text{TE}}}\mathcal{L}_{\text{TE}} + \frac{\sigma_{S+1}^4}{\sigma_K^4}\left(\log\frac{1}{\delta_T}\right)^2\mathcal{L}_{\text{TE}}\right\} + \frac{\delta_T}{2-\delta_T}\left\|f^*\right\|_{\mathcal{D}}^2.$$

For the proof of this theorem see Appendix F.

[**]The existence of such a decomposition follows from the Funk-Hecke formula

2-layer ReLU Networks. As a corollary, we have the bound for ReLU networks. For ReLU activation in Model 2.2, the expansion of σ in the basis of spherical harmonics $\phi_{k,i}$ will be $(\sigma_1, \dots, \sigma_1, \sigma_2, \dots, \sigma_2, \dots)$ where each σ_i repeats $N_k = \frac{(2k+d-2)(k+d-3)!}{k!(d-2)!}$ times, which is the order of the k-th spherical harmonics in d dimensions. Furthermore, σ_k is nonzero only for even k and $k = 1$. Therefore, Condition 6.1 in this case amounts to f^* being a sum of a linear functions and an even polynomial of order at most k, where $K = N_0 + \cdots + N_k$.

Corollary E.3 (Weak-to-Strong Generalization with 2-layer ReLU NN)**.** *Under Model 2.2, let Condition 6.1 be satisfied for f^* for some K and let k be the corresponding degree. If the teacher attains minimum loss $\mathcal{L}_{\mathrm{TE}} = \hat{\mathcal{L}}_{\min}$, then the student trained to early stopping time $T = \frac{1}{\sigma_k^2} \log \frac{1}{\delta_T}$ will have with probability at least $1 - \frac{2}{M_{\mathrm{TE}}} - K \exp(-c\frac{M_{\mathrm{TE}}}{K})$,*

$$\mathcal{L}_{\mathrm{ST}} \le O_d\left(\frac{1}{\sqrt{M_{\mathrm{TE}}}} (\log M_{\mathrm{TE}})^2\right) \cdot \mathcal{L}_{\mathrm{TE}} + O_d\left(\frac{1}{\sigma_k^4} \frac{1}{\sqrt{M_{\mathrm{TE}}}}\right) \cdot \left(\log \frac{1}{\delta_T}\right)^2 \mathcal{L}_{\mathrm{TE}} + \frac{\delta_T}{2 - \delta_T} \|f^*\|_{\mathcal{D}}^2.$$

Here c is an absolute constant.

For the proof of this theorem see Appendix F.1.

Linear Networks Finally, we can state the general bound for linear networks, Model 2.3.

Theorem E.4 (Weak-to-Strong Generalization in Linear Networks)**.** *For Model 2.3, if f^* is spanned by the first K coordinates and if the teacher attains minimum loss $\mathcal{L}_{\mathrm{TE}} = \hat{\mathcal{L}}_{\min}$, then for the student trained until time T we have with probability at least $1 - \frac{4}{m}$ that*

$$\mathcal{L}_{\mathrm{ST}} \le \inf_{S \ge K} \left\{ \frac{\left(\sum_{i=1}^{S} \psi_i\right)\left(\sum_{i=S+1}^{d} \psi_i\right)}{\psi_S^2} \frac{(\log m)^2}{m} \mathcal{L}_{\mathrm{TE}} + \psi_{S+1} T^2 \mathcal{L}_{\mathrm{TE}} \right\} + \frac{e^{-\psi_K T}}{2 - e^{-\psi_K T}} \left\|f^*\right\|_{\mathcal{D}}^2.$$

For the proof, see Appendix D.

F. Weak-to-Strong Bound Applied to the Case of 2-layer NN

This section deals with the cases of Model 2.2 and Model E.1.

Proof of Theorem 3.1. In Theorem F.2, if we select $\delta_T = \frac{1}{\sqrt{M_{\mathrm{TE}}}} \mathcal{L}_{\mathrm{TE}}$, then with stopping time $T = \frac{1}{\sigma_k^2}(\log M_{\mathrm{TE}} + \log \frac{1}{\mathcal{L}_{\mathrm{TE}}})$ we have that with probability $1 - \frac{2}{M_{\mathrm{TE}}} - K \exp(-c\frac{M_{\mathrm{TE}}}{K})$

$$\begin{aligned}\mathcal{L}_{\mathrm{ST}} &\le O_{d,k}\left(\frac{1}{\sqrt{M_{\mathrm{TE}}}} (\log M_{\mathrm{TE}})^2\right) \mathcal{L}_{\mathrm{TE}} \\ &\quad + O_{d,k}\left(\frac{1}{\sigma_k^4} \frac{1}{\sqrt{M_{\mathrm{TE}}}}\right) (\log M_{\mathrm{TE}} - \log \mathcal{L}_{\mathrm{TE}})^2 \mathcal{L}_{\mathrm{TE}} + \frac{1}{\sqrt{M_{\mathrm{TE}}}} \mathcal{L}_{\mathrm{TE}} \|f^*\|_{\mathcal{D}}^2.\end{aligned}$$

This will hold if $M_{\mathrm{TE}} / \log M_{\mathrm{TE}} > K = N_k + \cdots + N_0$ so it suffices to have $M_{\mathrm{TE}} \ge (N_0 + \cdots + N_k)^2$. For our setup, this is $(N_0 + \cdots + N_k)^2 = \Theta(d^{2k})$. From this we have that

$$\begin{aligned}\mathcal{L}_{\mathrm{ST}} &\le O_{d,k}\left(\frac{1}{\sqrt{M_{\mathrm{TE}}}} (\log M_{\mathrm{TE}})^2\right) \mathcal{L}_{\mathrm{TE}} \\ &\quad + O_{d,k}\left(\frac{1}{\sigma_k^4} \frac{1}{\sqrt{M_{\mathrm{TE}}}}\right) \left(\log^2 M_{\mathrm{TE}} + \log^2 \mathcal{L}_{\mathrm{TE}}\right) \mathcal{L}_{\mathrm{TE}} + \frac{1}{\sqrt{M_{\mathrm{TE}}}} \mathcal{L}_{\mathrm{TE}}\end{aligned}$$

where we used the fact that the target is nomalized. Note now that from Lemma F.8, we have that σ_k can be absorbed into the $O_{d,k}$. So we have that

$$\begin{aligned}\mathcal{L}_{\mathrm{ST}} &\le O_{d,k}\left(\frac{1}{\sqrt{M_{\mathrm{TE}}}} (\log M_{\mathrm{TE}})^2\right) \mathcal{L}_{\mathrm{TE}} \\ &\quad + O_{d,k}\left(\frac{1}{\sqrt{M_{\mathrm{TE}}}}\right) \log^2 \mathcal{L}_{\mathrm{TE}} \mathcal{L}_{\mathrm{TE}}.\end{aligned}$$

This can be written as

$$\mathcal{L}_{\rm ST} \le O_{d,k}\left(\frac{\log^2 M_{\rm TE} + \log^2 \mathcal{L}_{\rm TE}}{\sqrt{M_{\rm TE}}}\,\mathcal{L}_{\rm TE}\right).$$

Now, to get the right exponent, we need a sharp lower bound on $\mathcal{L}_{\rm TE}$. We can get that under the Gaussian Universality Ansatz. Under Gaussian Universality Ansatz, the teacher's test risk $\mathcal{L}_{\rm TE}$ behaves like the deterministic equivalent Equation (20). Therefore, Theorem D.6 applies, which tells us that

$$\mathcal{L}_{\rm TE} \ge \Theta\left(M_{\rm TE}^{-\alpha} d^{2K} \|f\|_{\mathcal{D}}^2\right).$$

In other words, $M_{\rm TE}^{-1} \le O_{d,k}(\mathcal{L}_{\rm TE}^{\frac{1}{\alpha}})$.

Note that under Gaussian Universality, since $\mathcal{L}_{\rm TE}$ is lower bounded polyonomially in $M_{\rm TE}$, we can absorb $\log^2 \mathcal{L}_{\rm TE}$ into $\log^2 M_{\rm TE}$. Therefore, we get

$$\mathcal{L}_{\rm ST} \le O_{d,k}\left(\frac{1}{\sqrt{M_{\rm TE}}}\left(\log M_{\rm TE}\right)^2\right)\mathcal{L}_{\rm TE}.$$

So we have that

$$\mathcal{L}_{\rm ST} \le O_{d,k}\left(\frac{\log^2 M_{\rm TE}}{\sqrt{M_{\rm TE}}}\right)\mathcal{L}_{\rm TE} \le \tilde{O}_{d,k}\left(\mathcal{L}_{\rm TE}^{1+\frac{1}{2\alpha}}\right).$$

Since we can take α to be close to $\frac{d+2}{d-2}$, for $d > 200$ we will have this be 1.02 so the final exponent is 1.49. Finally, note that here K is the number of nonzero σ_k up to index k, so by Corollary F.9, it is of order $\Theta(d^k)$. since c is an absolute constant, for $M_{\rm TE} > \Theta(d^{2k})$, $K\exp(-c\frac{M_{\rm TE}}{d^k}) < \frac{1}{M_{\rm TE}}$ and otherwise we can have the constant in $O_{d,k}$. □

Proof of Theorem E.2. This is given by Theorem F.1 for $t_A = \frac{1}{2}\sqrt{\frac{M_{\rm TE}}{K}}$. Note that since $M_{\rm TE} \ge K$, Condition 6.3 holds with probability one. □

Theorem F.1 (Master Bound for Weak-to-Strong Generalization with 2-layer NN)**.** *Under Model 2.2, if Condition 6.1 is satisfied for f^* and K. If $\sqrt{m} \ge t_A\sqrt{K}$ and the teacher attains minimum loss $\mathcal{L}_{\rm TE} = \hat{\mathcal{L}}_{\min}$, then the student trained to early stopping time $T = \frac{1}{\sigma_K^2}\log\frac{1}{\delta_T}$ will have with probability at least $1 - 2\delta - 2K\exp(-ct_A^2)$*

$$\mathcal{L}_{\rm ST} \le \inf_{S\ge K}\left\{\frac{m(\sum_{k=1}^S \sigma_k^2)(\sum_{k=S+1}^\infty \sigma_k^2)\log^2 m}{\sigma_S^4(\sqrt{m} - t_A\sqrt{K})^2}\,\mathcal{L}_{\rm TE} + \frac{\sigma_{S+1}^4}{\sigma_k^4}\left(\log\frac{1}{\delta_T}\right)^2\mathcal{L}_{\rm TE}\right\} + \frac{\delta_T}{2-\delta_T}\left\|f^*\right\|_{\mathcal{D}}^2.$$

Proof of Theorem F.1. Follows directly by instantiating Theorem 6.5 with the lower bound on κ_S given by Theorem F.4 and the lower bound on $\lambda_J(\boldsymbol{A})$ given by Theorem C.13. Finally, we just choose $T = \frac{1}{\sigma_K^2}\log\frac{1}{\delta_T}$. □

F.1. Weak-to-Strong Bound Applied to the Case of 2-layer ReLU NN

Proof of Corollary E.3. This is given by Theorem F.2. □

Theorem F.2 (Weak-to-Strong Generalization with 2-layer ReLU NN)**.** *Under Model 2.2, if Condition 6.1 is satisfied for f^* and K, let k be the unique number such that $N_0 + \cdots + N_{k-1} + 1 \le K \le N_0 + \cdots + N_k$. If $M_{\rm TE}/\log M_{\rm TE} \ge K$ and the teacher attains minimum loss $\mathcal{L}_{\rm TE} = \hat{\mathcal{L}}_{\min}$, then the student trained to early stopping time $T = \frac{1}{\sigma_k^2}\log\frac{1}{\delta_T}$ will have with probability at least $1 - \frac{2}{M_{\rm TE}} - K\exp(-c\frac{M_{\rm TE}}{k})$*

$$\begin{aligned}\mathcal{L}_{\rm ST} &\le O_{d,k}\left(\frac{1}{\sqrt{M_{\rm TE}}}\left(\log M_{\rm TE}\right)^2\right)\mathcal{L}_{\rm TE} \\ &\quad + O_{d,k}\left(\frac{1}{\sigma_k^4}\frac{1}{\sqrt{M_{\rm TE}}}\right)\left(\log\frac{1}{\delta_T}\right)^2\mathcal{L}_{\rm TE} + \frac{\delta_T}{2-\delta_T}\|f^*\|_{\mathcal{D}}^2.\end{aligned}$$

Here c is an absolute constant.

Proof of Theorem F.2. We will apply Theorem F.1. Note that the eigenvalues of the activation functions σ in this case are σ_k and note that for odd $k > 1$, we have $\sigma_k = 0$. Note that we can just ignore the zero eigenvalues. So in this case, σ_S will actually be equal to σ_s and σ_{S+1} to σ_{s+2} if we select $S = N_0 + \cdots + N_s$. Take s from Proposition F.3. Then the proposition says that we have

$$\frac{m(\sum_{k=1}^{s}\sigma_k^2 N_k)(\sum_{k=s}^{\infty}\sigma_k^2 N_k)\left(\log\frac{1}{\delta}\right)^2}{\sigma_s^4\left(\sqrt{m}-t_A\sqrt{K}\right)^4} \le \Theta_d\left(\frac{m\sqrt{m}\left(\log\frac{1}{\delta}\right)^2}{\left(\sqrt{m}-t_A\sqrt{K}\right)^4}\right)$$
$$\lambda_{S+1}^2 T^2 = \frac{\sigma_{s+2}^4}{\sigma_k^4}\left(\log\frac{1}{\delta_T}\right)^2 \le \Theta_d\left(\frac{1}{\sqrt{m}}\frac{1}{\sigma_k^4}\left(\log\frac{1}{\delta_T}\right)^2\right).$$

Therefore, from Theorem F.1, we have that with probability at least $1 - 2\delta - 2K\exp(-ct_A^2)$

$$\mathcal{L}_{\mathrm{ST}} \le \Theta_d\left(\frac{m\sqrt{m}\left(\log\frac{1}{\delta}\right)^2}{\left(\sqrt{m}-t_A\sqrt{K}\right)^4}\right)\mathcal{L}_{\mathrm{TE}}$$
$$+\Theta\left(\frac{1}{\sqrt{m}}\frac{1}{\sigma_k^2}\left(\log\frac{1}{\delta_T}\right)^2\right)\left(\log\frac{1}{\delta_T}\right)^2\mathcal{L}_{\mathrm{TE}} + \frac{\delta_T}{2-\delta_T}\|f^*\|_{\mathcal{D}}^2.$$

Take $\delta = \frac{1}{m}$ and take $t_a = m^{\frac{1}{2}}\frac{1}{2\sqrt{K}}$ so that $\sqrt{m} - t_A\sqrt{K} \ge \sqrt{m}/2$. Then $2K\exp(-ct_A^2) \le \frac{1}{m}$ if $\frac{m}{\log m} \ge K$. Note that therefore with probability at least $1 - \frac{4}{m}$, we have

$$\mathcal{L}_{\mathrm{ST}} \le \Theta_d\left(\frac{1}{\sqrt{M_{\mathrm{TE}}}}\left(\log M_{\mathrm{TE}}\right)^2\right)\mathcal{L}_{\mathrm{TE}}$$
$$+\Theta\left(\frac{1}{\sigma_k^4}\frac{1}{\sqrt{M_{\mathrm{TE}}}}\right)\left(\log\frac{1}{\delta_T}\right)^2\mathcal{L}_{\mathrm{TE}} + \frac{\delta_T}{2-\delta_T}\|f^*\|_{\mathcal{D}}^2,$$

which completes the proof. □

F.2. Proofs Related to 2-layer NN

Proposition F.3 (Optimal choice of s)**.** *We can take $s = m^{\frac{1}{4(d-4)}} + o_m(1)$ (s is closest even integer to $m^{\frac{1}{4(d-4)}}$) so that*

$$\frac{m(\sum_{k=1}^{s}\sigma_k^2 N_k)(\sum_{k=s}^{\infty}\sigma_k^2 N_k)\left(\log\frac{1}{\delta}\right)^2}{\sigma_s^4\left(\sqrt{m}-t_A\sqrt{K}\right)^4} \le \Theta_d\left(\frac{m\sqrt{m}\left(\log\frac{1}{\delta}\right)^2}{\left(\sqrt{m}-t_A\sqrt{K}\right)^4}\right)$$
$$\lambda_{S+1}^2 T^2 = \frac{\sigma_{s+2}^4}{\sigma_k^4}\left(\log\frac{1}{\delta_T}\right)^2 \le \Theta_d\left(\frac{1}{\sqrt{m}}\frac{1}{\sigma_k^4}\left(\log\frac{1}{\delta_T}\right)^2\right).$$

Note that this Θ hides absolute constants and d dependence.

Proof. Note that for $k \gg d$ we have that $N_k = \Theta(k^{d-2})$. This follows directly from the closed form $N_k = \frac{(2k+d-2)(k+d-3)!}{k!(d-2)!}$. Furthermore, by Lemma F.8, we have that $\sigma_k^2 = \Theta(k^{-(d+2)})$ for even $k \gg d$. Furthermore, $\sum_{k=1}^{\infty}\sigma_k^2 N_k = \sigma(1)$. So, we have that

$$\frac{m(\sum_{k=1}^{s}\sigma_k^2 N_k)(\sum_{k=s}^{\infty}\sigma_k^2 N_k)\left(\log\frac{1}{\delta}\right)^2}{\sigma_s^4\left(\sqrt{m}-t_A\sqrt{K}\right)^4} \le \frac{1}{\sigma_s^4}\frac{m\sigma(1)^2\left(\log\frac{1}{\delta}\right)^2}{\left(\sqrt{m}-t_A\sqrt{K}\right)^4}.$$

Therefore, taking $s = m^{\frac{1}{4(d+2)}} + o_m(1)$, where $o_m(1)$ is taken so that s is the closest integer to $m^{\frac{1}{4(d+2)}}$. Then since $\sigma_s^2 = \Theta(s^{-(d+2)})$, we have that $\sigma_s^2 = \Theta(m^{-\frac{1}{4}})$, so $\frac{1}{\sigma_s^4} = \Theta(\sqrt{m})$. Note that $\sigma_{s+2}^2 = \Theta((m^{\frac{1}{4(d+2)}}+2)^{-(d+2)}) = \Theta(m^{-\frac{1}{4}})$. Therefore $\sigma_{s+2}^4 = \Theta(\frac{1}{\sqrt{m}})$. □

Theorem F.4. *Under Model 2.2, if Condition 6.1 is satisfied for f^* and K, let s be such that $N_0 + \cdots + N_{s-1} + 1 \leq K \leq N_0 + \cdots + N_s$. Then with probability at least $1 - 2\delta - 2K\exp(-ct_A^2)$*

$$\frac{1}{\kappa_S} - 1 \leq \lambda_1\Big((\sqrt{\boldsymbol{A}_S})^+ \boldsymbol{B}_S (\sqrt{\boldsymbol{A}_S})^+\Big) \leq \frac{\|F_B^\dagger F_A\|_{\mathcal{H}}^2}{\lambda_S(\boldsymbol{A})^2} \leq \frac{m(\sum_{k=1}^s \sigma_k^2 N_k)(\sum_{k=s}^\infty \sigma_k^2 N_k)\left(\log\frac{1}{\delta}\right)^2}{\sigma_s^4\left(\sqrt{m} - t_A\sqrt{K}\right)^4}$$

where c is an absolute constant and $\delta \in (0,1)$. For $s \geq 2$ then withe probability at least $1 - 2\delta - 2K\exp(-ct_A^2)$

$$\lambda_1\Big((\sqrt{\boldsymbol{A}_S})^+ \boldsymbol{B}_S (\sqrt{\boldsymbol{A}_S})^+\Big) \leq \frac{m\sigma(1)^2\left(\log\frac{1}{\delta}\right)^2}{\sigma_s^4\left(\sqrt{m} - t_A\sqrt{K}\right)^4}.$$

Proof. The first bound follows from Theorem C.13 applied to $\lambda_J(\boldsymbol{A})$ and Proposition C.12 applied with Corollary F.7. This gives that with probability at least $1 - 2\delta - 2K\exp(-ct_A^2)$

$$\lambda_1\Big((\sqrt{\boldsymbol{A}_S})^+ \boldsymbol{B}_S (\sqrt{\boldsymbol{A}_S})^+\Big) \leq \frac{\|F_B^\dagger F_A\|_{\mathcal{H}}^2}{\lambda_S(\boldsymbol{A})^2} \leq \frac{m(\sum_{k=1}^s \sigma_k^2 N_k)(\sum_{k=s}^\infty \sigma_k^2 N_k)\left(\log\frac{1}{\delta}\right)^2}{\sigma_s^4\left(\sqrt{m} - t_A\sqrt{K}\right)^4}.$$

The second bound follows by noting that $\sum_{k=1}^\infty \sigma_k^2 N_k = \sigma(\langle \boldsymbol{x}, \boldsymbol{x}\rangle) = \sigma(1)$. □

In this case, the basis of eigenfunctions can be taken to be the spherical harmonics $\{\{\phi_{k,i}\}_{i=1}^{N_k}\}_{k=1}^\infty$ where $\phi_{k,i}$ is the i-th spherical harmonic of order k in d dimensions and $N_k = \frac{(2k+d-2)(k+d-3)!}{k!(d-2)!}$ is the dimensionality of spherical harmonics of order k in d dimensions.

Proposition F.5 (Eigendecomposition of the kernel induced by random features in the case of 2-layer NN)**.** *Under Model 2.2, the kernel induced by the random features is given by*

$$\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{k=1}^\infty \sigma_k^2 \langle \phi_k(\boldsymbol{x}), \phi_k(\boldsymbol{x}')\rangle.$$

Here, $\phi_k : \mathbb{R}^d \to \mathbb{R}^{N_k}$ is a vector of spherical harmonics of order k in d dimension, $N_k = \frac{(2k+d-2)(k+d-3)!}{k!(d-2)!}$, and σ_k is the k-th term in the decomposition of the ReLU functions σ to the basis of $\{\phi_k\}_{i=1}^\infty$.

We provide the estimates for the size of σ_k in Appendix F.3.

Finally, we find B and L in the spherical case.

Proposition F.6. *Under Model 2.2, if $J = N_0 + N_1 + \cdots + N_s$ then*

$$\|P_{\leq J} g\|_{\mathcal{D}}^2 = \sum_{k=1}^s \sigma_k^2 N_k$$
$$\|P_{\geq J+1} g\|_{\mathcal{D}}^2 = \sum_{k=s+1}^\infty \sigma_k^2 N_k.$$

So $L = B = \sqrt{(\sum_{k=1}^s \sigma_k^2 N_k)\left(\sum_{k=s+1}^\infty \sigma_k^2 N_k\right)}$.

Proof. We have that

$$P_{\leq J} g = \sum_{k=1}^s \sigma_k \phi_k(\boldsymbol{w})^T \phi_k(\boldsymbol{x})$$
$$\|P_{\leq J} g\|_{\mathcal{D}}^2 = \langle \sum_{k=1}^s \sigma_k \phi_k(\boldsymbol{w})^T \phi_k(\boldsymbol{x}), \sum_{l=1}^s \sigma_l \phi_l(\boldsymbol{w})^T \phi_l(\boldsymbol{x})\rangle_{\mathcal{D}} = \sum_{k=1}^s \sum_{l=1}^s \sigma_k \sigma_l \langle \phi_k(\boldsymbol{w})^T \phi_k(\boldsymbol{x}), \phi_l(\boldsymbol{w})^T \phi_l(\boldsymbol{x})\rangle_{\mathcal{D}}.$$

Now since $\phi_k(\boldsymbol{w})^T\phi_k(\boldsymbol{x}) = \sum_{j=1}^{N_k}\phi_{k,j}(\boldsymbol{w})\phi_{k,j}(\boldsymbol{x})$ and since $\{\phi_{k,j}\}$ are orthogonal w.r.t $\mathcal{D}$, only the terms of the form $\phi_{k,j}(\boldsymbol{w})\phi_{k,j}(\boldsymbol{x})\phi_{k,j}(\boldsymbol{w})\phi_{k,j}(\boldsymbol{x})$ remain. For these, we have that $\mathbb{E}_{\mathcal{D}}\left(\phi_{k,j}(\boldsymbol{w})\phi_{k,j}(\boldsymbol{x})\phi_{k,j}(\boldsymbol{w})\phi_{k,j}(\boldsymbol{x})\right) = \phi_{k,j}(\boldsymbol{w})^2\cdot 1$. So we have that

$$\begin{aligned}\|P_{\leq J}g\|_{\mathcal{D}}^2 &= \sum_{k=1}^{s}\sigma_k^2\langle\phi_k(\boldsymbol{w})^T\phi_k(\boldsymbol{x}),\phi_k(\boldsymbol{w})^T\phi_k(\boldsymbol{x})\rangle_{\mathcal{D}}\\ &= \sum_{k=1}^{s}\sigma_k^2\phi_{k,j}(\boldsymbol{w})^2 = \sum_{k=1}^{s}\sigma_k^2 N_k.\end{aligned}$$

The last step holds because $\phi_k(\boldsymbol{w})^T\phi_k(\boldsymbol{w}) = N_k P_k(\langle\boldsymbol{w},\boldsymbol{w}\rangle) = N_k P_k(1) = N_k$.

The same argument shows that $\|P_{J+1}g\|_{\mathcal{D}}^2 = \sum_{k=s+1}^{\infty}\sigma_k^2 N_k$. □

Corollary F.7. *Under Model 2.2, if $N_0+\cdots+N_{s-1}+1 \leq J \leq N_0+\cdots+N_s$ then*

$$L = B = \sqrt{(\sum_{k=1}^{s}\sigma_k^2 N_k)(\sum_{k=s}^{\infty}\sigma_k^2 N_k)}.$$

This simplifies to $L = B \leq \Theta(\frac{1}{d})\sqrt{\frac{2^s}{s!}}$.

Proof. Note that if $J = N_0+N_1+\cdots+N_{s-1}+p$ for some $1\leq p\leq N_s$ then we have that

$$\begin{aligned}\|P_{\leq J}g\|_{\mathcal{D}}^2 &= \sum_{k=1}^{s-1}\sigma_k^2 N_k + p\sigma_s^2\\ \|P_{\geq J+1}g\|_{\mathcal{D}}^2 &= (N_k-p)\sigma_s^2 + \sum_{k=s+1}^{\infty}\sigma_k^2 N_k.\end{aligned}$$

□

F.3. Computation of the Kernel Eigenvalues for the case 2-layer NN

Proof. of Proposition F.5: For ReLU activation, we can do the following decomposition:

$$\sigma(\langle\boldsymbol{w},\boldsymbol{x}\rangle) = \sum_{k\geq 0}N_k\sigma_k P_{d,k}(\langle\boldsymbol{w},\boldsymbol{x}\rangle) = \sum_{k\geq 0}\sigma_k\langle\boldsymbol{\phi}_{d,k}(\boldsymbol{w}),\boldsymbol{\phi}_{d,k}(\boldsymbol{x})\rangle,$$

Here σ_k is given by

$$\sigma_k = \frac{|\mathbb{S}^{d-2}|}{|\mathbb{S}^{d-1}|}\int_0^1 sP_{k,d}(s)(1-s^2)^{\frac{d-3}{2}}\mathrm{d}s.$$

Combining this with $|\mathbb{S}^{d-1}| = \frac{2\pi^{d/2}}{\Gamma(\frac{d}{2})}$, we have

$$\sigma_k = \frac{\Gamma(\frac{d}{2})}{\sqrt{\pi}\Gamma(\frac{d-1}{2})}\int_0^1 sP_{k,d}(s)(1-s^2)^{\frac{d-3}{2}}\mathrm{d}s.$$

The eigendecomposition of the kernel in this case is given by

$$\mathcal{K}(\boldsymbol{x},\boldsymbol{x}') = \mathbb{E}_{\boldsymbol{w}}[\sigma(\langle\boldsymbol{w},\boldsymbol{x}\rangle)\sigma(\langle\boldsymbol{w},\boldsymbol{x}'\rangle)] = \mathbb{E}_{\boldsymbol{w}}\left[\sum_{k\geq 0}\sigma_k\langle\boldsymbol{\phi}_k(\boldsymbol{w}),\boldsymbol{\phi}_k(\boldsymbol{x})\rangle\cdot\sigma_k\langle\boldsymbol{\phi}_k(\boldsymbol{w}),\boldsymbol{\phi}_k(\boldsymbol{x}')\rangle\right] = \sum_{k\geq 0}\sigma_k^2\langle\boldsymbol{\phi}_k(\boldsymbol{x}),\boldsymbol{\phi}_k(\boldsymbol{x}')\rangle.$$

□

Now, we find the size of σ_k.

Lemma F.8 (Size of σ_k)**.** *Depending on the size of d and k, the following holds for the size of coefficients of ReLU activation in the basis of Legendre polynomials $\sigma_k = \frac{\Gamma(\frac{d}{2})}{\sqrt{\pi}\Gamma(\frac{d-1}{2})} \cdot \frac{1}{2^k} \frac{1}{(\frac{k}{2}+\frac{d-1}{2})\dots(\frac{d-1}{2})}$ for even k, $\sigma_1 = \frac{1}{2d}$, and $\sigma_k = 0$ for odd $k > 1$. Asymptotically, this means*

1. $\sigma_0 = \frac{1}{\sqrt{2\pi d}} + \Theta(d^{-\frac{3}{2}})$, $\sigma_1 = \frac{1}{2d}$
2. *If $k = \Theta(1)$, then for even k, $\sigma_k = \Theta(d^{-\frac{k+1}{2}})$.*
3. *If $k \gg d$, then for even k, $\sigma_k = \Theta(k^{-\frac{d-1}{2}-\frac{3}{2}})$.*

Proof. The third case is shown in (Petrini et al., 2022). To prove the general claim, we use the formula derived in (Petrini et al., 2022)

$$\sigma_k = \frac{\Gamma(\frac{d}{2})}{\sqrt{\pi}\Gamma(\frac{d-1}{2})} \cdot \frac{\Gamma(\frac{d-1}{2})}{\Gamma(k+\frac{d-1}{2})} \left(-\left(-\frac{1}{2}\right)^k\right) \frac{d^{k-2}}{dt^{k-2}}(1-t^2)^{k+\frac{d-3}{2}}\Big|_0^1.$$

Note that all terms will be multiplied by $(1-t^2)^{2+\frac{d-3}{2}}$ which is 0 at $t=1$, so we only need to consider $t=0$. Note that if a term has a nonzero power of t, it will be evaluated to 0. When taking j-th ($j \le k-2$) derivative of the expression above, all the terms will be in the form of $(1-t^2)^{l+\frac{d-3}{2}} t^s$ times some constant that depends on d for some integers l and s. So if out of $k-2$ derivatives we take a on the term with $(1-t^2)$ and $k-2-a$ on t^s (note that the ordering doesn't matter as we only add constants), in order to have term with no t at the end, we need to have $k-2-a = s$, but $s = a$ so we have that $a = \frac{k-2}{2}$. Therefore, the evaluated derivate equals exactly

$$\frac{d^{k-2}}{dt^{k-2}}(1-t^2)^{k+\frac{d-3}{2}}\Big|_0^1 = \left(k+\frac{d-3}{2}\right)\dots\left(k-(\frac{k-2}{2}-1)+\frac{d-3}{2}\right) = \left(k+\frac{d-3}{2}\right)\dots\left(\frac{k+4}{2}+\frac{d-3}{2}\right).$$

Note that $\frac{\Gamma(\frac{d-1}{2})}{\Gamma(k+\frac{d-1}{2})} = \frac{1}{(k-1+\frac{d-1}{2})\dots(\frac{d-1}{2})}$, therefore we can compute that

$$\sigma_k = \frac{1}{2^k}\frac{\Gamma(\frac{d}{2})}{\sqrt{\pi}\Gamma(\frac{d-1}{2})} \cdot \frac{1}{(k-1+\frac{d-1}{2})\dots(\frac{d-1}{2})}\left(k+\frac{d-3}{2}\right)\dots\left(\frac{k+4}{2}+\frac{d-3}{2}\right)$$

$$\sigma_k = \frac{\Gamma(\frac{d}{2})}{\sqrt{\pi}\Gamma(\frac{d-1}{2})} \cdot \frac{1}{2^k}\frac{1}{(\frac{k}{2}+\frac{d-1}{2})\dots(\frac{d-1}{2})}.$$

$$\sigma_k = \frac{\Gamma(\frac{d}{2})}{\sqrt{\pi}\Gamma(\frac{d-1}{2})} \cdot \frac{1}{2}\frac{1}{(k+d-1)((k-2)+d-1)((k-4)+d-1)\dots(2+d-1)(d-1)}.$$

So, when $k = \Theta(1)$ we have that this is $\sigma_k = \Theta(d^{\frac{1}{2}}d^{-\frac{k+2}{2}}) = \Theta(d^{-\frac{k+1}{2}})$, since the first term is $\frac{\Gamma(\frac{d}{2})}{\sqrt{\pi}\Gamma(\frac{d-1}{2})} = \Theta(d^{\frac{1}{2}})$. □

Corollary F.9 (Size of $\sigma_k^2 N_k$)**.** *The following bound holds for $\sigma_k^2 N_k$ for k even or $k = 1$ If $k \gg d$ then $\sigma_k^2 N_k = \Theta(k^{-4})$. If $k \ll d$ then $\sigma_k^2 N_k = \Theta(d^{-1})$. If $k > 1$ is odd, then $\sigma_k^2 N_k = 0$.*

Proof. We have for $\sigma_k^2 N_k$ that

$$\sigma_k^2 N_k = \frac{1}{4}\left(\frac{\Gamma(\frac{d}{2})}{\sqrt{\pi}\Gamma(\frac{d-1}{2})}\right)^2 \cdot \frac{(d-1)(d-1+1)\dots(d-1+k-2)}{(k+d-1)^2\dots(d-1+2)^2(d-1)^2}(2k+d-2)$$

$$= \frac{1}{4}\left(\frac{\Gamma(\frac{d}{2})}{\sqrt{\pi}\Gamma(\frac{d-1}{2})}\right)^2 \cdot (2k+d-2)\frac{(d-1+1)(d-1+3)\dots(d-1+k-3)}{(d-1)(d-1+2)\dots(d-1+k-4)(d-1+k-2)}\frac{1}{(d-1+k)^2}.$$

When $k \ll d$, we can just count the terms with d in the numerator and denominator and note that $\left(\frac{\Gamma(\frac{d}{2})}{\sqrt{\pi}\Gamma(\frac{d-1}{2})}\right)^2 = \Theta(d)$. For $k \gg d$, we just need to count the terms with k in the numerator and denominator. The results follow.

□

G. Bootstrapping

A natural question to ask is whether extending the weak-to-strong setup described in Section 2 to use a chain of students can help break the limitations of Corollary 4.4. That is, training a teacher on the target, the first student on the teacher, the second student on the first student, and so on in a bootstraped game of "telephone", then finally asking if the final student can be much better than the teacher. The lower bound of Theorem 4.3 and Corollary 4.4 do not apply here as-is, because shrink-optimality is not transitive. That is, if f_2 is shrink-optimal w.r.t. f_1, and f_1 is shrin-optimal w.r.t. f_{teacher}, this does not imply f_2 is shrink-optimal w.r.t. f_{teacher}. Instead, we provide a slightly weaker lower bound against a more general proprety, which *is* transitive:

Theorem G.1 (Limitation of Weak-to-Strong Generalization with a Bounded Student)**.** *If the target f^* is normalized, $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[(f^*(\boldsymbol{x}))^2] = 1$, then for any teacher f_{teacher} shrinking-optimal w.r.t. f^* and any student f_{student} s.t. $\mathbb{E}_{\boldsymbol{x}}\left[f_{\text{student}}(\boldsymbol{x})^2\right] \leq \mathbb{E}_{\boldsymbol{x}}\left[f_{\text{teacher}}(\boldsymbol{x})^2\right]$, we have that*

$$\mathcal{L}_{\text{ST}} \geq \left(1 - \sqrt{1 - \mathcal{L}_{\text{TE}}}\right)^2 \geq \frac{1}{4}\mathcal{L}_{\text{TE}}^2$$

It is easy to verify that the shrink optimality implies $\mathbb{E}_{\boldsymbol{x}}\left[\hat{f}(\boldsymbol{x})^2\right] \leq \mathbb{E}_{\boldsymbol{x}}\left[f_{\text{train}}(\boldsymbol{x})^2\right]$, and the transitivity of the inequality now implies:

Corollary G.2 (Limitation of Weak-to-Strong Generalization with Bootstrapping)**.** *If the target f^* is normalized, $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[(f^*(\boldsymbol{x}))^2] = 1$, then for any teacher f_{teacher} shrinking-optimal w.r.t. f^* and any chain of student $f^{(i)}_{\text{student}}$, $i = 0, 2, \ldots, \infty$, where $f^{(i)}_{\text{student}} = f_{\text{teacher}}$ and $f^{(i+1)}_{\text{student}}$ is shrinking-optimal w.r.t. $f^{(i)}_{\text{student}}$, we have that for all $i \geq 1$*

$$\mathcal{L}_{\text{ST},i} \geq \left(1 - \sqrt{1 - \mathcal{L}_{\text{TE}}}\right)^2 \geq \frac{1}{4}\mathcal{L}_{\text{TE}}^2,$$

where $\mathcal{L}_{\text{ST},i}$ is the test loss of the n-th student. In particular, this includes the case where the student $f^{(i)}_{\text{student}}$ is trained by gradient flow w.r.t. $f^{(i-1)}_{\text{student}}$ for arbitrary stopping time T_i, and using arbitrarily (random or determnistic) features.

For the proof of Theorem G.1 and Corollary G.2 see Appendix H.

H. Proofs of Limitation of Weak-to-Strong Generalization

Proof of Corollary 4.4. This follows from Theorem 4.3 by noting that from Lemma 4.2 the teacher satisfies Definition 4.1 with respect to the target and the student satisfies Definition 4.1 with respect to the teacher. □

Proof of Lemma 4.2. We rewrite the claim of Lemma 4.2 as follows.

Lemma H.1 (Shrinkage Optimality of Gradient Flow Solutions)**.** *For any positive semi-definite kernel $\mathcal{K}$, time $T > 0$, and ground truth f^*, the gradient flow solution $f_T = \mathcal{T}_T^{\mathcal{K}}(f^*)$ is shrinkage optimal with respect to f^* in the sense of Definition 4.1.*

From Equation (9), the gradient flow solution is $f_T = \sum_{k\geq 1}(1 - e^{-\lambda_k T})\langle f^*, e_k\rangle_{\mathcal{D}} e_k$. For shrinkage optimality, we need to show that $\alpha = 1$ minimizes $\mathbb{E}_{\mathbf{x}}[(\alpha f_T(\mathbf{x}) - f^*(\mathbf{x}))^2]$ for $0 \leq \alpha \leq 1$.

Expanding this expression in the eigenbasis:

$$\mathbb{E}_{\mathbf{x}}[(\alpha f_T(\mathbf{x}) - f^*(\mathbf{x}))^2] = \left\| \alpha \sum_{k\geq 1} (1 - e^{-\lambda_k T}) \langle f^*, e_k \rangle_{\mathcal{D}} e_k - \sum_{k\geq 1} \langle f^*, e_k \rangle_{\mathcal{D}} e_k \right\|_{\mathcal{D}}^2 \tag{18}$$

$$= \sum_{k\geq 1} (\alpha(1 - e^{-\lambda_k T}) - 1)^2 \langle f^*, e_k \rangle_{\mathcal{D}}^2 \tag{19}$$

The proof is immediate by noting that $1 - e^{-\lambda_k T} \geq 0$ and $\alpha(1 - e^{-\lambda_k T}) \leq 1$ for all k and $0 \leq \alpha \leq 1$. □

Proof of Theorem 4.3. The proof is a straightforward geometric argument in the function space. Note first that Definition 4.1 gives a restiction of where the teacher predictor f_{teacher} can be in the function space with respect to the target f^* and the origin. Note that $\|f_{\text{teacher}} - f^*\|_{\mathcal{D}}^2 \leq \|\alpha f_{\text{teacher}} - f^*\|_{\mathcal{D}}^2$ for all $0 \leq \alpha \leq 1$ holds if and only if $\langle f_{\text{teacher}}, f_{\text{teacher}} - f^* \rangle \leq 0$. Similarlry, Definition 4.1 holds if and only if $\langle f_{\text{student}}, f_{\text{teacher}} - f_{\text{student}} \rangle \leq 0$. This implies that the teacher predictor f_{teacher} is inside the set $f_{\text{teacher}} \in \{f \mid \|f - \frac{f^*}{2}\|_{\mathcal{D}} \leq \frac{1}{2}\|f^*\|_{\mathcal{D}}\}$. Similarly, we have that $f_{\text{student}} \in \{f \mid \|f - \frac{f_{\text{teacher}}}{2}\|_{\mathcal{D}} \leq \frac{1}{2}\|f_{\text{teacher}}\|_{\mathcal{D}}\}$. Both of these sets are spheres in the function space w.r.t. $\mathcal{D}$-norm. Let $\beta \geq 1$ be such that $\tilde{f}_{\text{teacher}} = \beta f_{\text{teacher}}$ and such that $\langle \tilde{f}_{\text{teacher}}, f^* - \tilde{f}_{\text{teacher}} \rangle = 0$. Then we have that

$$\mathcal{L}_{\text{TE}} = 1 - \beta^2 f_{\text{teacher}}^2 + (\beta - 1)^2 f_{\text{teacher}}^2 = 1 - (2\beta - 1) f_{\text{teacher}}^2.$$

Let $\tilde{f}_{\text{student}}$ the predictor inside the student set with the smallest risk, i.e. on the intersection of the line connecting f^* and $\frac{f_{\text{teacher}}}{2}$ and the boundary of the allowed student set. Note then that

$$\mathcal{L}(\tilde{f}_{\text{student}}) = \left(\|f^* - \frac{f_{\text{teacher}}}{2}\|_{\mathcal{D}} - \|\frac{f_{\text{teacher}}}{2}\|_{\mathcal{D}} \right)^2$$

Note further that

$$\|f^* - \frac{f_{\text{teacher}}}{2}\|_{\mathcal{D}}^2 = 1 - \beta^2 f_{\text{teacher}}^2 + (\beta - \frac{1}{2})^2 f_{\text{teacher}}^2 = 1 - (\beta - \frac{1}{4}) f_{\text{teacher}}^2.$$

Therefore, we have that

$$f_{\text{teacher}}^2 = \frac{1 - \mathcal{L}_{\text{TE}}}{2\beta - 1}.$$

From this we get that

$$\mathcal{L}(\tilde{f}_{\text{student}}) = \left(\sqrt{1 - (\beta - \frac{1}{4}) \frac{1 - \mathcal{L}_{\text{TE}}}{2\beta - 1}} - \frac{1}{2} \sqrt{\frac{1 - \mathcal{L}_{\text{TE}}}{2\beta - 1}} \right)^2$$

$$= \frac{1}{8\beta - 4} \left(\sqrt{4\beta - 3 + (4\beta - 1)\mathcal{L}_{\text{TE}}} - \sqrt{1 - \mathcal{L}_{\text{TE}}} \right)^2$$

Note that for $\beta \geq 1$ and $\mathcal{L}_{\text{TE}} \leq 1$ we have that

$$\mathcal{L}(\tilde{f}_{\text{student}}) = \frac{1}{8\beta - 4} \left(\sqrt{4\beta - 3 + (4\beta - 1)\mathcal{L}_{\text{TE}}} - \sqrt{1 - \mathcal{L}_{\text{TE}}} \right)^2$$

$$\geq \frac{1}{2} \left(1 + \mathcal{L}_{\text{TE}} - \sqrt{1 + 2\mathcal{L}_{\text{TE}} - 3\mathcal{L}_{\text{TE}}^2} \right).$$

The last inequality holds by noticiting that the function is increasing in β, so it is minimized for $\beta = 1$ for $0 < \mathcal{L}_{\text{TE}} < 1$.

That is, since $\mathcal{L}_{\text{ST}} \geq \mathcal{L}(\tilde{f}_{\text{student}})$ by design, we have that

$$\mathcal{L}_{\text{ST}} \geq \frac{1}{2} \left(1 + \mathcal{L}_{\text{TE}} - \sqrt{1 + 2\mathcal{L}_{\text{TE}} - 3\mathcal{L}_{\text{TE}}^2} \right) \geq \frac{3}{4} \mathcal{L}_{\text{TE}}^2.$$

The last inequality $\frac{1}{2} \left(1 + \mathcal{L}_{\text{TE}} - \sqrt{1 + 2\mathcal{L}_{\text{TE}} - 3\mathcal{L}_{\text{TE}}^2} \right) \geq \frac{3}{4} \mathcal{L}_{\text{TE}}^2$ holds for $0 \leq \mathcal{L}_{\text{TE}} \leq 1$. □

Proof of Theorem G.1. The proof is similar to the proof of Theorem 4.3. Similarly, we have that the teacher predictor is bounded to the set $f_{\text{teacher}} \in \{f \mid \|f - \frac{f^*}{2}\|_{\mathcal{D}} \leq \frac{1}{2}\|f^*\|_{\mathcal{D}}\}$. The student predictor is in this case bounded to the set $f_{\text{student}} \in \{f \mid \|f\|_{\mathcal{D}} \leq \|f_{\text{teacher}}\|_{\mathcal{D}}\}$. Let $\beta = \langle f^* - f_{\text{teacher}}, f^*\rangle$. Note that from the condition $f_{\text{teacher}} \in \{f \mid \|f - \frac{f^*}{2}\|_{\mathcal{D}} \leq \frac{1}{2}\|f^*\|_{\mathcal{D}}\}$ we have that $\langle f_{\text{teacher}}, f^* - f_{\text{teacher}}\rangle \leq 0$ which implies that there is $\beta \geq 1$ for which we have $\langle \beta f_{\text{teacher}}, f^* - f_{\text{teacher}}\rangle$. Note that $\mathcal{L}_{\text{TE}} = \|f^* - f_{\text{teacher}}\|_{\mathcal{D}}^2$. Therefore, we have that since $\|f^*\|_{\mathcal{D}}^2 = 1$

$$\mathcal{L}_{\text{TE}} = 1 - \beta^2 f_{\text{teacher}}^2 + (\beta - 1)^2 f_{\text{teacher}}^2 = 1 - (2\beta - 1) f_{\text{teacher}}^2.$$

This implies that

$$f_{\text{teacher}}^2 = \frac{1 - \mathcal{L}_{\text{TE}}}{2\beta - 1}.$$

Note that

$$\mathcal{L}_{\text{ST}} \geq (\|f^*\|_{\mathcal{D}} - \|f_{\text{teacher}}\|_{\mathcal{D}})^2 = (1 - \|f_{\text{teacher}}\|_{\mathcal{D}})^2$$

by the same argument as in Theorem 4.3. Therefore, we have that

$$\mathcal{L}_{\text{ST}} \geq \left(1 - \sqrt{\frac{1 - \mathcal{L}_{\text{TE}}}{2\beta - 1}}\right)^2$$

Note that $1 - \sqrt{\frac{1-\mathcal{L}_{\text{TE}}}{2\beta-1}}$ is positive and increasing in β, so the above lower bound is minimized for $\beta = 1$. Therefore

$$\mathcal{L}_{\text{ST}} \geq \left(1 - \sqrt{1 - \mathcal{L}_{\text{TE}}}\right)^2 = 2 - \mathcal{L}_{\text{TE}} - 2\sqrt{1 - \mathcal{L}_{\text{TE}}} \geq \frac{1}{4}\mathcal{L}_{\text{TE}}^2.$$

□

Proof of Corollary G.2. The proof is immediate from Theorem G.1 and the fact that the student $f_{\text{student}}^{(i)}$ is shrinking-optimal w.r.t. $f_{\text{student}}^{(i-1)}$ implies that $\|f_{\text{student}}^{(i)}\|_{\mathcal{D}} \leq \|f_{\text{student}}^{(i-1)}\|_{cD}$, which further implies that $\|f_{\text{student}}^{(i)}\|_{\mathcal{D}} \leq \|f_{\text{teacher}}\|_{\mathcal{D}}$. □

I. Deterministic Equivalent of the Teacher Error

Let $\tilde{\mathcal{L}}(f^*, m)$ be the deterministic equivalent of the test risk of our model with m random features trained on population loss with target f^*. Let s_K be the eigenvalues of the activation function, i.e. the decompositon of the activation function in the eigenbasis $\{e_i\}_{i=1}^{\infty}$. Let β^* be the coefficients of the target f^* in the eigenbasis, $f^* = \sum_{i=1}^{\infty} \beta_i^* e_i$. To write down a closed form of $\tilde{\mathcal{L}}(f^*, m)$, we introduce the following notation $\boldsymbol{S}$ and ν

$$\boldsymbol{S} = \text{diag}(s_1, s_2, \dots)$$
$$m = \text{tr}(\boldsymbol{S}(\boldsymbol{S} + \nu)^{-1}).$$

That is, ν is the unique solution to this equation. According to Corollary 3.5 in Defilippis et al. (2024), the deterministic equivalent of test risk has the following closed form then

$$\tilde{\mathcal{L}}(f^*, m) = \nu\langle \beta^*, (\boldsymbol{S} + \nu)^{-1}\beta^*\rangle. \tag{20}$$

I.1. Lower Bound for Teacher Error

We show a lower bound on the deterministic equivalent of test risk of a Random Feature Model with m features that we consider in our setup. Under the Gaussian Universality Ansatz, the true test error of the teacher's predictor $\mathcal{L}_{\text{TE}}$ will behave like its deterministic equivalent.

Theorem I.1 (Lower bound for the Error of a 2-layer random feature ReLU)**.** *If f^* satisfies Condition 6.1 for Model 2.3 with K such that for some k, $N_0 + \cdots + N_{k-1} + 1 \leq K \leq N_0 + \cdots + N_k$, then for fixed d there exists $\alpha > \frac{d+2}{d-2}$ such that for the deterministic equivalent of the test risk we have*

$$\tilde{\mathcal{L}}(f^*, m) \geq \Theta\left(\frac{1}{m^\alpha} d^2 \|f^*\|_{\mathcal{D}}^2\right).$$

We can take $\alpha = \frac{d+2}{d-2} + o_d(1)$.

Proof of Theorem I.1. Note that the eigenvalues of the activation function are σ_s^2 with multiplicity N_s. We want to lower bound ν. Assume that for large m, $\nu \leq \frac{1}{m^\alpha}$. Note that we have the following two things for $s \gg d$,

$$\sigma_s^2 = \Theta(s^{-(d+2)}) \text{ and } N_s = \Theta(s^{d-2}).$$

The first follows from Lemma F.8 and the second one follows immediately from the closed form of N_s. Note that in this case we have that $s_i = \sigma_i^2$. Note that then for $\sigma_s^2 > \nu$ we have that $\frac{\sigma_s^2}{\sigma_s^2+\nu} > \frac{1}{2}$. Note that if $\nu < \frac{1}{m^\alpha}$ then we have that for all $s < m^{\frac{\alpha}{d+2}}$ that $\sigma_s^2 > \nu$. Note that there is at least $N_{s_{\max}}$ of those, where $s_{\max}$ is the largest such s. We have that $s_{\max} \geq m^{\frac{\alpha}{d+2}}$ But we have that $N_{s_{\max}} = \Theta(s_{\max}^{d-2}) \geq \Theta(m^{\alpha\frac{d-2}{d+2}})$. Note also that with this choice of $s_{\max}$

$$m = \operatorname{tr}(\boldsymbol{S}(\boldsymbol{S}+\nu)^{-1}) = \sum_{i=1}^{\infty} \frac{s_i}{s_i+\nu} = \sum_{s=1}^{\infty}\sum_{i=1}^{N_s} \frac{\sigma_{s,i}^2}{\sigma_{s,i}^2+\nu} \geq N_{s_{\max}} \frac{1}{2} \geq \Theta(m^{\alpha\frac{d-2}{d+2}}),$$

where $\sigma_{s,i} = \sigma_s$.

So if $\alpha = \frac{d+2}{d-2} + o_d(1)$ then we have a contradiction. Therefore $\nu \geq \frac{1}{m^\alpha}$ for this $\alpha = \frac{d+2}{d-2} + o_d(1)$. Plugging this back in we have

$$\tilde{\mathcal{L}}(f^*, m) = \nu \left(\sum_{i=1}^{\infty} \frac{1}{\sigma_i^2+\nu} \sum_{j=1}^{N_i} \beta_{i,j}^{*2} \right) \geq \frac{1}{2}\nu \frac{1}{\sigma_1^2 + \frac{1}{m}} \|f^*\|_{\mathcal{D}}^2.$$

For the last inequality, note first that $\sigma_i \leq \sigma_1$ so $\frac{1}{\sigma_i^2+\nu} \geq \frac{1}{\sigma_1^2+\nu}$. Also note that

$$m = \sum_{i=1}^{\infty} \frac{s_i}{s_i+\nu} \leq \frac{\sum_{i=1}^{\infty} s_i}{\nu},$$

so $\nu \leq \frac{\sum_{i=1}^{\infty} s_i}{m}$. Note that since $\sum_{s=1}^{\infty} \sigma_s^2 N_s = \sigma(1) = 1$ we have that $\sum_{i=1}^{\infty} s_i = 1$. Therefore, it holds that $\nu \leq \frac{1}{m}$. This finishes the proof. □

I.2. ReLU Network Error Upper Bound

Theorem I.2 (Upper Bound on Teacher Error with ReLU Network)**.** *If f^* is spanned by the spherical harmonics of order that is even and at mos k or 1, for Model 2.2 in fixed dimension d, we have the following bound on the deterministic equivalent of the test risk*

$$\tilde{\mathcal{L}}(f^*, m) \leq \Theta\left(\frac{1}{m} d^{(k+1)} \|f^*\|_{\mathcal{D}}^2 \right)$$

Proof of Theorem I.2. Note that

$$m = \operatorname{tr}(\boldsymbol{S}(\boldsymbol{S}+\nu)^{-1}) = \sum_{i=1}^{\infty} \frac{s_i}{s_i+\nu} \leq \frac{\sum_{i=1}^{\infty} s_i}{\nu}$$

so we have $\nu \leq \frac{\sum_{i=1}^{\infty} s_i}{m}$. Since $\sum_{i=1}^{\infty} s_i = \sigma(1) = 1$, we have that $\nu \leq \frac{1}{m}$. Then we have that since f^* is spanned by harmonics of order at most k

$$\tilde{\mathcal{L}}(f^*, m) = \nu \left(\sum_{i=1}^{\infty} \frac{1}{\sigma_i^2+\nu} \sum_{j=1}^{N_i} \beta_{i,j}^{*2} \right) = \nu \sum_{i=1}^{k} \frac{1}{\sigma_i^2+\nu} \sum_{j=1}^{N_s} \beta_{i,j}^{*2} \leq \frac{\nu}{\sigma_k^2+\nu} \sum_{i=1}^{k}\sum_{j=1}^{N_i} \beta_i^{*2} = \frac{\nu}{\sigma_k^2+\nu} \|f^*\|_{\mathcal{D}}^2.$$

Finally, note that $\sigma_k^2 = \Theta(d^{-(k+1)})$ so we have $\frac{\nu}{\nu+\sigma_k^2} \leq \frac{1}{m}\Theta(d^{(k+1)})$ □

Remark I.3. The same proof shows that in the Gaussian Features Model B.1 if the covairance structure is fixed we cannot have a $\Theta(1)$-error asymptotic for any function that is learnable by the model as $M_{\text{TE}} \to \infty$ in fixed dimension.

$\Theta(1)$-error asymptotics Theorem 3.2 shows a quadratic gap between the student and teacher errors, but both errors vanish as the model size M_{TE} increases. Perhaps a more interesting and relevant scaling it a proportional scaling where with larger model sizes we can handle increasingly more complex problems while still ensuring bounded error. A more variant of the seperation result shows that for any asymptotic error, we can have a quadratic error gap even as model size and dimensinoality jointly increase:

Theorem I.4 ($\Theta(1)$-error asymptotics for Linear Networks)**.** *For any $\alpha > 0$ there exists $d_\alpha \in \mathbb{N}$ such that for all $d > d_\alpha$, Model 2.3 with $f^* = x_1$, $\Psi_d = (1, \sqrt{\frac{\alpha}{d-1}}, \ldots, \sqrt{\frac{\alpha}{d-1}})$, and $M_{\mathrm{TE}} = \sqrt{\frac{d-1}{\alpha}}$, probability at least 0.99 we have*

$$\mathcal{L}_{\mathrm{TE}} > 0.98\frac{\alpha}{1+\alpha} \text{ and } \mathcal{L}_{\mathrm{ST}} \le 100\mathcal{L}_{\mathrm{TE}}^2.$$

The proof can be found in Appendix D.2.

J. Experiment Details

In this section, we provide experiment details for simulating weak-to-strong generalization for both 2-layer ReLU networks (Figure 2) and linear networks (Figure 3). All experiments are conducted on one H100 GPU, and are finished within 24 GPU hours.

To avoid the effect of randomness, experiments are repeated 5 times with different random seeds, 95% confidence intervals are shown in Figure 2(a) and Figure 3(a).

For curve fitting in Figure 2(b) and Figure 3(b), we use `numpy.polyfit` with degree 1 after taking the log of teacher loss $\mathcal{L}_{\mathrm{TE}}$ and student loss $\mathcal{L}_{\mathrm{ST}}$.

For linear network, we simulate the limit case where student size $M_{\mathrm{ST}} \to \infty$. The simulation follows Equation (9) with $\lambda_1 = 1$, $\lambda_i = (d-1)^{-2/3}$ for $i > 1$. Note this method is not viable for 2-layer ReLU network because λ_i is long-tailed for 2-layer ReLU network. Therefore, we set $M_{\mathrm{ST}} = 16384$ in Figure 2.

K. Additional Experimental Results

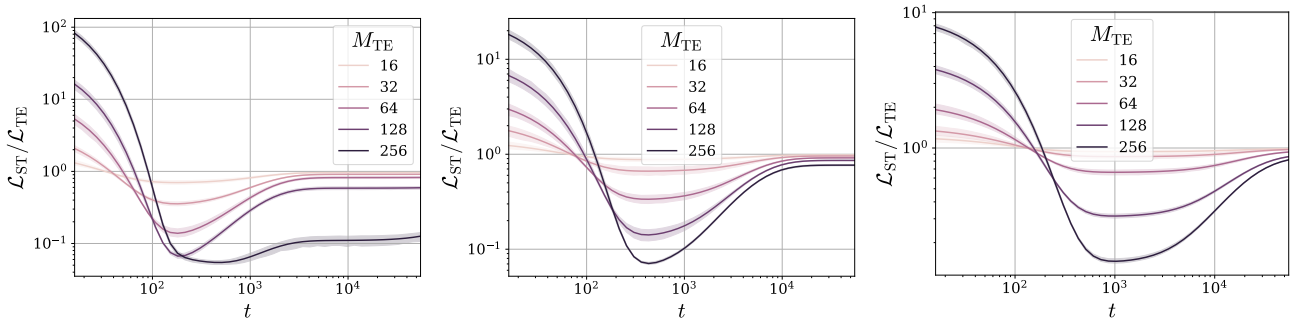
K.1. 2-Layer ReLU Network Experiments

Figure 4 shows full experimental results for Model 2.2. It additionally includes $d = 16$ and 64 compared to Figure 2. We observe that when $d = 64$, the student loss $\mathcal{L}_{\mathrm{ST}}$ is polynomially smaller than the teacher loss $\mathcal{L}_{\mathrm{TE}}$, with a higher estimated exponent compared to $d = 32$. When $d = 16$, the curve suggests a sudden change of optimal early stopping time, which we believe is caused by lack of student size M_{ST}.

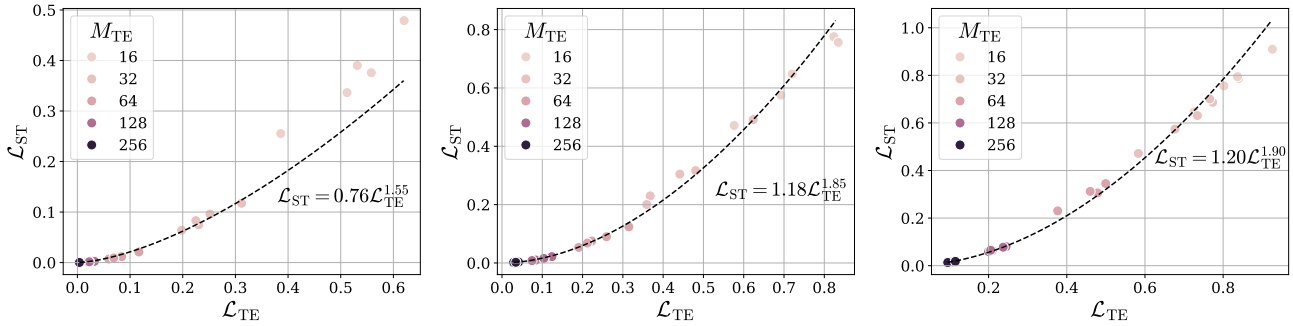
However, due to GPU memory limit, we are not able to increase M_{ST}. Thus we conduct ablation study with smaller M_{ST} to investigate the impact of M_{ST}. In Figure 5, we compare the loss ratio curve between $M_{\mathrm{ST}} = 8192$ and $M_{\mathrm{ST}} = 16384$. We observe larger gap between the two when teacher loss $\mathcal{L}_{\mathrm{TE}}$ is smaller, i.e., for small d and large M_{TE}. Therefore, it is not hard to believe that $\mathcal{L}_{\mathrm{ST}} = \tilde{\Theta}(\mathcal{L}_{\mathrm{TE}}^2)$ for this setting as well. We leave it as future work.

K.2. Linear ReLU Network Experiments

Figure 6 shows full experimental results for Model 2.3. It shows the same conclusions as Figure 3, but with $k = 10$ and 100, providing additional verification of Theorem 3.2.

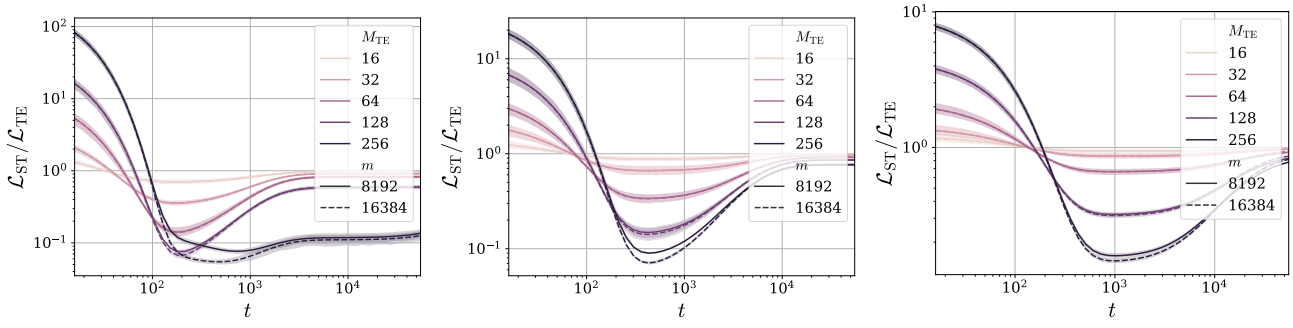


(a) $\mathcal{L}_{\rm ST}/\mathcal{L}_{\rm TE}$ v.s. t at $d = 16$ (b) $\mathcal{L}_{\rm ST}/\mathcal{L}_{\rm TE}$ v.s. t at $d = 32$ (c) $\mathcal{L}_{\rm ST}/\mathcal{L}_{\rm TE}$ v.s. t at $d = 64$



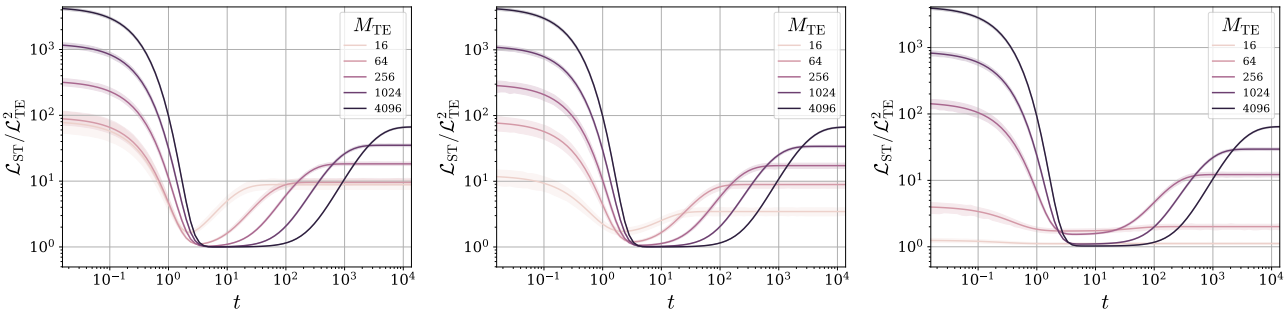
(d) $\mathcal{L}_{\rm ST}$ v.s. $\mathcal{L}_{\rm TE}$ at $d = 16$ (e) $\mathcal{L}_{\rm ST}$ v.s. $\mathcal{L}_{\rm TE}$ at $d = 32$ (f) $\mathcal{L}_{\rm ST}$ v.s. $\mathcal{L}_{\rm TE}$ at $d = 64$

Figure 4. Weak-to-strong generalization happens in 2-layer ReLU networks with input dimension $d = 16, 32, 64$, student size $M_{\rm ST} = 16384$, and teacher size $M_{\rm TE} \in \{16, \ldots, 256\}$. We consider target function f^* be a linear function, i.e., $f^* = \langle \beta, \boldsymbol{x} \rangle$ for some β of unit norm. The top figures plots the ratio between student loss $\mathcal{L}_{\rm ST}$ and teacher loss $\mathcal{L}_{\rm TE}$, with varying $M_{\rm TE}$ and gradient flow training time t. In bottom figures, we fit student loss $\mathcal{L}_{\rm ST}$ as a power law function of $\mathcal{L}_{\rm TE}$. The empirical observations align with Theorem 3.1.



(a) $M_{\rm ST} = 8192$, $d = 16$ (b) $M_{\rm ST} = 8192$, $d = 32$ (c) $M_{\rm ST} = 8192$, $d = 64$

Figure 5. Abalation of weak-to-strong generalization in 2-layer ReLU networks. We use same setting as Figure 4 and compare the results with smaller student size $M_{\rm ST}$.



(a) $\mathcal{L}_{\mathrm{ST}}/\mathcal{L}_{\mathrm{TE}}^2$ v.s. t at $k = 1$ (b) $\mathcal{L}_{\mathrm{ST}}/\mathcal{L}_{\mathrm{TE}}^2$ v.s. t at $k = 10$ (c) $\mathcal{L}_{\mathrm{ST}}/\mathcal{L}_{\mathrm{TE}}^2$ v.s. t at $k = 100$

Figure 6. Weak-to-Strong generalization happens in antitropic random linear feature networks (Model 2.3). Here we used an input distribution as in Theorem 3.2, with $k = 1, 10, 100$ and a target function $f^* = \langle \beta, \boldsymbol{x} \rangle$ where β is a unit norm vector with non-zero values for first k coordinates. The figures the ratio between the student loss $\mathcal{L}_{\mathrm{ST}}$ and squared teacher loss $\mathcal{L}_{\mathrm{TE}}^2$, with varying teacher size M_{TE}, and where the dimensionality $d = M_{\mathrm{TE}}^{3/2} + k$ as set as in the scaling of Theorem 3.2, as a function of the gradient flow time t. With proper early stopping time $\mathcal{L}_{\mathrm{ST}}/\mathcal{L}_{\mathrm{TE}}^2$ converges to approximately 1 as M_{TE} grows, confirming that for large M_{TE} we have $\mathcal{L}_{\mathrm{ST}} \propto \mathcal{L}_{\mathrm{TE}}^2$ as in Theorem 3.2.