
On the Mechanisms of Weak-to-Strong Generalization: A Theoretical Perspective

Behrad Moniri

University of Pennsylvania
Philadelphia, PA
bemoniri@seas.upenn.edu

Hamed Hassani

University of Pennsylvania
Philadelphia, PA
hassani@seas.upenn.edu

Abstract

Weak-to-strong generalization—where a student model trained on imperfect labels generated by a weaker teacher nonetheless surpasses that teacher—has been widely observed, but the mechanisms that enable it have remained poorly understood. In this paper, through a theoretical analysis of simple models, we uncover three core mechanisms that can drive this phenomenon. First, by analyzing ridge linear regression, we study the interplay between the teacher and student regularization parameters and prove that a student can compensate for a teacher’s under-regularization and achieve lower test error. We also analyze the role of the parameterization regime of the models and show that qualitatively different phenomena can happen in different regimes. Second, by analyzing weighted ridge linear regression, we show that a student model with a regularization structure better aligned to the target function, can outperform its teacher. Third, in a non-linear multi-index learning setting, we demonstrate that a student can learn easy, task-specific features from the teacher while leveraging its own broader pre-training to learn hard-to-learn features that the teacher cannot capture.

1 Introduction

Weak-to-strong generalization refers to the phenomenon where a strong (student) model trained on data produced by a weak (teacher) model can sometimes significantly surpass the teacher’s performance. This concept was first introduced by [Burns et al. \[2024\]](#), where the authors fine-tuned the GPT-2 model [\[Radford et al., 2019\]](#) (the teacher) for a specific task using ground-truth labels, subsequently employing the fine-tuned model to generate synthetic samples for the same task. These synthetic samples were then used to fine-tune GPT-4 [\[Achiam et al., 2023\]](#) (the student). Remarkably, the fine-tuned student model outperformed its teacher in certain settings despite having access only to the imperfect synthetic data generated by the teacher.

Weak-to-strong generalization is an especially important phenomenon from a practical perspective because of its implications for the emerging question of *superalignment* [\[OpenAI, 2023\]](#); i.e., can humans steer models with potentially superhuman capabilities to become aligned to human norms and values [\[Burns et al., 2024\]](#)? Considering the weak model as a proxy for humans, the possibility of the weak-to-strong generalization phenomenon suggests that the answer can be affirmative.

Despite its practical importance, the mechanisms that enable weak-to-strong generalization are still not fully understood. Regularization has empirically been shown to play a critical role in enabling weak-to-strong generalization. However, despite recent theoretical progress demonstrating that regularizing the student is necessary in some settings [\[Medvedev et al., 2025\]](#), the full picture of the effects and the interplay of the regularization of both the student and teacher models in weak-to-strong generalization is still unclear. For example, prior work on weak-to-strong generalization mainly focus on ridgeless regression (see e.g., [Dong et al. \[2025\]](#), [Xue et al. \[2025\]](#), [Ildiz et al. \[2025\]](#), etc.).

Additionally, prior work assumes that the teacher and student models have frozen representations, and only a linear head is trained through a convex objective (see e.g., [Ildiz et al. \[2025\]](#), [Medvedev et al. \[2025\]](#), [Dong et al. \[2025\]](#), [Xue et al. \[2025\]](#), [Charikar et al. \[2024\]](#), etc.). However, fine-tuning can in practice go beyond this linearized regime and update model features as well. [Burns et al. \[2024\]](#) empirically demonstrated that updating the features yields substantially stronger weak-to-strong gains than only updating a linear head. For these cases, a linearized theoretical model might not suffice to capture all the relevant phenomena. This motivates a theoretical study beyond the linearized regimes and an analysis of the role of feature learning on weak-to-strong generalization.

To take steps towards better understanding these aspects of training on weak-to-strong generalization, in this paper, we conduct a thorough theoretical study of this phenomenon in prototypical theoretical models. For the linear setting, we let the student and the teacher be high-dimensional standard and weighted ridge regression models and study how the explicit regularizations of the teacher and student models affect weak-to-strong generalization. We also investigate the role of the parameterization regime of the models. As the choice of regularization, we consider ridge and weighted ridge penalties. For the nonlinear case, we consider the problem of learning from multi-index models where the models learn relevant features through a non-convex optimization objective. By studying this setting, we characterize how knowledge propagates between the models.

1.1 Contributions

Here we discuss the main contributions of the paper. We characterize three mechanisms that can enable weak-to-strong generalization.

- In Section 2.1, we consider a setting where the student and the teacher are trained with ridge regression. We fully characterize the test error of the models in the high-dimensional proportional regime by deriving asymptotic expressions for the test errors. Using these expressions, we study the conditions where the student model outperforms the teacher. We show that the student model can outperform the teacher by *adequately compensating the under-regularization of the teacher*. We further prove that different parameterization regimes of the student model can result in qualitatively different phenomena.
- In Section 2.2, we consider a setting where the teacher is again trained with ridge regression. However, we train the student with a *weighted* ridge regularization. We again fully characterize the limiting test errors of the models in the high-dimensional proportional limit, and show that weak-to-strong generalization can happen *when the regularization structure of the student is better suited for the task*.
- In Section 3, we study learning from a nonlinear multi-index learning function that can be decomposed to a mix of *easy*- and *hard*-to-learn components by applying a single step of gradient descent on the first layer weights of two-layer neural networks. We assume that the easy component is highly specialized and task-specific, but, the hard component is a component shared across many tasks. We consider a case where the strong student model has already learned the hard component during pre-training, but not the easy (yet task-specific) component. We show that even if the teacher model is not able to learn the hard components on its own, a pre-trained student can learn the easy component from the teacher while still *retaining* the knowledge from pre-training for the hard component.

1.2 Related Works

The machine learning community has shown growing interest in weak-to-strong generalization. In this section, we review these results.

Theoretical Results. Prior work examines scenarios in which both the student and teacher rely on fixed, pre-trained feature representations. [Wu and Sahai \[2024\]](#) analyze a stylized classification task under an over-parameterized spiked-covariance model with Gaussian covariates, where the teacher model does not have the capability to fit the target function, and the student has a structure that is better aligned with the target. [Ildiz et al. \[2025\]](#) investigate weak-to-strong generalization for high-dimensional ridgeless regression. Building on this line and in a similar setting, [Dong et al. \[2025\]](#), [Xue et al. \[2025\]](#) study how mismatches between student and teacher features affect generalization: [Dong et al. \[2025\]](#) focus on a ridgeless, variance-dominated linear regime in which both models have negligible bias and show that weak-to-strong transfer occurs when the student’s features have lower intrinsic dimension. [Xue et al. \[2025\]](#) consider the same setting and propose that the overlap

between the subspace of features that teacher model has not learned, and the subspace of features that the student model has learned during pre-training govern weak-to-strong generalization. In contrast, this paper analyzes linear models in the high-dimensional proportional regime, covering both under- and over-parameterized cases, where models can have large bias. We also explicitly investigate the role of regularization on weak-to-strong generalization. Complementing these model-centric studies, [Shin et al. \[2025\]](#) take a data-centric approach and study the aspects of data enable weak-to-strong generalization.

Relatedly, [Medvedev et al. \[2025\]](#) consider two-layer neural networks with random first-layer weights (random-features models) and show that, when the student is much wider than the teacher, early stopping is essential for weak-to-strong generalization. However, they assume the teacher is already optimally trained and do not analyze the role of its training. [Charikar et al. \[2024\]](#) propose that the erroneous knowledge that the strong model does not obtain from the weak model characterizes how much the strong model improves over the weak model.

Empirical Studies. Following the pioneering work of [Burns et al. \[2024\]](#), different variants and applications of weak-to-strong generalization have been studied. [Bansal et al. \[2025\]](#), [Yang et al. \[2024\]](#) let the weak model generate data with chain-of-thought to supervise the student models. [Ji et al. \[2024\]](#), [Tao and Li \[2024\]](#) use weak-to-strong generalization for the problem of alignment. [Guo et al. \[2024\]](#) study this phenomenon in vision foundation models. [Liu and Alahi \[2024\]](#) propose a hierarchical mixture of experts method to boost weak-to-strong generalization. [Mulgund and Pabbaraju \[2025\]](#) characterize the gain in performance of the student model over the teacher model in terms of the misfit between the models.

1.3 Notation

We denote vector quantities by **bold** lower-case, and matrix quantities by **bold** upper-case. We use $\|\cdot\|_{\text{op}}$, $\|\cdot\|_{\text{Fr}}$ to denote the operator (spectral) and Frobenius norms. Given an indexed set of vectors $\{\mathbf{x}_i\}_{i=1}^n$, we use the upper case to denote the (row-wise) stacked matrix, e.g. $\mathbf{X} \triangleq [\mathbf{x}_1 \ \cdots \ \mathbf{x}_n]^\top$. Throughout the paper, we use the standard asymptotic notation $o(\cdot)$, $O(\cdot)$, $\Omega(\cdot)$, $\Theta(\cdot)$. Finally, we use $\rightarrow_{\mathbb{P}}$ to denote convergence in probability.

2 The Linearized Case

During the fine-tuning of pre-trained large-scale models, the training dynamic often falls into a kernel regime where the features are not evolved [[Wei et al., 2022](#), [Malladi et al., 2023](#)]. Motivated by these observations, in this section we cast the fine-tuning problem as a linear regression problem over Gaussian features. We aim to analyze the role of student and teacher *regularization*, and also the *parameterization regimes* of the models in weak-to-strong generalization.

Assume that the teacher model has access to n_t independent samples $\mathcal{S}_t = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n_t}$ drawn according to

$$\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_x}), \quad y_i = \beta_\star^\top \mathbf{x}_i + \varepsilon_i \quad (1)$$

where $\beta_\star \in \mathbb{R}^{d_x}$ is an unknown target vector, and $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ is an independent additive noise. The teacher model $\hat{f}_t : \mathbb{R}^{d_x} \rightarrow \mathbb{R}$ is fit on the features $\{\mathbf{x}_i\}_{i=1}^{n_t}$ using these labeled samples. The teacher is then used to generate synthetic labels for $n_s \in \mathbb{N}$ unlabeled covariates $\mathcal{S}_s = \{\tilde{\mathbf{x}}_i\}_{i=1}^{n_s}$ drawn independently from the same distribution according to $\tilde{\mathbf{x}}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_x})$ as $\tilde{y}_i = \hat{f}_t(\tilde{\mathbf{x}}_i)$. These samples are then used to train the student model $\hat{f}_s : \mathbb{R}^{d_x} \rightarrow \mathbb{R}$.

We focus on the following two settings, each showcasing a different mechanism that can enable weak-to-strong generalization.

Setting 1: Ridge Regression. We train the teacher $\hat{f}_t(\mathbf{x}) = \hat{\beta}_t^\top \mathbf{x}$ and the student $\hat{f}_s(\mathbf{x}) = \hat{\beta}_s^\top \mathbf{x}$ using (standard) ridge regression. We prove that a properly regularized student can outperform the teacher, in the case where the regularization parameter of the teacher is set to be smaller than the optimal regularization parameter. This is an example of weak-to-strong generalization through *adequately compensating under-regularization*. Furthermore, we show that *two qualitatively different scenarios* can arise depending whether the student model is *over- or under-parametrized*. This aligns with [Burns et al. \[2024\]](#) and [Medvedev et al. \[2025\]](#), which show that student regularization is necessary for weak-to-strong generalization; we extend their work by analyzing the role of teacher

regularization, and a finer-grained analysis of student regularization and model parameterization, revealing new phenomena.

Setting 2: Weighted Ridge Regression. We train $\hat{f}_t(\mathbf{x}) = \beta_t^\top \mathbf{x}$ using (standard) ridge regression and $\hat{f}_s(\mathbf{x}) = \beta_s^\top \mathbf{x}$ using weighted ridge regression [Hoerl and Kennard, 1970, Casella, 1980]. We show that the strong student model can *leverage better regularization structure* and outperform the weak teacher, *even if* the regularization parameter for the teacher is tuned optimally. We argue that a student model can have a more suitable regularization either by using an architecture that is better tailored to the task or by benefiting from more effective pre-training.

We consider growing n_s, n_t, d_x following the high-dimensional limit. Although our results are proven for this asymptotic regime, through numerical experiments, we show that they still match simulations very well, even for moderately large values of n_s, n_t, d_x .

Assumption 1. Assume that n_t, n_s and d_x all tend to infinity with a proportional rate; i.e.,

$$d_x/n_s \rightarrow \gamma_s > 0, \quad \text{and} \quad d_x/n_t \rightarrow \gamma_t > 0.$$

In this high-dimensional limit, we characterize the test errors achieved by the teacher and student models given by

$$\mathcal{L}_t = \mathbb{E}_{\mathbf{x}, y} \left(y - \hat{\beta}_t^\top \mathbf{x} \right)^2 = \sigma_\varepsilon^2 + \|\hat{\beta}_t - \beta_\star\|_2^2 \quad (2)$$

$$\mathcal{L}_s = \mathbb{E}_{\mathbf{x}, y} \left(y - \hat{\beta}_s^\top \mathbf{x} \right)^2 = \sigma_\varepsilon^2 + \|\hat{\beta}_s - \beta_\star\|_2^2$$

where $(\mathbf{x}, y)$ is an independent test sample drawn from (1). We then use these characterizations to study the conditions under which the student model outperforms the teacher.

2.1 Setting 1: High-Dimensional Ridge Regression

In this section, we assume that the teacher fits a linear regression model $\hat{f}_t(\mathbf{x}) = \hat{\beta}_t^\top \mathbf{x}$ trained on the samples $\mathcal{S}_t$, and is given by

$$\hat{\beta}_t = \operatorname{argmin}_{\beta \in \mathbb{R}^{d_x}} \left[\frac{1}{n_t} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{S}_t} (y_i - \beta^\top \mathbf{x}_i)^2 + \lambda_t \|\beta\|_2^2 \right] \quad (3)$$

where $\lambda_t \in \mathbb{R}$ is the teacher ridge regularization parameter. The student is also a linear model $\hat{f}_s(\mathbf{x}) = \hat{\beta}_s^\top \mathbf{x}$ trained on fresh samples $\mathcal{S}_s$ labeled by the teacher model, and is given by

$$\hat{\beta}_s = \operatorname{argmin}_{\beta \in \mathbb{R}^{d_x}} \left[\frac{1}{n_s} \sum_{\tilde{\mathbf{x}}_i \in \mathcal{S}_s} \left(\beta^\top \tilde{\mathbf{x}}_i - \hat{\beta}_t^\top \tilde{\mathbf{x}}_i \right)^2 + \lambda_s \|\beta\|_2^2 \right] \quad (4)$$

in which $\lambda_s \in \mathbb{R}$ is the student regularization parameter. We characterize the test error of these models in the high-dimensional proportional limit of Assumption 1. Our characterization of the test errors $\mathcal{L}_s, \mathcal{L}_t$ will be in terms of the following quantities from the random matrix theory literature (see e.g., Bai and Silverstein [2010]).

Definition 2. Let $m(\lambda; \gamma)$ be the Stieltjes transform of the Marchenko-Pastur law with parameter γ evaluated at $-\lambda$; i.e.,

$$m(\lambda; \gamma) = \int \frac{d\mu_{\text{MP}(\gamma)}(s)}{s + \lambda} = -\frac{1}{2\gamma\lambda} \left[1 - \gamma + \lambda - \sqrt{(1 + \gamma + \lambda)^2 - 4\gamma} \right].$$

Also, for $p \in \{s, t\}$, we define $m_{p,1} = m(\lambda_p, \gamma_p)$ and $m_{p,2} = -\frac{\partial m}{\partial \lambda} \big|_{\lambda_p, \gamma_p}$.

The test error of $\hat{\beta}_t$ in the high-dimensional proportional limit has been studied extensively in the literature [Tulino and Verdú, 2004, Dobriban and Wager, 2018, Hastie et al., 2022]. The following proposition characterizes the test error of $\hat{\beta}_t$ in our setting.

Proposition 3. Under the condition that $\beta_\star \sim \mathcal{N}(\mathbf{0}, d_x^{-1} \mathbf{I}_{d_x})$ independent of other sources of randomness in the problem, in the high-dimensional proportional limit of Assumption 1, we have

$$\mathcal{L}_t \rightarrow_{\mathbb{P}} \sigma_\varepsilon^2 + (\lambda_t - \sigma_\varepsilon^2 \gamma_t) \lambda_t m_{t,2} + \sigma_\varepsilon^2 \gamma_t m_{t,1},$$

where $m_{t,1}$ and $m_{t,2}$ are defined in Definition 2.

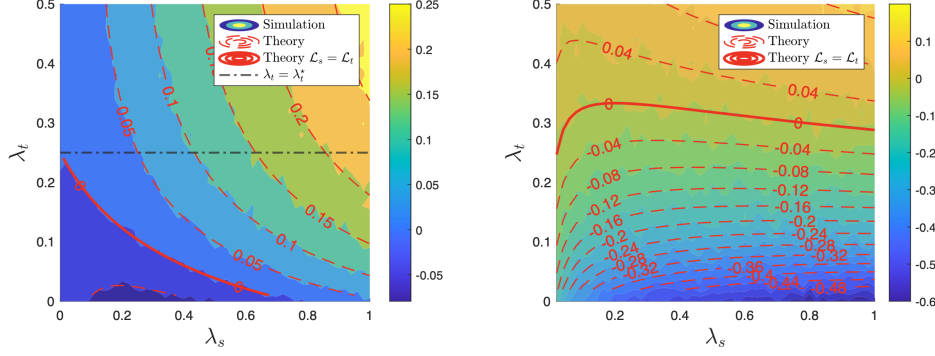


Figure 1: Test-error difference $\mathcal{L}_s - \mathcal{L}_t$ as a function of (λ_t, λ_s) in the setting of Section 2.1. Filled contours are numerical simulations, and the dashed red contours follow the expressions of Theorem 4. The solid curve marks $\mathcal{L}_s = \mathcal{L}_t$, and the dashed black curve is $\lambda_t = \lambda_t^*$. **Left:** under-parameterized student. **Right:** over-parameterized student. See Section 4 for more details.

In the following theorem, we study test error of the student model $\hat{\beta}_s$.

Theorem 4. *Under the same assumptions as Proposition 3, the test errors of $\hat{\beta}_s$ and $\hat{\beta}_t$ satisfy*

$$\mathcal{L}_s - \mathcal{L}_t \rightarrow_{\mathbb{P}} \Delta := (\sigma_\varepsilon^2 \gamma_t - \lambda_t) [(m_{t,1} - \lambda_t m_{t,2}) (\lambda_s^2 m_{s,2} - 2\lambda_s m_{s,1})] + \lambda_s^2 m_{s,2} (1 - \lambda_t m_{t,1}).$$

where $m_{t,1}, m_{t,2}, m_{s,1}, m_{s,2}$ are defined in Definition 2.

This theorem fully characterizes the limit of the test error of the student model in the high dimensional proportional limit. The formulas for the limiting errors derived in this theorem can be used to make numerical predictions for $\mathcal{L}_s - \mathcal{L}_t$. Figure 1 shows an example, supporting that the theoretical predictions of Theorem 4 match very well with simulations even for moderately large d, n_s, n_t . See Section 4 for more details on the experimental settings. We use this theorem to study the test error of the models as a function of overparameterization in Section E.

In the next theorem, we use the formula for the limiting value of $\mathcal{L}_s - \mathcal{L}_t$ from Theorem 4 to study the conditions on $\gamma_s, \gamma_t, \lambda_s, \lambda_t, \sigma_\varepsilon^2$ under which the student model outperforms the teacher; i.e., the conditions of weak-to-strong generalization.

Theorem 5. *Under the conditions of Theorem 4, the (limiting) test errors of the student and teacher models satisfy the following:*

- If $\lambda_t \geq \sigma_\varepsilon^2 \gamma_t$, we have $\mathcal{L}_s \geq \mathcal{L}_t$.
- If $\lambda_t < \sigma_\varepsilon^2 \gamma_t$, two cases can happen:
 - If $0 < \gamma_s < 1$, there exists $\bar{\lambda} \geq 0$ such that $\mathcal{L}_s < \mathcal{L}_t$ for all $\lambda_s \in (0, \bar{\lambda})$.
 - If $\gamma_s > 1$ and if the parameters $\gamma_t, \gamma_s, \lambda_t, \sigma_\varepsilon$ satisfy

$$\frac{\lambda_t - \gamma_t \sigma_\varepsilon^2}{\sqrt{(1 + \gamma_t + \lambda_t)^2 - 4\gamma_t}} > \frac{1}{1 - 4\gamma_s - 4\sqrt{\gamma_s^2 - \gamma_s}}, \quad (5)$$

then there exists $\bar{\lambda}_-, \bar{\lambda}_+ \geq 0$ such that $\mathcal{L}_s < \mathcal{L}_t$ for all $\lambda_s \in (\bar{\lambda}_-, \bar{\lambda}_+)$. Moreover, if (5) does not hold, we have $\mathcal{L}_s \geq \mathcal{L}_t$.

Note that under the setting of this section, the optimal ridge regularization parameter for the weak model is known to be equal to $\lambda_t^* = \sigma_\varepsilon^2 \gamma_t$ [Dobriban and Wager, 2018, Theorem 2.1]. Theorem 5 states that if the teacher is over-regularized ($\lambda_t \geq \lambda_t^*$), the student can never outperform it. In the case that the teacher is under-regularized ($\lambda_t < \lambda_t^*$), the parameterization regime of the student model γ_s plays a key role. In particular, if $\gamma_s < 1$ (i.e., the student is under-parameterized), the student model can outperform the teacher by further regularization as long as $0 < \lambda_s < \bar{\lambda}$. However, if $\gamma_s > 1$ (i.e., the student is over-parameterized), as long as (5) holds, λ_s should be larger than a certain threshold for it to outperform the teacher. Otherwise, the student will always have a worse performance compared to the teacher.

The phase transitions predicted in Theorem 5 can be seen in Figure 1, where for each (λ_t, λ_s) pair, we plot the contours of $\mathcal{L}_s - \mathcal{L}_t$ for a given $\gamma_s, \gamma_t, \sigma_\varepsilon$. In these plots, the solid red curves show the pairs (λ_w, λ_s) for which $\mathcal{L}_s = \mathcal{L}_t$. The left plot corresponds to the case where $\gamma_s < 1$. It can be seen that when $\lambda_t < \lambda_t^*$, the student model outperforms the teacher as long $\lambda_s < \bar{\lambda}(\lambda_t; \gamma_s, \gamma_t, \sigma_\varepsilon)$. Moreover, the student is always worse than the teacher when $\lambda_t > \lambda_t^*$. The right plot corresponds to the case with $\gamma_s > 1$. In this case, it is seen that as predicted in Theorem 5, for some values of λ_t , the student outperforms the teacher only if $\lambda_s \in (\bar{\lambda}_-, \bar{\lambda}_+)$ for some $0 < \bar{\lambda}_- < \bar{\lambda}_+$. See Section 4 for more details on the experimental setting.

Mechanism of Weak-to-Strong Generalization. In this section, we show that the student’s reduced error stems from *compensating for the teacher’s insufficient regularization*. Thus, intuitively, similar to what is proven in Theorem 5, when the teacher is already over-regularized, the student is unable to achieve a better performance by leveraging this mechanism. Also, note that when $\gamma_s > 1$, the student model is over-parameterized and as a result, some information is lost. Thus, more regularization is required for the student to outperform the teacher. This can be seen as the reason why in this regime, a non-zero lower bound exists for λ_s to ensure this. Our results complement the results of [Medvedev et al. \[2025\]](#) who demonstrated that regularizing the student is essential to avoid overfitting to the mistakes in a setting where the teacher is optimally trained.

Non-Monotone Student Test Error Curves. Studying the test error $\mathcal{L}_s$ as a function of the teacher overparameterization γ_t using Theorem 4, we observe that student model also exhibits the *double descent* phenomenon, where we can see a second bias-variance tradeoff in the test error beyond the interpolation limit [[Belkin et al., 2019](#)]. See Section E for more details. This is in line with findings in different linear regression setups such as standard ridge/ridgeless regression [[Hastie et al., 2022](#), [Nakkiran et al., 2021](#)] (which corresponds to double descent in the teacher model in our setting), ridge regression with correlated samples [[Atanasov et al., 2024](#), [Moniri and Hassani, 2025](#)], and weighted ridge regression [[Wu and Xu, 2020](#)].

Universality. Similar to other high-dimensional ridge regression problems (see e.g., [Hastie et al. \[2022\]](#), [Hu and Lu \[2023\]](#), [Montanari and Saeed \[2022\]](#), etc.), we can use a simple Lindeberg exchange argument [[Lindeberg, 1922](#), [Korada and Montanari, 2011](#)] to show that Δ from Theorem 4 exhibits universality; i.e., the actual distribution of the samples does not matter and the limiting value of $\mathcal{L}_s - \mathcal{L}_t$ is determined only by the first and second moment of the covariates. Thus, our results in theory do hold much more broadly than the Gaussian case. The study of this universality is not central to our discussion and we leave it as future work.

Other Related Work. The high-dimensional ridge regression setting considered in this section is related to the linear regression setting considered by [Dohmatob et al. \[2024\]](#) to study model collapse. However, in their setting, only the downstream model (which corresponds to the student model in our setting) has a non-zero ridge regularization. Similar settings have also been studied in the self-distillation literature (see e.g., [Das and Sanghavi \[2023\]](#), [Pareek et al. \[2024\]](#), etc.). However, the training procedure of the models are different; e.g. in their setting, the teacher generates synthetic labels for its own training set and not for a fresh set of covariates. Additionally, in the self distillation setting, the student model still has access to ground truth labels.

2.2 Setting 2: High-Dimensional Weighted Ridge Regression

In this section, we consider a setting where the strong model is a linear model trained using weighted ridge regression [[Hoerl and Kennard, 1970](#), [Casella, 1980](#), [Wu and Xu, 2020](#), [Richards et al., 2021](#)]. Given samples $\mathcal{S} \subset \mathbb{R}^{d_x} \times \mathbb{R}$, the estimator $\text{WRidge}(\mathcal{S}, \lambda, \mathbf{\Gamma})$ is defined as

$$\text{WRidge}(\mathcal{S}, \lambda, \mathbf{\Gamma}) := \underset{\beta \in \mathbb{R}^{d_x}}{\operatorname{argmin}} \left[\frac{1}{n} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{S}} (y_i - \beta^\top \mathbf{x}_i)^2 + \lambda \|\mathbf{\Gamma}^{-1} \beta\|_2^2 \right], \quad (6)$$

where $\mathbf{\Gamma} \in \mathbb{R}^{d_x \times d_x}$ is a weighting matrix and $\lambda \in \mathbb{R}$ is a scalar. In this section, we assume that the teacher is still a (standard) ridge regression estimator. However, unlike the previous section, we let the student model be a weighted ridge estimator; i.e.,

$$\hat{\beta}_t = \text{WRidge}(\mathcal{S}_t, \lambda_t, \mathbf{I}_{d_x}), \quad \hat{\beta}_s = \text{WRidge}(\mathcal{S}_s, \lambda_s, \mathbf{\Gamma}). \quad (7)$$

In this model, the matrix $\mathbf{\Gamma}$ is assumed to be given and fixed. The matrix $\mathbf{\Gamma}$ determines the structure of the student regularization enforcing different levels of regularization in different directions. In the next remark, we provide a linear neural-network interpretation for $\mathbf{\Gamma}$.

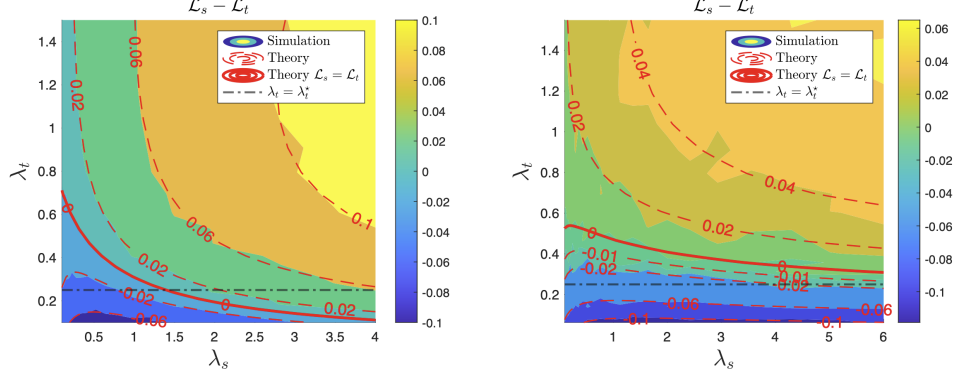


Figure 2: Test-error difference $\mathcal{L}_s - \mathcal{L}_t$ as a function of (λ_t, λ_s) in the setting of Section 2.2. Filled contours are numerical simulations; dashed red contours follow the theory of Theorem 8. The solid curve marks $\mathcal{L}_s = \mathcal{L}_t$, and the dashed black curve is $\lambda_t = \lambda_t^*$. **Left:** under-parameterized student. **Right:** over-parameterized student. See Section 4 for more details.

Remark 6. The weighted ridge estimator $\text{WRidge}(\mathcal{S}, \lambda, \Gamma)$ can also be seen as training the second layer of a two-layer linear neural network $f_{\text{NN}}(\mathbf{x}) = \mathbf{x}^\top \Gamma \alpha$; i.e., $\hat{\beta} = \Gamma \hat{\alpha}$ in which

$$\hat{\alpha} = \underset{\alpha \in \mathbb{R}^{d_x}}{\text{argmin}} \left[\frac{1}{n} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{S}} (y_i - \mathbf{x}_i^\top \Gamma \alpha)^2 + \lambda \|\alpha\|_2^2 \right].$$

In light of the connection to linear neural networks in Remark 6, one can think of the student model as a pre-trained neural network. During the pre-training, we assume that the student model has had access to data from various sources with a shared structure with β_* . The goal of the pre-training is to use this data to learn features that align well with the underlying task β_* [Sun et al., 2021]. Motivated by a recent line of results in deep learning theory where the updated first layer weights are shown to have a *spiked structure* with a few directions having information about the target function [Ba et al., 2022, Moniri et al., 2024, Cui et al., 2024, Zhang et al., 2025, Ba et al., 2024, Demir and Doğan, 2024, Moniri and Hassani, 2024, Li and Sonthalia, 2024, Mousavi-Hosseini et al., 2023, Radhakrishnan et al., 2024], we model the alignment of Γ with the task structure using a non-informative bulk component plus an informative low-rank component.

Assumption 7. We assume that the matrix $\Gamma \in \mathbb{R}^{d_x \times d_x}$ is given by

$$\Gamma = \mathbf{I}_{d_x} + d_x \hat{\beta} \hat{\beta}^\top \quad \text{with} \quad \frac{|\hat{\beta}^\top \beta_*|}{\|\hat{\beta}\|_2 \|\beta_*\|_2} \xrightarrow{\mathbb{P}} \zeta \quad (8)$$

where $\zeta \in [0, 1]$ is the correlation of the learned direction $\hat{\beta}$ to the target direction β_* which is a measure of how much $\hat{\beta}$ aligns with the target direction β_* .

The prefactor d_x for the spike term in (8) is chosen in a way to ensure that $\|\mathbf{I}\|_{\text{Fr}} \asymp \|d_x \hat{\beta} \hat{\beta}^\top\|_{\text{Fr}}$. This closely resembles the scaling of the updated weights with maximal update parameterization in the feature learning theory literature [Yang and Hu, 2021, Ba et al., 2022]. In the following theorem, we characterize the test error difference of the models $\mathcal{L}_s - \mathcal{L}_t$ for this setting in the high-dimensional proportional regime of Assumption 1.

Theorem 8. Under the conditions of Proposition 3, the test errors of the student and teacher model from (7) with Γ from Assumption 7 satisfy $\mathcal{L}_s - \mathcal{L}_t \xrightarrow{\mathbb{P}} \Delta - \zeta^2 \Delta_\Gamma$ where the expression for Δ is given in Theorem 4, and

$$\Delta_\Gamma := \lambda_s (-1 + \lambda_t m_{t,1}) \left[-2\lambda_t m_{s,1} m_{t,1} + \lambda_s m_{s,2} (-1 + \lambda_t m_{t,1}) \right].$$

where $m_{t,1}, m_{t,2}, m_{s,1}, m_{s,2}$ are defined in Definition 2. Additionally, we have $\Delta_\Gamma \geq 0$.

In the limiting formula for $\mathcal{L}_s - \mathcal{L}_t$, the term Δ is equal to the limiting value $\mathcal{L}_s - \mathcal{L}_t$ in the setting of Section 2.1 where both models are trained using (standard) ridge regression, and the benefit of the learned features for the student is due to the term $-\zeta^2 \Delta_{\mathbf{F}}$, which is always non-positive. Because of this term, even in the settings that $\Delta \geq 0$ (i.e., the mechanism of Section 2.1 is not enough on its own for weak-to-strong generalization), the student may still outperform the teacher.

Figures 2 and 3 demonstrate that the asymptotic characterization of Theorem 8 match simulations very well even for moderately large n_s, n_t, d_x . In Figure 2, we fix γ_s, γ_t and σ_ε^2 and plot the contours of $\mathcal{L}_s - \mathcal{L}_t$ for different (λ_t, λ_s) pairs. We show that unlike Section 2.1, in both settings with $\gamma_s > 1$ or $\gamma_s < 1$, pairs (λ_t, λ_s) with $\lambda_t > \lambda_t^* = \sigma_\varepsilon^2 \gamma_t$ (i.e., over-regularized teacher) exist where the student model outperforms the teacher. In Figure 3, we set $\lambda_t = \lambda_t^*$ and plot $\mathcal{L}_s$ as a function of λ_s for different values of the feature quality parameter ζ . We see that for small ζ , the student never outperforms the teacher, similar to the case in Section 2.1. However, this changes when ζ is increased, and the student can have a smaller test error for some values of λ_s .

Mechanism of Weak-to-Strong Generalization. Theorem 8 shows that if the student model has been pre-trained and has learned features that are better suited for the task of predicting the target function (or equivalently is fine-tuned using a better regularization structure), it can leverage this advantage to achieve a better performance compared to the teacher, despite being trained on labels generated by the teacher. This shows yet another mechanism of weak-to-strong generalization.

3 The Nonlinear Case

In Section 2, we considered a setting where both the student and the teacher model trained a linear head through a convex objective. Although this is a very effective model for the analysis of the roles of regularization and over-parameterization, it coincides to the linearized regime of neural network training where features are frozen at their initialization. As a result, the linearized models are not rich enough to study phenomena that happen as a result of (nonlinear) *feature learning*.

Consider the problem of learning a few distinct skills and abilities using data from a compositional task, where a blend of different skills are needed in order to succeed. We model each skill as a vector in the high-dimensional input space $\mathbb{R}^{d_x}$, and learning a skill as learning the corresponding direction. Take, as a running example the task of holding a coherent conversation in an unfamiliar language. To succeed, the model needs to blend abilities such as logical reasoning, with language-specific skills. Abilities such as logical reasoning are difficult to acquire, but transferable across domains. However, language-specific abilities such as vocabulary and grammar, are conceptually straightforward, but they are task-specific and often cannot be learned from other related tasks.

Motivated by this setting, in our model, we let the samples for the teacher model be generated according to a multi-index function (e.g., [Box and Cox \[1964\]](#), [Bickel and Doksum \[1981\]](#)) with an easy and a hard component. Here, the teacher has access to $\mathcal{S}_t = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n_t}$ drawn from

$$\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_x}), \quad \text{and} \quad y_i = \sigma_e(\mathbf{x}_i^\top \beta_e) + \sigma_h(\mathbf{x}_i^\top \beta_h), \quad (9)$$

where $\beta_e, \beta_h \in \mathbb{R}^{d_x}$ are two orthonormal directions that we want to learn, and $\sigma_e, \sigma_h : \mathbb{R} \rightarrow \mathbb{R}$ are two link functions. In this problem, the hardness of learning each direction β_e, β_h from these samples is known to be characterized by the *information-exponent* of their link function [[Dudeja and Hsu, 2018](#), [Ben Arous et al., 2021](#)]. For any real-valued function $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ with Hermite coefficients $\{c_{\sigma,k}\}_{k=0}^\infty$, the information-exponent is defined as $\kappa_\sigma := \min\{k \in \mathbb{N} : c_{\sigma,k} \neq 0\}$. We make the following assumption on σ_e and σ_h .

Assumption 9. Assume that $\kappa_{\sigma_e} = 1$ and $\kappa_{\sigma_h} > 1$ (i.e., σ_e is an easy and σ_h is a hard link function).

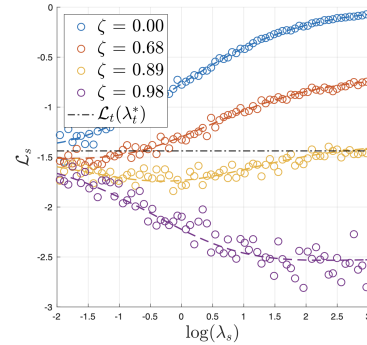


Figure 3: Student error $\mathcal{L}_s$ versus $\log \lambda_s$ in the setting of Section 2.2, plotted for several values of ζ with the teacher optimally regularized. Circles show simulation results, and dashed curves are the predictions of Theorem 8. The dashed black line marks the teacher error $\mathcal{L}_t$. See Section 4 for details.

We let the student $\hat{f}_s$ and teacher $\hat{f}_t$ models be neural networks given by $\hat{f}_s(\mathbf{x}) = \mathbf{a}_s^\top \sigma(\mathbf{W}_s \mathbf{x})$ and $\hat{f}_t(\mathbf{x}) = \mathbf{a}_t^\top \sigma(\mathbf{W}_t \mathbf{x})$, where $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ is an activation function with $\kappa_\sigma = 1$, and $\mathbf{a}_t \in \mathbb{R}^{p_t}$, $\mathbf{a}_s \in \mathbb{R}^{p_s}$, $\mathbf{W}_t \in \mathbb{R}^{p_t \times d_x}$, and $\mathbf{W}_s \in \mathbb{R}^{p_s \times d_x}$. To learn the relevant directions β_e, β_h , we update the first layer weights to align them to these directions; a task also referred to as *weak recovery* [Ben Arous et al., 2021, Dandi et al., 2023, 2024, Arnaboldi et al., 2024, Lee et al., 2024].

For training, we update the teacher model $\hat{f}_t$ using the samples $\mathcal{S}$ and use $\hat{f}_t$ to generate synthetic labels for $n_s \in \mathbb{N}$ unlabeled covariates $\mathcal{S}_s = \{\tilde{\mathbf{x}}_i\}_{i=1}^{n_s}$ drawn from the same distribution according to $\tilde{\mathbf{x}}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_x})$ as $\tilde{y}_i = \hat{f}_t(\tilde{\mathbf{x}}_i)$. We then use these samples to update the student model $\hat{f}_s$. We consider the correlation loss defined as

$$\hat{\mathcal{L}}_t := -n_t^{-1} \sum_{i=1}^{n_t} y_i \hat{f}_t(\mathbf{x}_i), \quad \text{and} \quad \hat{\mathcal{L}}_s := -n_s^{-1} \sum_{i=1}^{n_s} \tilde{y}_i \hat{f}_s(\tilde{\mathbf{x}}_i). \quad (10)$$

Fixing $\mathbf{a}_t \in \mathbb{R}^{p_t}$ with $\|\mathbf{a}_t\|_2 = \Theta(1)$, and initializing $\mathbf{W}_t$ at $\mathbf{W}_{t,0}$ with i.i.d. $\mathcal{N}(0, d_x^{-1})$ entries, we update $\mathbf{W}_t$ using one-step of gradient descent on $\mathcal{L}_t$ given by

$$\widehat{\mathbf{W}}_t = \mathbf{W}_{t,0} - \eta_t \nabla_{\mathbf{W}_t} \hat{\mathcal{L}}_t|_{\mathbf{W}_{t,0}, \mathbf{a}_t}.$$

The following result, which is a corollary of Ba et al. [2022, Proposition 2], shows that the after this update, $\widehat{\mathbf{W}}_t$ aligns to the easy direction β_e , but does not align to the hard direction β_h ; i.e., the teacher model *could not learn the hard direction* using these samples.

Proposition 10. *Under the high-dimensional proportional limit of Assumption 1, and assuming that $\eta_t = O(1)$ and $p_t = \Theta(d_x)$, we have*

$$\|\widehat{\mathbf{W}}_t \beta_e\|_2 \rightarrow_{\mathbb{P}} c > 0 \quad \text{and} \quad \|\widehat{\mathbf{W}}_t \beta_h\|_2 \rightarrow_{\mathbb{P}} 0.$$

After this update, we set $\hat{f}_t(\mathbf{x}) = \mathbf{a}_t^\top \sigma(\widehat{\mathbf{W}}_t \mathbf{x})$. We assumed that the teacher has not gone through extensive pre-training on other tasks that depend on the directions β_e, β_h ; thus, we made the assumption that the $\mathbf{W}_t$ is initialized at random. Recall that $\hat{\beta}_h$ is a hard direction, but it is relevant for a variety of tasks. Thus, we assume that the student has been pre-trained on a variety of different relevant tasks, and has already learned the hard but cross-domain ability that corresponds to β_h ; thus $\mathbf{W}_s$ at initialization is aligned to β_h . In particular, we set $\mathbf{W}_{s,0} = \mathbf{W}_{s,0} + \tau \bar{\mathbf{a}} \beta_h^\top$ where $\mathbf{W}_{s,0} \in \mathbb{R}^{p_s \times d_x}$ has $\mathcal{N}(0, d_x^{-1})$ entries, $\tau \in \mathbb{R}$ and $\bar{\mathbf{a}} \in \mathbb{R}^{p_s}$ is a unit norm vector. We update $\mathbf{W}_s$ using one-step of gradient descent on $\mathcal{L}_s$; i.e.,

$$\widehat{\mathbf{W}}_s = \mathbf{W}_{s,0} - \eta_s \nabla_{\mathbf{W}_s} \hat{\mathcal{L}}_s|_{\mathbf{W}_{s,0}, \mathbf{a}_s}.$$

Note that training $\mathbf{W}_s$ excessively on data generated from the teacher can result in the student to forget the direction β_h . However, in the next theorem, we show that a single step of SGD on $\hat{\mathcal{L}}_s$ can induce non-trivial alignment between $\widehat{\mathbf{W}}_s$ and the easy direction β_e while keeping the weights aligned to β_h . This aligns to the empirical and theoretical findings of Burns et al. [2024], Medvedev et al. [2025] that show that early stopping the teacher is required for weak-to-strong generalization.

Theorem 11. *In the asymptotic regime of Assumption 1 with $p_t, p_s = \Theta(d_x)$, assuming that $\eta_t, \eta_s = O(1)$ and $\tau = o(\sqrt{d_x})$, we have*

$$\|\widehat{\mathbf{W}}_s \beta_e\|_2 \rightarrow_{\mathbb{P}} c_e > 0 \quad \text{and} \quad \|\widehat{\mathbf{W}}_s \beta_h\|_{\text{op}} \rightarrow_{\mathbb{P}} c_h > 0$$

This theorem is an extension of [Ba et al., 2022, Proposition 2] to the case where the first-layer weight is initialized as a spiked random matrix and the labels are also generated by another one-step updated two-layer neural network, which can be also of independent interest. This theorem shows that, under this setting, the student model is still able to learn the direction β_e from the imperfect labels generated by the teacher, and achieve non-vanishing alignment to both the directions β_e, β_h , although the student was unable to learn the hard direction β_h .

Mechanism of Weak-to-Strong Generalization. In this setting, a teacher model can acquire the easier, yet specialized skills through fine-tuning, even though it cannot learn abilities that are challenging to learn. By contrast, the student is pretrained on vast, heterogeneous corpora and already possesses those hard, yet cross-domain, skills. As a result, weak-to-strong generalization can happen when the teacher teaching the student the specialized language abilities, thereby complementing the student's strengths from pre-training.

4 Numerical Validation

In this section, we provide the details of the simulations presented throughout the papers.

Figures 1 and 2. We fix the values of d_x, n_t, n_s and σ_ε and plot the contours of $\mathcal{L}_s - \mathcal{L}_t$ using numerical simulations and also the results of Theorem 4 and 8. The simulation results are averaged over ten trials. See Section 2.1 and 2.2 for discussions of the results. In Figure 1, for the under-parameterized regime (**left**), we set $d_x = 500, n_t = n_s = 2000, \sigma_\varepsilon = 1$ and for the over-parameterized regime (**right**), we set $d_x = 500, n_t = 2000, n_s = 416, \sigma_\varepsilon = 2$. In Figure 2, for the under-parameterized regime (**left**), we set $\zeta = 0.8, d_x = 500, n_t = n_s = 2000, \sigma_\varepsilon = 1$, and the over-parameterized regime (**right**), we set $\zeta = 0.88, d_x = 500, n_t = 2000, n_s = 416, \sigma_\varepsilon = 1$.

Figure 3. In these experiments, we set $d_x = 500, n_t = n_s = 2000, \sigma_\varepsilon = 1$ and set $\lambda_t = \sigma_\varepsilon^2 \gamma_t = 0.25$. We compare the theoretical curves of $\mathcal{L}_s$ as a function of γ_s with numerical simulation, for $\zeta \in \{0, 0.68, 0.89, 0.98\}$. The simulations in this experiment have not been averaged over multiple trials. See Section 2.2 for discussion of the results.

5 Conclusion

In this paper, we show that weak-to-strong generalization is not unique to complex language models and can happen even in much simpler theoretical setups. By studying three natural and tractable learning theoretical settings, we identified and theoretically analyzed three distinct routes by which a student model can outperform its teacher: (1) *compensating for under-regularization*, (2) *harnessing a more task-aligned regularization structure*, and (3) *combining teacher-taught, easy-to-learn components with pretrained, hard-to-learn features*. Our results clarify when and why these effects arise, complementing prior empirical and theoretical insights about the roles of regularization and overparameterization, and the role of feature adaptation.

6 Acknowledgments

Behrad Moniri gratefully acknowledges the gift from AWS AI to Penn Engineering’s ASSET Center for Trustworthy AI. The work of Behrad Moniri and Hamed Hassani is supported by The Institute for Learning-enabled Optimization at Scale (TILOS), under award number NSF-CCF-2112665, and the NSF CAREER award CIF-1943064.

References

- Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office, 1968.
- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *OpenAI Technical Reports*, 2023.
- Luca Arnaboldi, Yatin Dandi, Florent Krzakala, Luca Pesce, and Ludovic Stephan. Repetita iuvant: Data repetition allows SGD to learn high-dimensional multi-index functions. *arXiv preprint arXiv:2405.15459*, 2024.
- Alexander Atanasov, Jacob A Zavatore-Veth, and Cengiz Pehlevan. Risk and cross validation in ridge regression with correlated samples. *arXiv preprint arXiv:2408.04607*, 2024.
- Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In *Advances in Neural Information Processing Systems*, 2022.
- Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, and Denny Wu. Learning in the presence of low-dimensional structure: a spiked random matrix perspective. In *Advances in Neural Information Processing Systems*, 2024.
- Zhidong Bai and Jack W Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. Springer, 2010.

- Hritik Bansal, Arian Hosseini, Rishabh Agarwal, Vinh Q Tran, and Mehran Kazemi. Smaller, weaker, yet better: Training LLM reasoners via compute-optimal sampling. In *International Conference on Learning Representations*, 2025.
- Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.
- Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research*, 22(106):1–51, 2021.
- Peter J Bickel and Kjell A Doksum. An analysis of transformations revisited. *Journal of the American Statistical Association*, 76(374):296–311, 1981.
- George EP Box and David R Cox. An analysis of transformations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 26(2):211–243, 1964.
- Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. Weak-to-strong generalization: eliciting strong capabilities with weak supervision. In *International Conference on Machine Learning*, 2024.
- Mireille Capitaine and Catherine Donati-Martin. Strong asymptotic freeness for wigner and wishart matrices. *Indiana University mathematics journal*, pages 767–803, 2007.
- George Casella. Minimax ridge regression estimation. *The Annals of Statistics*, 8(5), 1980.
- Moses Charikar, Chirag Pabbaraju, and Kirankumar Shiragur. Quantifying the gain in weak-to-strong generalization. In *Advances in neural information processing systems*, volume 37, 2024.
- Hugo Cui, Luca Pesce, Yatin Dandi, Florent Krzakala, Yue Lu, Lenka Zdeborova, and Bruno Loureiro. Asymptotics of feature learning in two-layer networks after one gradient-step. In *International Conference on Machine Learning*, 2024.
- Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. Learning two-layer neural networks, one (giant) step at a time. *arXiv preprint arXiv:2305.18270*, 2023.
- Yatin Dandi, Emanuele Troiani, Luca Arnaboldi, Luca Pesce, Lenka Zdeborova, and Florent Krzakala. The benefits of reusing batches for gradient descent in two-layer networks: Breaking the curse of information and leap exponents. In *Forty-first International Conference on Machine Learning*, 2024.
- Rudrajit Das and Sujay Sanghavi. Understanding self-distillation in the presence of label noise. In *International Conference on Machine Learning*, 2023.
- Samet Demir and Zafer Doğan. Random features outperform linear models: Effect of strong input-label correlation in spiked covariance data. *arXiv preprint arXiv:2409.20250*, 2024.
- Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.
- Elvis Dohmatob, Yunzhen Feng, and Julia Kempe. Model collapse demystified: The case of regression. In *Advances in Neural Information Processing Systems*, 2024.
- Yijun Dong, Yicheng Li, Yunai Li, Jason D Lee, and Qi Lei. Discrepancies are virtue: Weak-to-strong generalization through lens of intrinsic dimension. In *International Conference on Machine Learning*, 2025.
- Rishabh Dudeja and Daniel Hsu. Learning single-index models in gaussian space. In *Conference On Learning Theory*, 2018.
- Adina Roxana Feier. *Methods of proof in random matrix theory*, volume 9. 2012.

- Michael J Feldman. Spectral properties of elementwise-transformed spiked matrices. *SIAM Journal on Mathematics of Data Science*, 7(2):542–571, 2025.
- Alice Guionnet, Justin Ko, Florent Krzakala, Pierre Mergny, and Lenka Zdeborová. Spectral phase transitions in non-linear wigner spiked models. *arXiv preprint arXiv:2310.14055*, 2023.
- Jianyuan Guo, Hanting Chen, Chengcheng Wang, Kai Han, Chang Xu, and Yunhe Wang. Vision superalignment: Weak-to-strong generalization for vision foundation models. *arXiv preprint arXiv:2402.03749*, 2024.
- David Lee Hanson and Farroll Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083, 1971.
- Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.
- Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.
- Hong Hu and Yue M Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 69(3), 2023.
- M Emrullah Ildiz, Halil Alperen Gozeten, Ege Onur Taga, Marco Mondelli, and Samet Oymak. High-dimensional analysis of knowledge distillation: Weak-to-strong generalization and scaling laws. In *International Conference on Learning Representations*, 2025.
- Jiaming Ji, Boyuan Chen, Hantao Lou, Donghai Hong, Borong Zhang, Xuehai Pan, Juntao Dai, and Yaodong Yang. Aligner: Achieving efficient alignment through weak-to-strong correction. In *Advances in Neural Information Processing Systems*, 2024.
- Satish Babu Korada and Andrea Montanari. Applications of the lindeberg principle in communications and statistical learning. *IEEE Transactions on Information Theory*, 57(4):2440–2450, 2011.
- Arun Kumar Kuchibhotla and Abhishek Chakraborty. Moving beyond sub-gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *Information and Inference: A Journal of the IMA*, 11(4):1389–1456, 2022.
- Jason D Lee, Kazusato Oko, Taiji Suzuki, and Denny Wu. Neural network learns low-dimensional polynomials with SGD near the information-theoretic limit. In *Advances in Neural Information Processing Systems*, 2024.
- Jiping Li and Rishi Sonthalia. Generalization for least squares regression with simple spiked covariances. *arXiv preprint arXiv:2410.13991*, 2024.
- Jarl Waldemar Lindeberg. Eine neue herleitung des exponentialgesetzes in der wahrscheinlichkeit-srechnung. *Mathematische Zeitschrift*, 15(1):211–225, 1922.
- Yuejiang Liu and Alexandre Alahi. Co-supervised learning: Improving weak-to-strong generalization with hierarchical mixture of experts. *arXiv preprint arXiv:2402.15505*, 2024.
- Sadhika Malladi, Alexander Wettig, Dingli Yu, Danqi Chen, and Sanjeev Arora. A kernel-based view of language model fine-tuning. In *International Conference on Machine Learning*, 2023.
- V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues in certain sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967.
- Marko Medvedev, Kaifeng Lyu, Dingli Yu, Sanjeev Arora, Zhiyuan Li, and Nathan Srebro. Weak-to-strong generalization even in random feature networks, provably. *arXiv preprint arXiv:2503.02877*, 2025.
- Behrad Moniri and Hamed Hassani. Signal-plus-noise decomposition of nonlinear spiked random matrix models. *arXiv preprint arXiv:2405.18274*, 2024.
- Behrad Moniri and Hamed Hassani. Asymptotics of linear regression with linearly dependent data. In *Annual Learning for Dynamics and Control Conference*, 2025.

- Behrad Moniri, Donghwan Lee, Hamed Hassani, and Edgar Dobriban. A theory of non-linear feature learning with one gradient step in two-layer neural networks. In *International Conference on Machine Learning*, 2024.
- Andrea Montanari and Basil N Saeed. Universality of empirical risk minimization. In *Conference on Learning Theory*, 2022.
- Alireza Mousavi-Hosseini, Denny Wu, Taiji Suzuki, and Murat A Erdogdu. Gradient-based feature learning under structured data. In *Advances in Neural Information Processing Systems*, 2023.
- Abhijeet Mulgund and Chirag Pabbaraju. Relating misfit to gain in weak-to-strong generalization beyond the squared loss. *arXiv preprint arXiv:2501.19105*, 2025.
- Preetum Nakkiran, Prayaag Venkat, Sham M Kakade, and Tengyu Ma. Optimal regularization can mitigate double descent. In *International Conference on Learning Representations*, 2021.
- OpenAI. Introducing superalignment. *OpenAI Blog*, 2023.
- Divyansh Pareek, Simon Shaolei Du, and Sewoong Oh. Understanding the gains from repeated self-distillation. In *Advances in Neural Information Processing Systems*, 2024.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI Blog*, 2019.
- Adityanarayanan Radhakrishnan, Daniel Beaglehole, Parthe Pandit, and Mikhail Belkin. Mechanism for feature learning in neural networks and backpropagation-free machine learning models. *Science*, 383(6690):1461–1467, 2024.
- Dominic Richards, Jaouad Mourtada, and Lorenzo Rosasco. Asymptotics of ridge (less) regression under general source condition. In *International Conference on Artificial Intelligence and Statistics*, 2021.
- Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18:1–9, 2013.
- Changho Shin, John Cooper, and Frederic Sala. Weak-to-strong generalization through the data-centric lens. In *International Conference on Learning Representations*, 2025.
- Yue Sun, Adhyayan Narang, Ibrahim Gulluk, Samet Oymak, and Maryam Fazel. Towards sample-efficient overparameterized meta-learning. In *Advances in Neural Information Processing Systems*, 2021.
- Leitian Tao and Yixuan Li. Your weak LLM is secretly a strong teacher for alignment. *arXiv preprint arXiv:2409.08813*, 2024.
- Antonio M Tulino and Sergio Verdú. Random matrix theory and wireless communications. *Communications and Information Theory*, 1(1):1–182, 2004.
- Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Yonina C. Eldar and Gitta Kutyniok, editors, *Compressed Sensing: Theory and Applications*, page 210–268. Cambridge University Press, 2012.
- Dan Voiculescu. Limit laws for random matrices and free products. *Inventiones mathematicae*, 104(1):201–220, 1991.
- Zhichao Wang, Andrew Engel, Anand Sarwate, Ioana Dumitriu, and Tony Chiang. Spectral evolution and invariance in linear-width neural networks. *arXiv preprint arXiv:2211.06506*, 2022.
- Alexander Wei, Wei Hu, and Jacob Steinhardt. More than a toy: Random matrix models predict how real-world neural representations generalize. In *International Conference on Machine Learning*, 2022.
- David Xing Wu and Anant Sahai. Provable weak-to-strong generalization via benign overfitting. In *NeurIPS 2024 Workshop on Mathematics of Modern Machine Learning*, 2024.

- Denny Wu and Ji Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. In *Advances in Neural Information Processing Systems*, 2020.
- Yihao Xue, Jiping Li, and Baharan Mirzasoleiman. Representations shape weak-to-strong generalization: Theoretical insights and empirical predictions. In *International Conference on Machine Learning*, 2025.
- Greg Yang and Edward J Hu. Tensor programs IV: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning*, 2021.
- Yuqing Yang, Yan Ma, and Pengfei Liu. Weak-to-strong reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024.
- Thomas T Zhang, Behrad Moniri, Ansh Nagwekar, Faraz Rahman, Anton Xue, Hamed Hassani, and Nikolai Matni. On the concurrence of layer-wise preconditioning methods and provable feature learning. In *International Conference on Machine Learning*, 2025.

A Preliminaries

Given two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n_1 \times n_2}$, we denote their Hadamard product (element-wise product) by $\mathbf{A} \odot \mathbf{B} \in \mathbb{R}^{n_1 \times n_2}$. Also, for $k \in \mathbb{N}$, we define the Hadamard power as

$$\mathbf{A}^{\odot k} = \underbrace{\mathbf{A} \odot \dots \odot \mathbf{A}}_{k \text{ times}} \in \mathbb{R}^{n_1 \times n_2}.$$

Lemma 12. For any $\mathbf{v} \in \mathbb{R}^{n_1}$, $\mathbf{u} \in \mathbb{R}^{n_2}$, and $\mathbf{C} \in \mathbb{R}^{n_1 \times n_2}$, we have

$$(\mathbf{v}\mathbf{u}^\top) \odot \mathbf{C} = \text{diag}(\mathbf{v}) \mathbf{C} \text{diag}(\mathbf{u}).$$

Proof. The proof is immediate by writing the entries of the two sides. $\square$

We also heavily leverage the following theorem to prove the concentration inequality for quadratic forms. See e.g., [Rudelson and Vershynin \[2013\]](#) for a modern proof.

Theorem 13 (Hanson-Wright Inequality [[Hanson and Wright, 1971](#)]). Let $\mathbf{x} = (X_1, \dots, X_n) \in \mathbb{R}^d$ be a random vector with independent sub-gaussian components X_i with $\mathbb{E}X_i = 0$. Let $\mathbf{D}$ be an $n \times n$ matrix. Then, for every $t \geq 0$, we have

$$\mathbb{P}\left[|\mathbf{x}^\top \mathbf{D} \mathbf{x} - \mathbb{E}[\mathbf{x}^\top \mathbf{D} \mathbf{x}]| > t\right] \leq 2 \exp\left[-c \min\left(\frac{t^2}{\|\mathbf{D}\|_F^2}, \frac{t}{\|\mathbf{D}\|_{\text{op}}}\right)\right],$$

where c is a constant that depends only on the sub-gaussian constants of X_i .

To analyze spiked random matrices, we will use the following matrix identity.

Lemma 14. (Sherman-Morrison Formula). Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be an invertible matrix, and let $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ be column vectors such that $1 + \mathbf{v}^\top \mathbf{A}^{-1} \mathbf{u} \neq 0$. Then the inverse of the rank-one update $\mathbf{A} + \mathbf{u}\mathbf{v}^\top$ is given by:

$$(\mathbf{A} + \mathbf{u}\mathbf{v}^\top)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1} \mathbf{u} \mathbf{v}^\top \mathbf{A}^{-1}}{1 + \mathbf{v}^\top \mathbf{A}^{-1} \mathbf{u}}.$$

A.1 Hermite Polynomials

We let H_k be the k -th (probabilist's) Hermite polynomial on $\mathbb{R}$ defined by

$$H_k(x) = (-1)^k \exp(x^2/2) \frac{d^k}{dx^k} \exp(-x^2/2) \quad \forall x \in \mathbb{R}.$$

These polynomials form an orthogonal basis in the Hilbert space L^2 of measurable functions $f : \mathbb{R} \rightarrow \mathbb{R}$ such that

$$\int f^2(x) e^{-\frac{x^2}{2}} dx < \infty$$

with inner product

$$\langle f, g \rangle = \int f(x) g(x) e^{-\frac{x^2}{2}} dx.$$

The first few Hermite polynomials are

$$H_0(x) = 1, \quad H_1(x) = x, \quad \text{and} \quad H_2(x) = x^2 - 1.$$

Lemma 15. For any $k \in \mathbb{N}$ and $x, y \in \mathbb{R}$, we have

$$H_k(x + y) = \sum_{j=0}^k \binom{k}{j} x^j H_{k-j}(y)$$

Proof. Note that using [[Abramowitz and Stegun, 1968](#), Equation 22.8.8] we have

$$\frac{d}{dx} H_k(x) = k H_{k-1}(x).$$

Thus, the j -th derivative of H_k is given by

$$\frac{d^j}{dx^j} H_k(x) = \frac{k!}{(k-j)!} H_{k-j}(x).$$

By Taylor expanding $H_k(x+y)$ at y , we find

$$H_k(x+y) = \sum_{j=0}^k \frac{x^j}{j!} \frac{d^j}{dy^j} H_k(y) = \sum_{j=0}^k \binom{k}{j} x^j H_{k-j}(y),$$

proving the lemma.

A.2 Random Matrix Theory

We first define the following empirical covariance and resolvent matrices that will appear throughout the proofs.

$$\begin{aligned} \hat{\Sigma} &= \mathbf{X}^\top \mathbf{X} / n_t, \quad \tilde{\Sigma} = \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} / n_s, \\ \hat{\mathbf{R}} &= (\hat{\Sigma} + \lambda_t \mathbf{I}_{d_X})^{-1}, \quad \text{and} \quad \tilde{\mathbf{R}} = (\tilde{\Sigma} + \lambda_s \mathbf{I}_{d_X})^{-1}. \end{aligned}$$

We will use the following characterization of the eigenvalues of $\hat{\Sigma}, \tilde{\Sigma}$ in the high-dimensional proportional limit by [Marchenko and Pastur \[1967\]](#).

Theorem 16 (Marchenko–Pastur Theorem). *In the high-dimensional proportional limit where $n_t, n_s, d_X \rightarrow \infty$ such that $\frac{d_X}{n_s} \rightarrow \gamma_s$ and $\frac{d_X}{n_t} \rightarrow \gamma_t$, the empirical spectral distribution (ESD) of $\hat{\Sigma}$ (and $\tilde{\Sigma}$) converges almost surely to the Marchenko–Pastur distribution $\mu_{MP(\gamma_t)}$ (and $\mu_{MP(\gamma_s)}$).*

Recall from definition 2 that the function $m(\cdot, \cdot) \rightarrow \mathbb{R}$ is defined as

$$m(\lambda; \gamma) = \int \frac{d\mu_{MP(\gamma)}(s)}{s + \lambda}.$$

Hence, taking derivatives with respect to λ , we get

$$m'(\lambda; \gamma) = \frac{\partial}{\partial \lambda} m(\lambda; \gamma) = - \int \frac{d\mu_{MP(\gamma)}(s)}{(s + \lambda)^2}.$$

In the following section, we write the test error $\mathcal{L}_s$ and $\mathcal{L}_t$ as a function of m and its derivatives. In particular, note that using the Marchenko–Pastur theorem, we have

$$\begin{aligned} d_X^{-1} \text{Tr}(\hat{\mathbf{R}}) &\rightarrow_{\mathbb{P}} m_{t,1}, \quad d_X^{-1} \text{Tr}(\hat{\mathbf{R}}^2) \rightarrow_{\mathbb{P}} m_{t,2}, \quad \text{and} \\ d_X^{-1} \text{Tr}(\tilde{\mathbf{R}}) &\rightarrow_{\mathbb{P}} m_{s,1}, \quad d_X^{-1} \text{Tr}(\tilde{\mathbf{R}}^2) \rightarrow_{\mathbb{P}} m_{s,2}, \end{aligned}$$

where for $p \in \{s, t\}$, we define $m_{p,1} = m(\lambda_p, \gamma_p)$ and $m_{p,2} = -\frac{\partial m}{\partial \lambda} \big|_{\lambda_p, \gamma_p}$.

In the following proofs, we will also use the asymptotic freeness of independent Wishart random matrices [\[Voiculescu, 1991, Capitaine and Donati-Martin, 2007\]](#). See [\[Feier, 2012, Section 3.4 and 4.4\]](#) for a brief overview of free probability theory for random matrix theory.

B Proof of Proposition 3

To prove this theorem, note that the vector $\hat{\beta}_t$ can be written as

$$\begin{aligned} \hat{\beta}_t &= (\mathbf{X}^\top \mathbf{X} + \lambda_t n_t \mathbf{I}_{d_X})^{-1} \mathbf{X}^\top \mathbf{y} \\ &= (\mathbf{X}^\top \mathbf{X} + \lambda_t n_t \mathbf{I}_{d_X})^{-1} \mathbf{X}^\top (\mathbf{X} \beta_\star + \varepsilon) \\ &= (\hat{\Sigma} + \lambda_t \mathbf{I}_{d_X})^{-1} \hat{\Sigma} \beta_\star + n_t^{-1} (\hat{\Sigma} + \lambda_t \mathbf{I}_{d_X})^{-1} \mathbf{X}^\top \varepsilon, \end{aligned}$$

where we have used the fact that $\mathbf{y} = \mathbf{X}\beta_\star + \varepsilon$. Thus, recalling the definition of $\hat{\Sigma}$ and $\hat{\mathbf{R}}$ from the Section A, we can write

$$\hat{\beta}_t - \beta_\star = (\hat{\mathbf{R}}\hat{\Sigma} - \mathbf{I}_{d_X})\beta_\star + n_t^{-1}\hat{\mathbf{R}}\mathbf{X}^\top\varepsilon$$

The test error of the teacher model is given by $\mathcal{L}_t = \sigma_\varepsilon^2 + \|\hat{\beta}_t - \beta_\star\|_2^2$ in which

$$\|\hat{\beta}_t - \beta_\star\|_2^2 = \beta_\star^\top (\hat{\mathbf{R}}\hat{\Sigma} - \mathbf{I}_{d_X})^\top (\hat{\mathbf{R}}\hat{\Sigma} - \mathbf{I}_{d_X})\beta_\star + n_t^{-2}\varepsilon^\top (\mathbf{X}\hat{\mathbf{R}}^2\mathbf{X}^\top)\varepsilon + o_{\mathbb{P}}(1),$$

where we have used the Hanson-Wright inequality and the facts that ε and $\beta_\star$ are independent mean zero random vectors. Again, by using the Hanson-Wright inequality and recalling that $(\varepsilon, \beta_\star)$ is independent of other sources of randomness in the problem, we can further simplify the above expression to arrive at

$$\|\hat{\beta}_t - \beta_\star\|_2^2 \rightarrow_{\mathbb{P}} d_X^{-1} \text{Tr} \left[(\hat{\mathbf{R}}\hat{\Sigma} - \mathbf{I}_{d_X})^\top (\hat{\mathbf{R}}\hat{\Sigma} - \mathbf{I}_{d_X}) \right] + \sigma_\varepsilon^2 \gamma_t d_X^{-1} \text{Tr} \left[\hat{\Sigma}\hat{\mathbf{R}}^2 \right]$$

Note that $\hat{\mathbf{R}}\hat{\Sigma} = \mathbf{I}_{d_X} - \lambda_t\hat{\mathbf{R}}$. Thus, denoting the eigenvalues of $\hat{\Sigma}$ by $\{\sigma_k\}_{k=1}^d$, we can use the Marchenko-Pastur Theorem for covariance matrices to write

$$d_X^{-1} \text{Tr} \left[(\hat{\mathbf{R}}\hat{\Sigma} - \mathbf{I}_{d_X})^\top (\hat{\mathbf{R}}\hat{\Sigma} - \mathbf{I}_{d_X}) \right] = \lambda_t^2 d_X^{-1} \text{Tr} \left[\hat{\mathbf{R}}^2 \right] \rightarrow_{\mathbb{P}} \lambda_t^2 m_{t,2}.$$

Similarly, we have

$$d_X^{-1} \text{Tr} \left[\hat{\Sigma}\hat{\mathbf{R}}^2 \right] = d_X^{-1} \text{Tr} \left[(\hat{\mathbf{R}} - \lambda_t\hat{\mathbf{R}}^2) \right] \rightarrow_{\mathbb{P}} m_{t,1} - \lambda_t m_{t,2}.$$

Putting these together yields

$$\begin{aligned} \|\hat{\beta}_t - \beta_\star\|_2^2 &\rightarrow_{\mathbb{P}} \lambda_t^2 m_{t,2} + \sigma_\varepsilon^2 \gamma_t (m_{t,1} - \lambda_t m_{t,2}). \\ &= \lambda_t m_{t,2} (\lambda_t - \sigma_\varepsilon^2 \gamma_t) + \sigma_\varepsilon^2 \gamma_t m_{t,1}. \end{aligned}$$

Plugging this into the expression of $\mathcal{L}_t$ gives

$$\mathcal{L}_t \rightarrow_{\mathbb{P}} \sigma_\varepsilon^2 + \lambda_t m_{t,2} (\lambda_t - \sigma_\varepsilon^2 \gamma_t) + \sigma_\varepsilon^2 \gamma_t m_{t,1},$$

which concludes the proof for Proposition 3.

C Proof of Theorem 4

Recalling that $\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\hat{\beta}_t$ and $\hat{\beta}_t = (\mathbf{X}^\top\mathbf{X} + \lambda_t n_t \mathbf{I}_{d_X})^{-1}\mathbf{X}^\top\mathbf{y}$, the vector $\hat{\beta}_s$ can be written as

$$\begin{aligned} \hat{\beta}_s &= (\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \lambda_s n_s \mathbf{I}_{d_X})^{-1}\tilde{\mathbf{X}}^\top\tilde{\mathbf{y}} \\ &= (\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \lambda_s n_s \mathbf{I}_{d_X})^{-1}\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}(\mathbf{X}^\top\mathbf{X} + \lambda_t n_t \mathbf{I}_{d_X})^{-1}\mathbf{X}^\top\mathbf{y} \\ &= (\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \lambda_s n_s \mathbf{I}_{d_X})^{-1}\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}(\mathbf{X}^\top\mathbf{X} + \lambda_t n_t \mathbf{I}_{d_X})^{-1}\mathbf{X}^\top(\mathbf{X}\beta_\star + \varepsilon), \end{aligned}$$

where we have used $\mathbf{y} = \mathbf{X}\beta_\star + \varepsilon$. Using the definition of the matrices $\tilde{\Sigma}, \hat{\mathbf{R}}, \hat{\Sigma}, \tilde{\mathbf{R}}$ from Section A, we can simplify the above expression as

$$\hat{\beta}_s - \beta_\star = (\tilde{\mathbf{R}}\tilde{\Sigma}\hat{\mathbf{R}}\hat{\Sigma} - \mathbf{I}_{d_X})\beta_\star + n_t^{-1}\tilde{\mathbf{R}}\tilde{\Sigma}\hat{\mathbf{R}}\mathbf{X}^\top\varepsilon.$$

Hence, we have

$$\begin{aligned} \|\hat{\beta}_s - \beta_\star\|_2^2 &= \beta_\star^\top (\tilde{\mathbf{R}}\tilde{\Sigma}\hat{\mathbf{R}}\hat{\Sigma} - \mathbf{I}_{d_X})^\top (\tilde{\mathbf{R}}\tilde{\Sigma}\hat{\mathbf{R}}\hat{\Sigma} - \mathbf{I}_{d_X})\beta_\star + n_t^{-2}\varepsilon^\top \mathbf{X}\hat{\mathbf{R}}\tilde{\Sigma}\tilde{\mathbf{R}}^2\tilde{\Sigma}\hat{\mathbf{R}}\mathbf{X}^\top\varepsilon + o_{\mathbb{P}}(1) \\ &= d_X^{-1} \text{Tr} \left[(\tilde{\mathbf{R}}\tilde{\Sigma}\hat{\mathbf{R}}\hat{\Sigma} - \mathbf{I}_{d_X})^\top (\tilde{\mathbf{R}}\tilde{\Sigma}\hat{\mathbf{R}}\hat{\Sigma} - \mathbf{I}_{d_X}) \right] \\ &\quad + \sigma_\varepsilon^2 \gamma_t d_X^{-1} \text{Tr} \left[(\hat{\mathbf{R}}\hat{\Sigma}\hat{\mathbf{R}})(\tilde{\Sigma}\tilde{\mathbf{R}}^2\tilde{\Sigma}) \right] + o_{\mathbb{P}}(1), \end{aligned}$$

where we have used the Hanson-Wright inequality and the facts that ε and $\beta_\star$ are independent of other sources of randomness in the problem. Now, it remain to analyze the following traces in the high-dimensional proportional limit:

$$\begin{aligned} \tau_1 &= d_X^{-1} \text{Tr} \left[(\tilde{\mathbf{R}}\tilde{\Sigma}\hat{\mathbf{R}}\hat{\Sigma} - \mathbf{I}_{d_X})^\top (\tilde{\mathbf{R}}\tilde{\Sigma}\hat{\mathbf{R}}\hat{\Sigma} - \mathbf{I}_{d_X}) \right], \quad \text{and} \\ \tau_2 &= d_X^{-1} \text{Tr} \left[(\hat{\mathbf{R}}\hat{\Sigma}\hat{\mathbf{R}})(\tilde{\Sigma}\tilde{\mathbf{R}}^2\tilde{\Sigma}) \right]. \end{aligned}$$

Analysis of τ_1 . Note that $\tilde{\mathbf{R}}\tilde{\Sigma} = \mathbf{I}_{d_X} - \lambda_s \tilde{\mathbf{R}}$ and $\hat{\mathbf{R}}\hat{\Sigma} = \mathbf{I}_{d_X} - \lambda_t \hat{\mathbf{R}}$, which gives

$$\tilde{\mathbf{R}}\tilde{\Sigma}\hat{\mathbf{R}}\hat{\Sigma} - \mathbf{I}_{d_X} = -\lambda_s \tilde{\mathbf{R}} - \lambda_t \hat{\mathbf{R}} + \lambda_s \lambda_t \tilde{\mathbf{R}}\hat{\mathbf{R}}.$$

Plugging this into the expression for τ_1 , we arrive at

$$\begin{aligned} \tau_1 = & \lambda_s^2 d_X^{-1} \text{Tr} [\tilde{\mathbf{R}}^2] + \lambda_t^2 d_X^{-1} \text{Tr} [\hat{\mathbf{R}}^2] + \lambda_s^2 \lambda_t^2 d_X^{-1} \text{Tr} [\tilde{\mathbf{R}}^2 \hat{\mathbf{R}}^2] \\ & + 2\lambda_s \lambda_t d_X^{-1} \text{Tr} [\tilde{\mathbf{R}}\hat{\mathbf{R}}] - 2\lambda_s \lambda_t^2 d_X^{-1} \text{Tr} [\tilde{\mathbf{R}}\hat{\mathbf{R}}^2] - 2\lambda_s^2 \lambda_t d_X^{-1} \text{Tr} [\hat{\mathbf{R}}\tilde{\mathbf{R}}^2] \end{aligned}$$

The limiting values of these traces can be computed as follows:

- **Term 1 and 2:** Let $\tilde{s}_k$ be the k -th eigenvalue of $\tilde{\Sigma}$. We can use the arguments in Section A to write

$$d_X^{-1} \text{Tr} [\tilde{\mathbf{R}}^2] \rightarrow_{\mathbb{P}} m_{s,2},$$

where $m_{s,2}$ is defined in Definition 2. Similarly, for the second term, we have

$$d_X^{-1} \text{Tr} [\hat{\mathbf{R}}^2] \rightarrow_{\mathbb{P}} m_{t,2},$$

where the term $m_{t,2}$ is defined in Definition 2.

- **Term 3, 4, 5, and 6:** To analyze $d_X^{-1} \text{Tr} [\tilde{\mathbf{R}}^2 \hat{\mathbf{R}}^2]$, we can use the asymptotic freeness of independent Wishart random matrices [Voiculescu, 1991] (see also Capitaine and Donati-Martin [2007]), and the Stone-Weierstrass theorem to approximate the function $f(x) = (x + s)^{-2}$ using polynomials, to write

$$\begin{aligned} d_X^{-1} \text{Tr} [\tilde{\mathbf{R}}^2 \hat{\mathbf{R}}^2] &= \left(d_X^{-1} \text{Tr} [\tilde{\mathbf{R}}^2] \right) \left(d_X^{-1} \text{Tr} [\hat{\mathbf{R}}^2] \right) + o_{\mathbb{P}}(1) \\ &\rightarrow_{\mathbb{P}} m_{s,2} m_{t,2}. \end{aligned}$$

Similarly, for the remaining terms, we have

$$\begin{aligned} d_X^{-1} \text{Tr} [\tilde{\mathbf{R}}\hat{\mathbf{R}}] &\rightarrow_{\mathbb{P}} m_{t,1} m_{s,1} \\ d_X^{-1} \text{Tr} [\tilde{\mathbf{R}}\hat{\mathbf{R}}^2] &\rightarrow_{\mathbb{P}} m_{t,1} m_{s,2}, \quad \text{and} \\ d_X^{-1} \text{Tr} [\hat{\mathbf{R}}^2 \tilde{\mathbf{R}}] &\rightarrow_{\mathbb{P}} m_{t,2} m_{s,1}. \end{aligned}$$

Putting these together, we arrive at the following conclusion:

$$\begin{aligned} \tau_1 \rightarrow_{\mathbb{P}} & \lambda_t^2 m_{t,2} + \lambda_s^2 m_{s,2} + \lambda_s^2 \lambda_t^2 m_{s,2} m_{t,2} \\ & + 2\lambda_t \lambda_s m_{t,1} m_{s,1} - 2\lambda_s \lambda_t^2 m_{s,1} m_{t,2} - 2\lambda_s^2 \lambda_t m_{s,2} m_{t,1}. \end{aligned}$$

Analysis of τ_2 . Again, we can use the identities $\tilde{\mathbf{R}}\tilde{\Sigma} = \mathbf{I}_{d_X} - \lambda_s \tilde{\mathbf{R}}$ and $\hat{\mathbf{R}}\hat{\Sigma} = \mathbf{I}_{d_X} - \lambda_t \hat{\mathbf{R}}$, and the asymptotic freeness of independent Wishart random matrices to write

$$\begin{aligned} \tau_2 &= d_X^{-1} \text{Tr} [(\hat{\mathbf{R}}\hat{\Sigma}\hat{\mathbf{R}})(\tilde{\Sigma}\tilde{\mathbf{R}}^2\tilde{\Sigma})] \\ &= d_X^{-1} \text{Tr} \left[\left(\hat{\mathbf{R}} - \lambda_t \hat{\mathbf{R}}^2 \right) \left(\mathbf{I}_{d_X} - \lambda_s \tilde{\mathbf{R}} \right)^2 \right] \\ &= d_X^{-1} \text{Tr} \left[\left(\hat{\mathbf{R}} - \lambda_t \hat{\mathbf{R}}^2 \right) \right] \cdot d_X^{-1} \text{Tr} \left[\left(\mathbf{I}_{d_X} - \lambda_s \tilde{\mathbf{R}} \right)^2 \right] \end{aligned}$$

Hence, we can use an argument similar to the argument for the term 3 above to write

$$\tau_2 \rightarrow_{\mathbb{P}} \left(m_{t,1} - \lambda_t m_{t,2} \right) \left(1 + \lambda_s^2 m_{s,2} - 2\lambda_s m_{s,1} \right).$$

Putting everything together. Putting the conclusions above together, the limiting loss difference can be written as

$$\mathcal{L}_s - \mathcal{L}_t = \sigma_\varepsilon^2 + \tau_1 + \tau_2 + o_{\mathbb{P}}(1)$$

$$\rightarrow_{\mathbb{P}} (\sigma_\varepsilon^2 \gamma_t - \lambda_t) [(m_{t,1} - \lambda_t m_{t,2}) (\lambda_s^2 m_{s,2} - 2\lambda_s m_{s,1})] + \lambda_s^2 m_{s,2} (1 - \lambda_t m_{t,1}) := \Delta$$

which completes the proof of the theorem.

D Proof of Theorem 5

From Theorem 4, we know that in the high-dimensional proportional limit, we have

$$\mathcal{L}_s - \mathcal{L}_t \rightarrow_{\mathbb{P}} \Delta = (\sigma_\varepsilon^2 \gamma_t - \lambda_t) [(m_{t,1} - \lambda_t m_{t,2}) (\lambda_s^2 m_{s,2} - 2\lambda_s m_{s,1})] + \lambda_s^2 m_{s,2} (1 - \lambda_t m_{t,1}).$$

To study the (λ_s, λ_t) pairs for which the strong model can outperform the teacher, we study the non-zero roots of the nonlinear equation $\Delta = 0$, which are the solutions to

$$(\sigma_\varepsilon^2 \gamma_t - \lambda_t) \left(\frac{m_{t,1} - \lambda_t m_{t,2}}{\lambda_t m_{t,1} - 1} \right) = \frac{\lambda_s m_{s,2}}{\lambda_s m_{s,2} - 2m_{s,1}} \quad (11)$$

where the left-hand side is a function of teacher parameters, and the right-hand side is a function of the student parameters.

Right-Hand Side. First, recall from Definition 2 that

$$m_{s,1} = m(\lambda_s; \gamma_s), \quad m_{s,2} = -\frac{\partial}{\partial \lambda} m(\lambda; \gamma)|_{\lambda_s, \gamma_s}.$$

in which

$$m(\lambda; \gamma) = -\frac{1}{2\gamma\lambda} \left[1 - \gamma + \lambda - \sqrt{(1 + \gamma + \lambda)^2 - 4\gamma} \right].$$

Thus, after simple Algebraic manipulations, we can write the right-hand side of (11) as

$$H_1 := \frac{\lambda_s m_{s,2}}{\lambda_s m_{s,2} - 2m_{s,1}}$$

$$= \frac{\lambda_s}{1 + \gamma_s^2 + 2\gamma_s(\lambda_s - 1) + \lambda_s + \lambda_s^2 + (1 - \gamma_s - \lambda_s)\sqrt{-4\gamma_s + (1 + \gamma_s + \lambda_s)^2}} \leq 0.$$

We plot H_1 as a function of λ_s for different values of γ_s in Figure 4. It can be seen that this function H_1 undergoes a phase transition at $\gamma_s = 1$. When $\gamma_s < 1$, the equation $H_1(\lambda, \gamma) = c$ for any $c \leq 0$ has only one solution for λ_s and the function is strictly decreasing. However, for $\gamma_s > 0$, the function H_1 is non-monotone and $H_1(\lambda, \gamma) = c$ can have either none, one, or two solutions for λ_s . We will algebraically prove this fact below.

After simple algebraic manipulations, we find that setting $H_1(\lambda_s, \gamma_s) = c$ for some $c \leq 0$, we can have two potential solutions for λ_s given by

$$\lambda_s^\pm(c, \gamma_s) = -\frac{1 + 2c + 5c^2 + 4c\gamma_s + 4c^2\gamma_s \pm (1 + 3c)\sqrt{(c - 1)^2 + 8c(1 + c)\gamma_s}}{4c(1 + c)}. \quad (12)$$

By inspecting the solutions, we find that when $\gamma_s \leq 1$, only the solution λ_s^- is valid. However, when $\gamma_s > 1$, both λ_s^+ and λ_s^- become valid solutions as long as

$$c < \frac{1}{1 - 4\gamma_s + \sqrt{\gamma_s(\gamma_s - 1)}}, \quad (13)$$

which ensures that $(c - 1)^2 + 8c(1 + c)\gamma_s \geq 0$.

Figure 5 shows the values of λ_s for which $H_2(\lambda_s, \gamma_s) = c$ as a function of c , for different values of γ_s . The case of $\gamma_s = 1$ is shown with a blue dashed line. When $\gamma_s > 1$, two solutions can exist for λ_s . In this case, the largest c for which two solutions exists are given by (13). However, for $\gamma_s < 1$, there is always one solution.

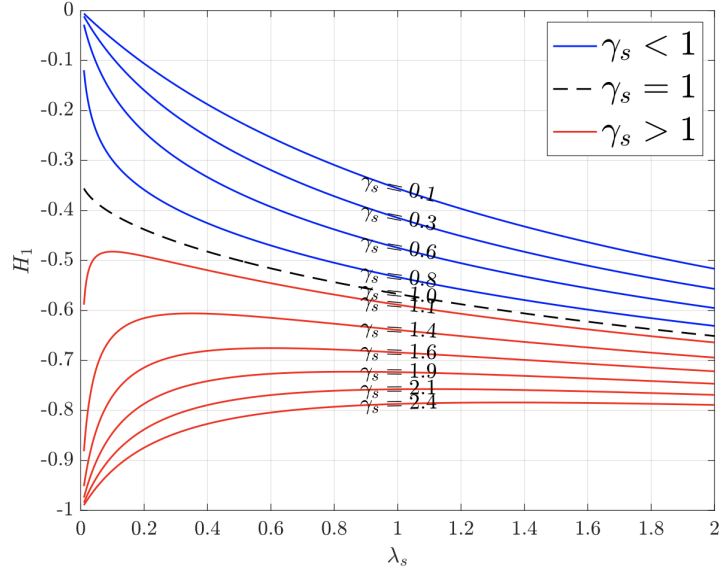


Figure 4: The function H_1 , as a function of λ_s , for different values of γ_s . For the case with $\gamma_s > 1$, the equation $H_1(\lambda_s) = c$ with $c < 0$ can have two solutions. However, for $\gamma_s < 1$, there is always one solution.

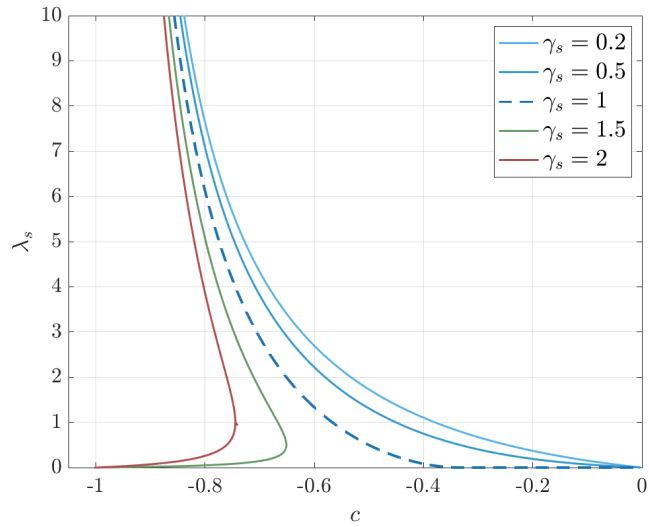


Figure 5: The parameters λ_s for which $H_1(\lambda_s) = c$, for different values of c . Two solutions can exist when $\gamma_s > 1$. However, for $\gamma_s < 1$, only one solution can exist.

Left-Hand Side. We will now turn our attention to the left-hand side of (11). Recall that

$$m_{t,1} = m(\lambda_t; \gamma_t), \quad m_{t,2} = -\frac{\partial}{\partial \lambda} m(\lambda; \gamma)|_{\lambda_t, \gamma_t}.$$

We plug these into the right-hand side of (11), and after simplification, we arrive at

$$H_2 := (\sigma_\varepsilon^2 \gamma_t - \lambda_t) \left(\frac{m_{t,1} - \lambda_t m_{t,2}}{\lambda_t m_{t,1} - 1} \right) = \frac{\sigma_\varepsilon^2 \gamma_t - \lambda_t}{\sqrt{-4\gamma_t + (1 + \gamma_t + \lambda_t)^2}}.$$

Hence, $\sigma_\varepsilon^2 \gamma_t - \lambda_t$ determines the sign of H_2 .

Putting Everything Together. After characterizing the functions H_1 and H_2 , we can use these characterizations to prove the theorem.

- When $\sigma_\varepsilon^2 \gamma_t < \lambda_t$, we have $H_2 > 0$. Noting that $H_1 \leq 0$, we find that in this case, there is no solution for λ_s such that $H_1 = H_2$.
- When $\sigma_\varepsilon^2 \gamma_t \geq \lambda_t$, we have $H_2 \leq 0$. Based on the analysis above for the right-hand side of (11), two cases can happen:
 - If $\gamma_s < 1$, we always have a solution $\bar{\lambda}$ given by $\lambda_s^-(c, \gamma_s)$ from (12) with

$$c = \frac{\sigma_\varepsilon^2 \gamma_t - \lambda_t}{\sqrt{-4\gamma_t + (1 + \gamma_t + \lambda_t)^2}} \quad (14)$$

such that $H_1 = H_2$.

- If $\gamma_s > 1$, as long (13) holds; i.e.,

$$c = \frac{\sigma_\varepsilon^2 \gamma_t - \lambda_t}{\sqrt{-4\gamma_t + (1 + \gamma_t + \lambda_t)^2}} \leq \frac{1}{1 - 4\gamma_s + \sqrt{\gamma_s(\gamma_s - 1)}},$$

two solutions exist for λ_s that satisfy for $H_1 = H_2$. The solutions are given by $(\lambda_s^-, \lambda_s^+)$ from (12). Consequently, we have $\Delta < 1$ as long as $\lambda_s \in (\lambda_s^-, \lambda_s^+)$.

These together finish the proof.

E Non-Monotone Error Curves

We use the theoretical prediction from Theorem 4 and plot the test error of the student model $\mathcal{L}_s$ as a function of γ_t . In Figure 6, we set the teacher regularizer to $\lambda_t \rightarrow 0$ (ridgeless regression) and set $\gamma_s = 0.1, \sigma_\varepsilon = 0.2$. We consider the same setting but with $\sigma_\varepsilon = 1$ in Figure 7. We observe that in both settings, the student has a non-monotone behavior for different values of λ_s , with a peak happening at the interpolation threshold $\gamma_t = 1$.

In Figure 8, we consider the same setting as Figure 7 with $\gamma_s = 0.1, \sigma_\varepsilon = 1$ and set $\lambda_t = \lambda_t^* = \sigma_\varepsilon^2 \gamma_t$ (i.e., the optimal ridge regularizer). We observe that optimal regularization of the teacher model completely mitigates double descent in the student model; i.e., the test loss of the student model becomes monotone as a function of γ_t . This is in line with the findings of [Nakkiran et al., 2021] for standard ridge regression. Also, we observe that as predicted by Theorem 5, the student can never outperform the teacher.

In Figure 9, we consider the same setting as 8 but we set $\lambda_t = 0.15\lambda_t^*$. We observe that in this setting where the teacher is still under-regularized, the student model still exhibits a non-monotone test error as a function of γ_t .

F Proof of Theorem 8

First, recall from Definition 8 that

$$\Gamma = \mathbf{I}_{d_x} + d_x \hat{\beta} \hat{\beta}^\top.$$

Using the Sherman-Morrison formula, we have

$$\Gamma^{-1} = \mathbf{I}_{d_x} - \frac{d_x \hat{\beta} \hat{\beta}^\top}{1 + d_x \hat{\beta}^\top \hat{\beta}} = \mathbf{I}_{d_x} - (1 + o(1)) \hat{\beta} \hat{\beta}^\top.$$

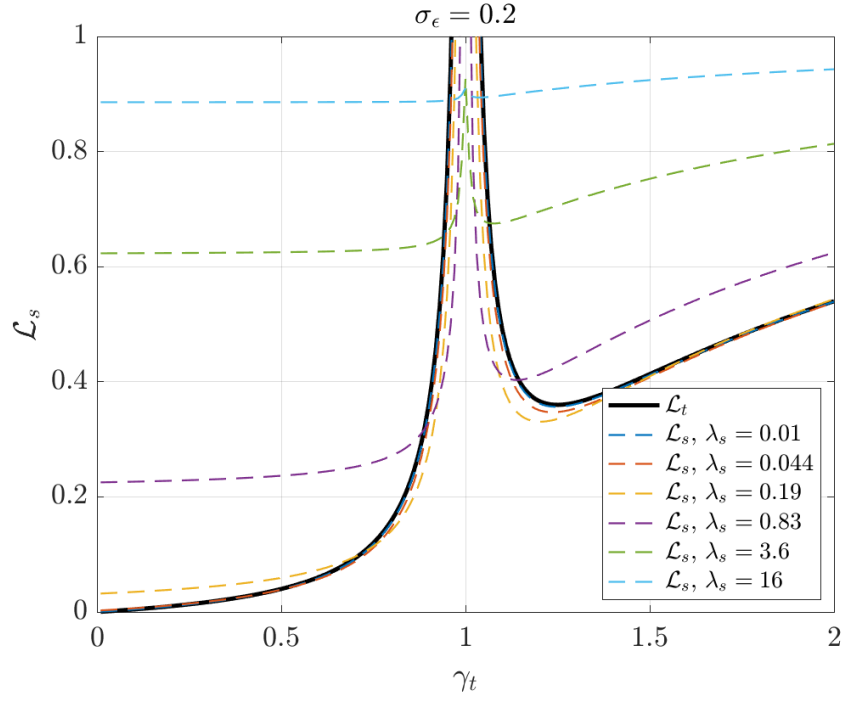


Figure 6: The test error of the student model $\mathcal{L}_s$ as a function of γ_t for $\sigma_\epsilon = 0.2, \gamma_s = 0.1, \lambda_t \rightarrow 0$ (ridgeless), and different values of λ_s .

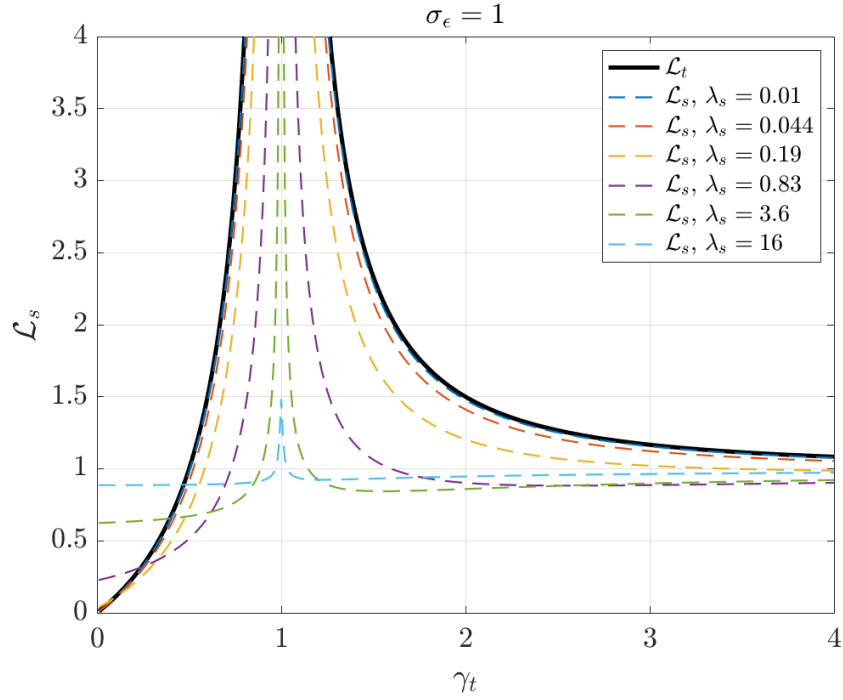


Figure 7: The test error of the student model $\mathcal{L}_s$ as a function of γ_t for $\sigma_\epsilon = 1, \gamma_s = 0.1, \lambda_t \rightarrow 0$ (ridgeless), and different values of λ_s .

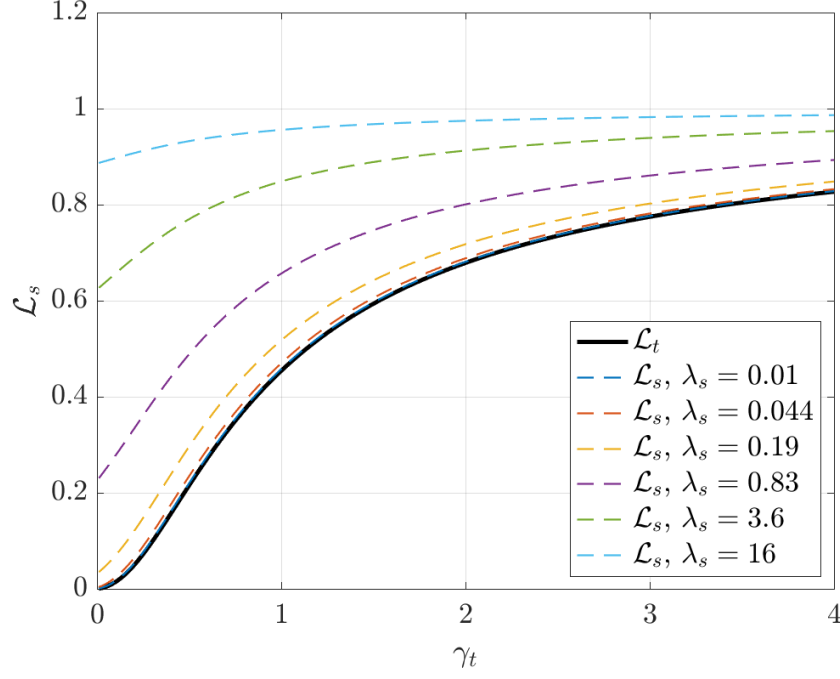


Figure 8: The test error of the student model $\mathcal{L}_s$ as a function of γ_t for $\sigma_\varepsilon = 1, \gamma_s = 0.1, \lambda_t = \lambda_t^* = \sigma_\varepsilon^2 \gamma_t$ (optimal ridge regularizer), and different values of λ_s .

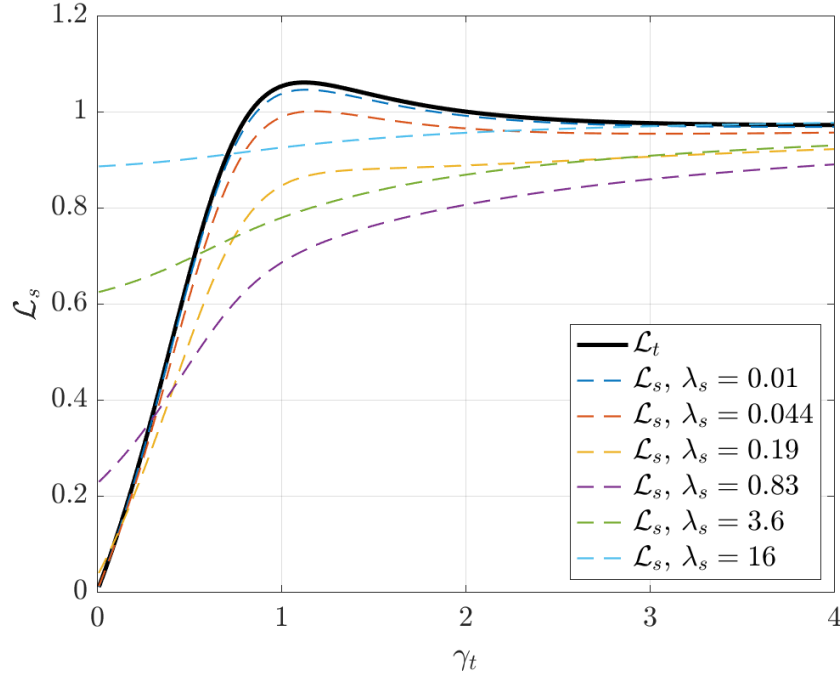


Figure 9: The test error of the student model $\mathcal{L}_s$ as a function of γ_t for $\sigma_\varepsilon = 1, \gamma_s = 0.1, \lambda_t = 0.15\lambda_t^* = 0.15\sigma_\varepsilon^2 \gamma_t$, and different values of λ_s .

In the setting of this theorem, we have $\hat{\beta}_s = (\tilde{\Sigma} + \lambda_s \Gamma^{-1})^{-1} \tilde{\mathbf{X}}^\top \tilde{\mathbf{y}}/n_s$. Thus, we now focus on the generalized resolvent matrix $(\tilde{\Sigma} + \lambda_s \Gamma^{-1})^{-1}$. Using the Sherman-Morrison formula, this matrix can be expanded as

$$(\tilde{\Sigma} + \lambda_s \Gamma^{-1})^{-1} = (\tilde{\Sigma} + \lambda_s \mathbf{I}_{\text{d}_X} - \lambda_s \hat{\beta} \hat{\beta}^\top)^{-1} = \tilde{\mathbf{R}} + \frac{\lambda_s}{1 - \lambda_s \hat{\beta}^\top \tilde{\mathbf{R}} \hat{\beta}} \tilde{\mathbf{R}} \hat{\beta} \hat{\beta}^\top \tilde{\mathbf{R}}.$$

For simplicity, we define the scalar

$$\nu := \frac{\lambda_s}{1 - \lambda_s \hat{\beta}^\top \tilde{\mathbf{R}} \hat{\beta}}.$$

Hence, plugging the Sherman-Morrison expression back into the expression for $\hat{\beta}_s$, we have

$$\begin{aligned} \hat{\beta}_s &= \left[\tilde{\mathbf{R}} + \nu \tilde{\mathbf{R}} \hat{\beta} \hat{\beta}^\top \tilde{\mathbf{R}} \right] \tilde{\mathbf{X}} \tilde{\mathbf{y}}/n_s = \left[\tilde{\mathbf{R}} + \nu \tilde{\mathbf{R}} \hat{\beta} \hat{\beta}^\top \tilde{\mathbf{R}} \right] \tilde{\Sigma} \hat{\beta}_t \\ &= \left[\tilde{\mathbf{R}} + \nu \tilde{\mathbf{R}} \hat{\beta} \hat{\beta}^\top \tilde{\mathbf{R}} \right] \tilde{\Sigma} \left(\mathbf{R} \hat{\Sigma} \beta_\star + \mathbf{R} \mathbf{X}^\top \varepsilon/n_t \right) \\ &= \tilde{\mathbf{R}} \tilde{\Sigma} \mathbf{R} \hat{\Sigma} \beta_\star + \tilde{\mathbf{R}} \tilde{\Sigma} \mathbf{R} \frac{\mathbf{X}^\top \varepsilon}{n_t} + \nu \left[\tilde{\mathbf{R}} \hat{\beta} \hat{\beta}^\top \tilde{\mathbf{R}} \tilde{\Sigma} \mathbf{R} \hat{\Sigma} \beta_\star + \tilde{\mathbf{R}} \hat{\beta} \hat{\beta}^\top \tilde{\mathbf{R}} \tilde{\Sigma} \mathbf{R} \frac{\mathbf{X}^\top \varepsilon}{n_t} \right]. \end{aligned}$$

We define $\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3 \in \mathbb{R}^{\text{d}_X}$ as

$$\begin{aligned} \mathbf{t}_1 &= \tilde{\mathbf{R}} \tilde{\Sigma} \mathbf{R} \hat{\Sigma} \beta_\star - \beta_\star, \quad \mathbf{t}_2 = \tilde{\mathbf{R}} \tilde{\Sigma} \mathbf{R} \frac{\mathbf{X}^\top \varepsilon}{n_t}, \\ \mathbf{t}_3 &= \nu \left[\tilde{\mathbf{R}} \hat{\beta} \hat{\beta}^\top \tilde{\mathbf{R}} \tilde{\Sigma} \mathbf{R} \hat{\Sigma} \beta_\star + \tilde{\mathbf{R}} \hat{\beta} \hat{\beta}^\top \tilde{\mathbf{R}} \tilde{\Sigma} \mathbf{R} \frac{\mathbf{X}^\top \varepsilon}{n_t} \right]. \end{aligned}$$

With this definition, $\mathcal{L}_s = \sigma_\varepsilon^2 + \|\beta_s - \beta_\star\|_2^2$ can be written as

$$\begin{aligned} \mathcal{L}_s &= \sigma_\varepsilon^2 + \|\mathbf{t}_1 + \mathbf{t}_2 + \mathbf{t}_3\|_2^2 \\ &= \|\mathbf{t}_1\|_2^2 + \|\mathbf{t}_2\|_2^2 + \|\mathbf{t}_3\|_2^2 + 2\mathbf{t}_1^\top \mathbf{t}_2 + 2\mathbf{t}_2^\top \mathbf{t}_3 + 2\mathbf{t}_1^\top \mathbf{t}_3. \end{aligned}$$

We will analyze each term separately:

- The first and second terms $\|\mathbf{t}_1\|_2^2 + \|\mathbf{t}_2\|_2^2$ have already been calculated in the proof of Theorem 4, we have

$$\|\mathbf{t}_1\|_2^2 + \|\mathbf{t}_2\|_2^2 - \mathcal{L}_t \rightarrow_{\mathbb{P}} \Delta,$$

where Δ is defined in Theorem 4.

- For the third term, we can write

$$\|\mathbf{t}_3\|_2^2 = (\hat{\beta}^\top \tilde{\mathbf{R}}^2 \hat{\beta}) \cdot \left(\frac{\lambda_s}{1 - \lambda_s \hat{\beta}^\top \tilde{\mathbf{R}} \hat{\beta}} \right)^2 \cdot \left(\hat{\beta}^\top \tilde{\mathbf{R}} \tilde{\Sigma} \mathbf{R} \hat{\Sigma} \beta_\star \right)^2.$$

From the definition of $m_{s,1}$ and $m_{s,2}$ and using the Marchenko-Pastur theorem, we have $\hat{\beta}^\top \tilde{\mathbf{R}}^2 \hat{\beta} \rightarrow_{\mathbb{P}} m_{s,2}$, and also $\hat{\beta}^\top \tilde{\mathbf{R}} \hat{\beta} \rightarrow_{\mathbb{P}} m_{s,1}$. Also, we can write

$$\begin{aligned} \hat{\beta}^\top \tilde{\mathbf{R}} \tilde{\Sigma} \mathbf{R} \hat{\Sigma} \beta_\star &= \zeta \text{d}_X^{-1} \text{Tr} \left(\tilde{\mathbf{R}} \tilde{\Sigma} \mathbf{R} \hat{\Sigma} \right) = \zeta \text{d}_X^{-1} \text{Tr} \left(\tilde{\mathbf{R}} \tilde{\Sigma} \right) \cdot \text{d}_X^{-1} \text{Tr} \left(\hat{\mathbf{R}} \hat{\Sigma} \right) \\ &\rightarrow_{\mathbb{P}} \zeta \cdot (1 - \lambda_s m_{s,1}) \cdot (1 - \lambda_t m_{t,1}), \end{aligned}$$

where ζ is defined in Assumption 7, and we have used the asymptotic freeness of independent Wishart random matrices [Voiculescu, 1991, Capitaine and Donati-Martin, 2007]. Hence,

$$\|\mathbf{t}_3\|_2^2 \rightarrow_{\mathbb{P}} \frac{\lambda_s^2 \zeta^2 m_{s,2}}{(1 - \lambda_s m_{s,1})^2} (1 - \lambda_s m_{s,1})^2 (1 - \lambda_t m_{t,1})^2 = \lambda_s^2 \zeta^2 m_{s,2} (1 - \lambda_t m_{t,1})^2.$$

- For the fourth term, note that

$$2\mathbf{t}_1^\top \mathbf{t}_3 = 2((\tilde{\mathbf{R}} \tilde{\Sigma} \mathbf{R} \hat{\Sigma} - \mathbf{I}_{\text{d}_X}) \hat{\Sigma} \beta_\star)^\top \tilde{\Sigma} \mathbf{R} \mathbf{X}^\top \varepsilon/n_t \rightarrow_{\mathbb{P}} 0,$$

using the Hanson-Wright inequality and the fact that ε is mean zero and independent of all other sources of randomness in the problem.

- Similarly to the fourth term, for the fifth term we can write

$$\begin{aligned}
2\mathbf{t}_2^\top \mathbf{t}_3 &= 2\nu \left(\tilde{\mathbf{R}} \tilde{\Sigma} \hat{\mathbf{R}} \frac{\mathbf{X}^\top \boldsymbol{\varepsilon}}{n_t} \right)^\top \left[\tilde{\mathbf{R}} \hat{\beta} \hat{\beta}^\top \tilde{\mathbf{R}} \tilde{\Sigma} \hat{\mathbf{R}} \hat{\beta}_* + \tilde{\mathbf{R}} \hat{\beta} \hat{\beta}^\top \tilde{\mathbf{R}} \tilde{\Sigma} \hat{\mathbf{R}} \frac{\mathbf{X}^\top \boldsymbol{\varepsilon}}{n_t} \right] \\
&= 2\nu \left(\hat{\beta}^\top \tilde{\mathbf{R}} \tilde{\Sigma} \hat{\mathbf{R}} \hat{\beta}_* \right) \left(n_t^{-1} \boldsymbol{\varepsilon}^\top \mathbf{X} \tilde{\mathbf{R}} \tilde{\Sigma} \hat{\mathbf{R}}^2 \hat{\beta} \right) \\
&\quad + 2\nu \left(n_t^{-1} \hat{\beta}^\top \tilde{\mathbf{R}} \tilde{\Sigma} \hat{\mathbf{R}} \mathbf{X}^\top \boldsymbol{\varepsilon} \right) \left(n_t^{-1} \boldsymbol{\varepsilon}^\top \mathbf{X} \tilde{\mathbf{R}} \tilde{\Sigma} \hat{\mathbf{R}}^2 \hat{\beta} \right).
\end{aligned}$$

Using the Hanson-Wright inequality and the fact that $\boldsymbol{\varepsilon}$ is mean zero and independent of all other sources of randomness in the problem, we have $n_t^{-1} \boldsymbol{\varepsilon}^\top \mathbf{X} \tilde{\mathbf{R}} \tilde{\Sigma} \hat{\mathbf{R}}^2 \hat{\beta} \rightarrow_{\mathbb{P}} 0$. Hence,

$$2\mathbf{t}_2^\top \mathbf{t}_3 \rightarrow_{\mathbb{P}} 0.$$

- The sixth term can be expanded as follows:

$$\begin{aligned}
2\mathbf{t}_1^\top \mathbf{t}_3 &= 2\nu \left(\hat{\beta}^\top \tilde{\mathbf{R}} \tilde{\Sigma} \hat{\mathbf{R}} \hat{\beta}_* + n_t^{-1} \hat{\beta}^\top \tilde{\mathbf{R}} \tilde{\Sigma} \hat{\mathbf{R}} \mathbf{X}^\top \boldsymbol{\varepsilon} \right) \cdot \left(\hat{\beta}^\top \tilde{\mathbf{R}} (\tilde{\mathbf{R}} \tilde{\Sigma} \hat{\mathbf{R}} \hat{\beta}_* - \mathbf{I}_{d_X}) \beta_* \right) \\
&= 2\nu \left(\hat{\beta}^\top \tilde{\mathbf{R}} \tilde{\Sigma} \hat{\mathbf{R}} \hat{\beta}_* \right) \cdot \left(\hat{\beta}^\top \tilde{\mathbf{R}}^2 \tilde{\Sigma} \hat{\mathbf{R}} \hat{\beta}_* - \hat{\beta}^\top \tilde{\mathbf{R}} \beta_* \right) + o_{\mathbb{P}}(1),
\end{aligned}$$

where again we have used the Hanson-Wright inequality and the fact that $\boldsymbol{\varepsilon}$ is mean zero and independent of all other sources of randomness in the problem. Above we have already shown above that

$$\hat{\beta}^\top \tilde{\mathbf{R}} \tilde{\Sigma} \hat{\mathbf{R}} \hat{\beta}_* \rightarrow_{\mathbb{P}} \zeta \cdot (1 - \lambda_s m_{s,1}) \cdot (1 - \lambda_t m_{t,1}).$$

With a similar argument, we have

$$\begin{aligned}
\hat{\beta}^\top \tilde{\mathbf{R}}^2 \tilde{\Sigma} \hat{\mathbf{R}} \hat{\beta}_* &= \zeta \cdot d_X^{-1} \text{Tr} \left(\tilde{\mathbf{R}}^2 \tilde{\Sigma} \hat{\mathbf{R}} \hat{\beta}_* \right) + o_{\mathbb{P}}(1) \\
&= \zeta \cdot d_X^{-1} \text{Tr} \left(\tilde{\mathbf{R}}^2 \tilde{\Sigma} \right) \cdot d_X^{-1} \text{Tr} \left(\hat{\mathbf{R}} \hat{\beta}_* \right) + o_{\mathbb{P}}(1) \\
&\rightarrow_{\mathbb{P}} \zeta (1 - \lambda_t m_{t,1}) \cdot (1 - \lambda_t m_{t,1}) \cdot (m_{s,1} - \lambda_s m_{s,2}).
\end{aligned}$$

Also, $\hat{\beta}^\top \tilde{\mathbf{R}} \beta_* \rightarrow_{\mathbb{P}} \zeta m_{s,1}$. Hence, putting all together, we get

$$2\mathbf{t}_1^\top \mathbf{t}_3 \rightarrow_{\mathbb{P}} 2\zeta^2 \lambda_s (1 - \lambda_t m_{t,1}) [\lambda_t \lambda_s m_{t,1} m_{s,2} - \lambda_t m_{t,1} m_{s,1} - \lambda_s m_{s,1}].$$

Thus, adding all the terms together, we have

$$\mathcal{L}_s - \mathcal{L}_t \rightarrow_{\mathbb{P}} \Delta - \zeta^2 \Delta_{\mathbf{R}}$$

where the expression for Δ is given in Theorem 4, and $\Delta_{\mathbf{R}}$ is given by

$$\Delta_{\mathbf{R}} := \lambda_s (-1 + \lambda_t m_{t,1}) \left[-2\lambda_t m_{s,1} m_{t,1} + \lambda_s m_{s,2} (-1 + \lambda_t m_{t,1}) \right].$$

This concludes the proof.

G Proof of Proposition 10

The training loss for the teacher model is given by

$$\hat{\mathcal{L}}_t := -\frac{1}{n_t} \sum_{i=1}^{n_t} y_i \hat{f}_t(\mathbf{x}_i) = -\frac{1}{n_t} \sum_{i=1}^{n_t} y_i \mathbf{a}_t^\top \sigma(\mathbf{W}_t \mathbf{x}_i).$$

Taking derivatives with respect to the matrix $\mathbf{W}_t$, we arrive at

$$\nabla_{\mathbf{W}_t} \hat{\mathcal{L}}_t = -\frac{1}{n_t} [(\mathbf{a}_t \mathbf{y}^\top) \odot \sigma'(\mathbf{W}_t \mathbf{X}^\top)] \mathbf{X}$$

Let $c_{\sigma,1}$ be the first Hermite coefficient of the activation function σ , and define $\sigma_{\perp} : \mathbb{R} \rightarrow \mathbb{R}$ as

$$\sigma_{\perp}(z) = \sigma(z) - c_{\sigma,1} z, \quad \forall z \in \mathbb{R},$$

where $\mathbb{E}_{z \sim \mathcal{N}(0,1)}[\sigma_\perp(z)] = 0$. Thus, we can write

$$\begin{aligned}\nabla_{\mathbf{W}_t} \widehat{\mathcal{L}}_t &= -\frac{1}{n_t} [(\mathbf{a}_t \mathbf{y}^\top) \odot (c_{\sigma,1} + \sigma'_\perp(\mathbf{W}_{t,0} \mathbf{X}^\top))] \mathbf{X} \\ &= -\frac{c_{\sigma,1}}{n_t} \mathbf{a}_t \mathbf{y}^\top \mathbf{X} - \frac{1}{n_t} [(\mathbf{a}_t \mathbf{y}^\top) \odot \sigma'_\perp(\mathbf{W}_{t,0} \mathbf{X}^\top)] \mathbf{X}\end{aligned}$$

By construction, the matrix $\sigma'_\perp(\mathbf{W}_{t,0} \mathbf{X}^\top)$ has mean zero entries. Thus, using [Vershynin, 2012, Theorem 5.44], we have $\|\sigma'_\perp(\mathbf{W}_{t,0} \mathbf{X}^\top)\|_{\text{op}} = O(\sqrt{n_t})$. Hence, using Lemma 12, we have

$$\begin{aligned}\frac{1}{n_t} \left\| [(\mathbf{a}_t \mathbf{y}^\top) \odot \sigma'_\perp(\mathbf{W}_{t,0} \mathbf{X}^\top)] \mathbf{X} \right\|_{\text{op}} &= \frac{1}{n_t} \left\| \text{diag}(\mathbf{a}_t) \sigma'_\perp(\mathbf{W}_{t,0} \mathbf{X}^\top) \text{diag}(\mathbf{y}) \mathbf{X} \right\|_{\text{op}} \\ &= \frac{1}{n_t} \cdot \frac{\text{polylog}(p_t)}{\sqrt{p_t}} \cdot \sqrt{n_t} \cdot \text{polylog}(n_t) \sqrt{n_t} \\ &= \tilde{O}\left(\frac{1}{\sqrt{p_t}}\right),\end{aligned}$$

where we have used the fact that $\|\mathbf{X}\|_{\text{op}} = O(\sqrt{n_t})$ [Vershynin, 2012, Theorem 7.3.1], and the sub-gaussian maximal inequality to get $\|\mathbf{a}\|_\infty = p_t^{-1/2} \text{polylog}(p_t)$. Similarly, using the sub-Weibull maximal inequality [Kuchibhotla and Chakraborty, 2022, Proposition A.6 and Remark A.1], we have $\|\mathbf{y}\|_\infty = O(\text{polylog}(n_t))$. As a result, for any $\beta \in \mathbb{R}^{\text{dx}}$ with $\|\beta\|_2 = 1$, we have

$$\|\nabla_{\mathbf{W}_t} \widehat{\mathcal{L}}_t \beta\|_2 = n_t^{-1} \beta^\top \mathbf{X}^\top \mathbf{y} + o_{\mathbb{P}}(1).$$

We will now study the case where β is the easy or the hard direction.

Easy direction. First, we let $\beta = \beta_e$. In this case, we have

$$\|\nabla_{\mathbf{W}_t} \widehat{\mathcal{L}}_t \beta_e\|_2 = n_t^{-1} \beta_e^\top \mathbf{X}^\top (\sigma_e(\mathbf{X} \beta_e) + \sigma_h(\mathbf{X} \beta_h)) + o_{\mathbb{P}}(1).$$

Note that $\mathbf{X} \beta_e \in \mathbb{R}^{n_t}$ is a vector of i.i.d. $\mathcal{N}(0, 1)$ entries. Thus, using the weak law of large numbers, we have

$$n_t^{-1} \beta_e^\top \mathbf{X}^\top \sigma_e(\mathbf{X} \beta_e) \rightarrow_{\mathbb{P}} \mathbb{E}_{z \sim \mathcal{N}(0,1)}[z \sigma_e(z)] = c_{\sigma_e,1}.$$

Also, recall the assumption that β_e and β_h are orthonormal vectors and $\mathbf{X}$ is a matrix with i.i.d. $\mathcal{N}(0, 1)$ entries. Thus, $\mathbf{X} \beta_e$ and $\mathbf{X} \beta_h$ are independent and we have

$$n_t^{-1} \beta_e^\top \mathbf{X}^\top \sigma_h(\mathbf{X} \beta_h) \rightarrow_{\mathbb{P}} 0.$$

Thus, the gradient has a non-trivial alignment to the easy direction. Consequently, for $\widehat{\mathbf{W}}_t = \mathbf{W}_{t,1} - \eta_t \nabla_{\mathbf{W}_0} \widehat{\mathcal{L}}_t$, with $\eta_t = \Theta(1)$, we have $\|\widehat{\mathbf{W}}_t \beta_e\|_{\text{op}} \rightarrow_{\mathbb{P}} c > 0$, proving the first part of the proposition.

Hard direction. For the hard direction $\beta = \beta_h$, we have

$$\|\nabla_{\mathbf{W}_t} \widehat{\mathcal{L}}_t \beta_h\|_2 = n_t^{-1} \beta_h^\top \mathbf{X}^\top (\sigma_e(\mathbf{X} \beta_e) + \sigma_h(\mathbf{X} \beta_h)) + o_{\mathbb{P}}(1).$$

The first term $n_t^{-1} \beta_h^\top \mathbf{X}^\top \sigma_e(\mathbf{X} \beta_e)$ can be shown to be $o(1)$ with an argument identical to the argument above. For the second term, note that $\mathbf{X} \beta_h \in \mathbb{R}^{n_t}$ is a vector with independent $\mathcal{N}(0, 1)$ entries. Using the weak law of large numbers, we have

$$n_t^{-1} \beta_h^\top \mathbf{X}^\top \sigma_h(\mathbf{X} \beta_h) \rightarrow_{\mathbb{P}} \mathbb{E}_{z \sim \mathcal{N}(0,1)}[z \sim \sigma_h(z)] = c_{\sigma_h,1} = 0,$$

where we have used the fact that the information exponent of σ_h is larger than one; i.e., $c_{\sigma_h,1} = 0$. This shows that the gradient has no alignment to the hard direction, completing the proof.

H Proof of Theorem 11

From the proof of Proposition 10, we have

$$\widehat{\mathbf{W}}_t = \mathbf{W}_{t,0} + c_{\sigma,1} \eta_t \mathbf{a}_t \hat{\beta}_e^\top + \Delta,$$

where $\hat{\beta}_e = n_t^{-1} \mathbf{X}^\top \mathbf{y}$ and $\|\Delta\|_{\text{op}} = o(1)$. Given the fresh independent set of samples $\tilde{\mathbf{X}}$, the updated teacher model labels them as

$$\tilde{\mathbf{y}} = \tilde{\mathbf{F}} \mathbf{a}_t, \quad \text{with} \quad \tilde{\mathbf{F}} = \sigma(\tilde{\mathbf{X}} \widehat{\mathbf{W}}_t^\top) \in \mathbb{R}^{n_s \times p_t}$$

and the training loss for the student model given by

$$\hat{\mathcal{L}}_s := -\frac{1}{n_s} \sum_{i=1}^{n_s} \tilde{y}_i \hat{f}_t(\tilde{\mathbf{x}}_i) = -\frac{1}{n_s} \sum_{i=1}^{n_s} \tilde{y}_i \mathbf{a}_s^\top \sigma(\mathbf{W}_s \tilde{\mathbf{x}}_i).$$

Taking derivatives with respect to the matrix $\mathbf{W}_t$, we arrive at

$$\nabla_{\mathbf{W}_s} \hat{\mathcal{L}}_s \Big|_{\mathbf{W}_{s,0}} = -\frac{1}{n_s} \left[(\mathbf{a}_s \tilde{\mathbf{y}}^\top) \odot \sigma'(\mathbf{W}_{s,0} \tilde{\mathbf{X}}^\top) \right] \tilde{\mathbf{X}}. \quad (15)$$

To analyze the gradient, we should first characterize $\sigma'(\mathbf{W}_{s,0} \tilde{\mathbf{X}}^\top)$ and $\tilde{\mathbf{y}}$.

Analysis of $\tilde{\mathbf{y}}$. The feature matrix $\tilde{\mathbf{F}}$ is given by

$$\tilde{\mathbf{F}} = \sigma(\tilde{\mathbf{X}} \widehat{\mathbf{W}}_t^\top) = \sigma(\tilde{\mathbf{X}} \mathbf{W}_{t,0}^\top + c_{\sigma,1} \eta_t \tilde{\mathbf{X}} \hat{\beta}_e \mathbf{a}_t^\top),$$

which is a nonlinear transform applied element-wise to a spiked random matrix. Following the recent results in nonlinear random matrix theory (e.g., [Moniri et al. \[2024\]](#), [Wang et al. \[2022\]](#), [Moniri and Hassani \[2024\]](#), [Guionnet et al. \[2023\]](#), [Feldman \[2025\]](#)), in the regime where $\eta_t = \Theta(1)$, we Hermite expand the nonlinearity as follows:

$$\begin{aligned} \tilde{\mathbf{F}} &= \sigma(\tilde{\mathbf{X}} \widehat{\mathbf{W}}_t^\top) = \sigma(\tilde{\mathbf{X}} \mathbf{W}_{t,0}^\top + c_{\sigma,1} \eta_t \tilde{\mathbf{X}} \hat{\beta}_e \mathbf{a}_t^\top) \\ &= \sum_{k=1}^{\infty} c_{\sigma,k} H_k(\tilde{\mathbf{X}} \mathbf{W}_{t,0}^\top + c_{\sigma,1} \eta_t \tilde{\mathbf{X}} \hat{\beta}_e \mathbf{a}_t^\top). \end{aligned}$$

Using Lemma 15 element-wise, we can expand this matrix further

$$\begin{aligned} \tilde{\mathbf{F}} &= \sum_{k=1}^{\infty} \sum_{j=0}^k \binom{k}{j} c_{\sigma,1}^j \eta_t^j c_{\sigma,k} H_{k-j}(\tilde{\mathbf{X}} \mathbf{W}_{t,0}^\top) \odot ((\tilde{\mathbf{X}} \hat{\beta}_e)^{\odot j} \mathbf{a}_t^{\odot j \top}) \\ &= \sum_{k=1}^{\infty} c_{\sigma,k} H_k(\tilde{\mathbf{X}} \mathbf{W}_{t,0}^\top) + \sum_{k=1}^{\infty} c_{\sigma,1}^k \eta_t^k c_{\sigma,k} ((\tilde{\mathbf{X}} \hat{\beta}_e)^{\odot k} \mathbf{a}_t^{\odot k \top}) \\ &\quad + \sum_{k=1}^{\infty} \sum_{j=1}^{k-1} \binom{k}{j} c_{\sigma,1}^j \eta_t^j c_{\sigma,k} H_{k-j}(\tilde{\mathbf{X}} \mathbf{W}_{t,0}^\top) \odot ((\tilde{\mathbf{X}} \hat{\beta}_e)^{\odot j} \mathbf{a}_t^{\odot j \top}). \end{aligned}$$

Note that the first sum can be written as

$$\sum_{k=1}^{\infty} c_{\sigma,k} H_k(\tilde{\mathbf{X}} \mathbf{W}_{t,0}^\top) = \sigma(\tilde{\mathbf{X}} \mathbf{W}_{t,0}^\top).$$

In the second sum, by a simple sub-multiplicativity argument, the k -th term has an operator norm bounded by

$$\left\| c_{\sigma,1}^k \eta_t^k c_{\sigma,k} ((\tilde{\mathbf{X}} \hat{\beta}_e)^{\odot k} \mathbf{a}_t^{\odot k \top}) \right\|_{\text{op}} = O(p_t^{1-k/2})$$

which is $o(\sqrt{p_t})$ when $k > 1$. Moreover, using Lemma 12, the (k, j) -th term of the third sum has an operator upper bounded by

$$\begin{aligned} &\left\| \binom{k}{j} c_{\sigma,1}^j \eta_t^j c_{\sigma,k} H_{k-j}(\tilde{\mathbf{X}} \mathbf{W}_{t,0}^\top) \odot ((\tilde{\mathbf{X}} \hat{\beta}_e)^{\odot j} \mathbf{a}_t^{\odot j \top}) \right\|_{\text{op}} \\ &= \left\| \binom{k}{j} c_{\sigma,1}^j \eta_t^j c_{\sigma,k} \text{diag}((\tilde{\mathbf{X}} \hat{\beta}_e)^{\odot j}) H_{k-j}(\tilde{\mathbf{X}} \mathbf{W}_{t,0}^\top) \text{diag}(\mathbf{a}_t^{\odot j}) \right\|_{\text{op}} \\ &= \tilde{O}(p_t^{-j/2} \cdot n_s^{1/2}) = o(p_t^{1/2}). \end{aligned}$$

Putting everything together, we have

$$\tilde{\mathbf{F}} = \sigma(\tilde{\mathbf{X}}\mathbf{W}_{t,0}^\top) + c_{\sigma,1}^2 \eta_t (\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_e) \mathbf{a}_t^\top + \boldsymbol{\Delta},$$

where $\|\boldsymbol{\Delta}\|_{\text{op}} = o(\sqrt{n_s})$. Hence, recalling that $\|\mathbf{a}_t\|_2 = \Theta(1)$, we have

$$\tilde{\mathbf{y}} = \tilde{\mathbf{F}}\mathbf{a}_t = \sigma(\tilde{\mathbf{X}}\mathbf{W}_{t,0}^\top) \mathbf{a}_t + c_{\sigma,1}^2 \eta_t (\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_e) + \boldsymbol{\delta} \quad (16)$$

in which $\|\boldsymbol{\delta}\|_2 = o(\sqrt{n_s})$.

Derivative Term $\sigma'(\mathbf{X}\mathbf{W}_{s,0}^\top)$. Recall that we have $\mathbf{W}_{s,0} = \bar{\mathbf{W}}_{s,0} + \tau \bar{\mathbf{a}} \boldsymbol{\beta}_h^\top$. Hence,

$$\sigma'(\tilde{\mathbf{X}}\mathbf{W}_{s,0}^\top) = \sigma'(\tilde{\mathbf{X}}\bar{\mathbf{W}}_{s,0}^\top + \tau(\tilde{\mathbf{X}}\boldsymbol{\beta}_h) \bar{\mathbf{a}}^\top).$$

This is again a nonlinearity applied element-wise to a spiked random matrix. Similar to the argument for $\tilde{\mathbf{F}}$, we can use Lemma 15 to write

$$\begin{aligned} \sigma'(\tilde{\mathbf{X}}\mathbf{W}_{s,0}^\top) &= \sum_{k=1}^{\infty} \sum_{j=0}^k \binom{k}{j} \tau^j c_{\sigma',k} H_{k-j} \left(\tilde{\mathbf{X}}\bar{\mathbf{W}}_{s,0}^\top \right) \odot \left((\tilde{\mathbf{X}}\boldsymbol{\beta}_h)^{\odot j} \bar{\mathbf{a}}^{\odot j\top} \right) \\ &= \sum_{k=1}^{\infty} c_{\sigma',k} H_k \left(\tilde{\mathbf{X}}\bar{\mathbf{W}}_{s,0}^\top \right) + \sum_{k=1}^{\infty} \tau^k c_{\sigma',k} \left((\tilde{\mathbf{X}}\boldsymbol{\beta}_h)^{\odot k} \bar{\mathbf{a}}^{\odot k\top} \right) \\ &\quad + \sum_{k=1}^{\infty} \sum_{j=1}^{k-1} \binom{k}{j} \tau^j c_{\sigma',k} H_{k-j} \left(\tilde{\mathbf{X}}\bar{\mathbf{W}}_{s,0}^\top \right) \odot \left((\tilde{\mathbf{X}}\boldsymbol{\beta}_h)^{\odot j} \bar{\mathbf{a}}^{\odot j\top} \right). \end{aligned}$$

Similar to the reasoning used for $\tilde{\mathbf{F}}$, we have

$$\left\| \binom{k}{j} \tau^j c_{\sigma',k} H_{k-j} \left(\tilde{\mathbf{X}}\bar{\mathbf{W}}_{s,0}^\top \right) \odot \left((\tilde{\mathbf{X}}\boldsymbol{\beta}_h)^{\odot j} \bar{\mathbf{a}}^{\odot j\top} \right) \right\|_{\text{op}} = \tilde{O} \left(n_s^{1/2} \left(\frac{\tau}{\sqrt{p_s}} \right)^j \right)$$

which is $o(\sqrt{n_s})$ as long as $\tau = o(\sqrt{p_s})$. Thus, we have

$$\sigma'(\tilde{\mathbf{X}}\mathbf{W}_{s,0}^\top) = \sigma'(\tilde{\mathbf{X}}\bar{\mathbf{W}}_{s,0}^\top) + \sum_{k=1}^{\infty} \tau^k c_{\sigma',k} \left((\tilde{\mathbf{X}}\boldsymbol{\beta}_h)^{\odot k} \bar{\mathbf{a}}^{\odot k\top} \right) + \bar{\boldsymbol{\Delta}} \quad (17)$$

in which $\|\bar{\boldsymbol{\Delta}}\|_{\text{op}} = o(\sqrt{n_s})$.

Gradient of the Loss. Now, we have all the ingredients to study the gradient of the loss function of the student model. Plugging (17) into (15), we have

$$\nabla_{\mathbf{W}_s} \hat{\mathcal{L}}_s \Big|_{\mathbf{W}_{s,0}} = -\frac{1}{n_s} \left[(\mathbf{a}_s \tilde{\mathbf{y}}^\top) \odot \left[\sigma'(\bar{\mathbf{W}}_{s,0} \tilde{\mathbf{X}}^\top) + \sum_{k=1}^{\infty} \tau^k c_{\sigma',k} \left(\bar{\mathbf{a}}^{\odot k} (\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_h)^{\odot k\top} \right) + \bar{\boldsymbol{\Delta}} \right] \right] \tilde{\mathbf{X}},$$

which we decompose as $\nabla_{\mathbf{W}_s} \hat{\mathcal{L}}_s \Big|_{\mathbf{W}_{s,0}} = \mathbf{G}_1 + \mathbf{G}_2$ where $\mathbf{G}_1$ and $\mathbf{G}_2$ are defined as

$$\begin{aligned} \mathbf{G}_1 &= -\frac{1}{n_s} \left[(\mathbf{a}_s \tilde{\mathbf{y}}^\top) \odot \sigma'(\bar{\mathbf{W}}_{s,0} \tilde{\mathbf{X}}^\top) \right] \tilde{\mathbf{X}}, \\ \mathbf{G}_2 &= -\frac{1}{n_s} \left[(\mathbf{a}_s \tilde{\mathbf{y}}^\top) \odot \left(\sum_{k=1}^{\infty} \tau^k c_{\sigma',k} \left(\bar{\mathbf{a}}^{\odot k} (\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_h)^{\odot k\top} \right) + \bar{\boldsymbol{\Delta}} \right) \right] \tilde{\mathbf{X}}. \end{aligned}$$

We will analyze each component separately.

Analysis of $\mathbf{G}_1$. This term can be written as

$$\mathbf{G}_1 = -\frac{1}{n_s} \left[(\mathbf{a}_s \tilde{\mathbf{y}}^\top) \odot \sigma'(\bar{\mathbf{W}}_{s,0} \tilde{\mathbf{X}}^\top) \right] \tilde{\mathbf{X}}.$$

Let $c_{\sigma,1}$ be the first Hermite coefficient of the activation function σ , and define $\sigma_\perp : \mathbb{R} \rightarrow \mathbb{R}$ as

$$\sigma_\perp(z) = \sigma(z) - c_{\sigma,1}z, \quad \forall z \in \mathbb{R},$$

where $\mathbb{E}_{z \sim \mathcal{N}(0,1)}[\sigma_\perp(z)] = 0$. We can write

$$\mathbf{G}_1 = -\mathbf{a}_s \left(\frac{\tilde{\mathbf{y}}^\top \tilde{\mathbf{X}}}{n_s} \right) - \frac{1}{n_s} \left[(\mathbf{a}_s \tilde{\mathbf{y}}^\top) \odot \sigma'_\perp \left(\bar{\mathbf{W}}_{s,0} \tilde{\mathbf{X}}^\top \right) \right] \tilde{\mathbf{X}}. \quad (18)$$

By using Lemma 12 and by a similar argument to the one used in the proof of Proposition 10, the operator norm of the second term of (18) be upper bounded as

$$\left\| \frac{1}{n_s} \left[(\mathbf{a}_s \tilde{\mathbf{y}}^\top) \odot \sigma'_\perp \left(\bar{\mathbf{W}}_{s,0} \tilde{\mathbf{X}}^\top \right) \right] \tilde{\mathbf{X}} \right\|_{\text{op}} = o(1).$$

Using the characterization of $\tilde{\mathbf{y}}$ in (16), the first term of (18) can be written as $-\mathbf{a}_s \hat{\boldsymbol{\beta}}^\top$ with

$$\hat{\boldsymbol{\beta}} = \frac{1}{n_s} \tilde{\mathbf{X}}^\top \tilde{\mathbf{y}} = \frac{1}{n_s} \tilde{\mathbf{X}}^\top \left[\sigma(\tilde{\mathbf{X}} \mathbf{W}_{t,0}^\top) \mathbf{a}_t + c_{\sigma,1}^2 \eta_t(\tilde{\mathbf{X}} \hat{\boldsymbol{\beta}}_e) + \boldsymbol{\delta} \right]$$

This vector aligns to the easy target direction:

$$\begin{aligned} \boldsymbol{\beta}_e^\top \hat{\boldsymbol{\beta}} &= \frac{1}{n_s} (\tilde{\mathbf{X}} \boldsymbol{\beta}_e)^\top \left[\sigma(\tilde{\mathbf{X}} \mathbf{W}_{t,0}^\top) \mathbf{a}_t + c_{\sigma,1}^2 \eta_t(\tilde{\mathbf{X}} \hat{\boldsymbol{\beta}}_e) + \boldsymbol{\delta} \right] \\ &= \frac{1}{n_s} (\tilde{\mathbf{X}} \boldsymbol{\beta}_e)^\top \left[\sigma(\tilde{\mathbf{X}} \mathbf{W}_{t,0}^\top) \mathbf{a}_t + c_{\sigma,1}^2 \eta_t(\tilde{\mathbf{X}} \hat{\boldsymbol{\beta}}_e) \right] + o(1) \\ &= \frac{c_{\sigma,1}^2 \eta_t}{n_s} (\tilde{\mathbf{X}} \boldsymbol{\beta}_e)^\top (\tilde{\mathbf{X}} \hat{\boldsymbol{\beta}}_e) + o(1) \xrightarrow{\mathbb{P}} c_e > 0. \end{aligned} \quad (19)$$

Analysis of $\mathbf{G}_2$. To analyze this component, first note that we can write

$$\mathbf{G}_2 = - \left(\sum_{k=1}^{\infty} \frac{\tau^k c_{\sigma',k}}{n_s} \left((\mathbf{a}_s \odot \bar{\mathbf{a}}^{\odot k}) \left(\tilde{\mathbf{y}} \odot (\tilde{\mathbf{X}} \hat{\boldsymbol{\beta}}_h) \right)^{\odot k^\top} \right) + \bar{\boldsymbol{\Delta}} \right) \tilde{\mathbf{X}}.$$

The operator norm of the k -th term of the sum is given by

$$\left\| \left[\frac{\tau^k c_{\sigma',k}}{n_s} \left((\mathbf{a}_s \odot \bar{\mathbf{a}}^{\odot k}) \left(\tilde{\mathbf{y}} \odot (\tilde{\mathbf{X}} \hat{\boldsymbol{\beta}}_h) \right)^{\odot k^\top} \right) + \bar{\boldsymbol{\Delta}} \right] \tilde{\mathbf{X}} \right\|_{\text{op}} = O \left(\left(\frac{\tau}{\sqrt{p_s}} \right)^k \right) = o(1),$$

where we have used the sub-multiplicativity of the operator norm, $\tau = o(\sqrt{p_s})$, and the fact that $\|\tilde{\mathbf{y}} \odot (\tilde{\mathbf{X}} \hat{\boldsymbol{\beta}}_h)\|_2 = \Theta(\sqrt{n_s})$ and $\|\mathbf{a}_s \odot \bar{\mathbf{a}}^{\odot k}\|_2 = O(p_s^{-k/2})$.

Putting everything together. Using the the results of the analyses above, we have

$$\widehat{\mathbf{W}}_s = \bar{\mathbf{W}}_{s,0} + \tau \bar{\mathbf{a}} \boldsymbol{\beta}_h^\top + \eta_s \mathbf{a}_s \hat{\boldsymbol{\beta}}^\top + \mathring{\boldsymbol{\Delta}}$$

where $\|\mathring{\boldsymbol{\Delta}}\|_{\text{op}} = o(1)$. Hence, recalling (19), the updated weight matrix has non-vanishing correlation with both the easy and hard directions, completing the proof.

NeurIPS Paper Checklist

1. Claims

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

Answer: [\[Yes\]](#)

Guidelines:

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

2. Limitations

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

Answer: [\[Yes\]](#)

Guidelines:

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

3. Theory assumptions and proofs

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

Answer: [\[Yes\]](#)

Guidelines:

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

4. Experimental result reproducibility

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

Answer: [\[Yes\]](#)

Guidelines:

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

5. Open access to data and code

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

Answer: [NA]

Justification: The paper only has simple synthetic experiments with a clear instruction on how to replicate them.

Guidelines:

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

6. Experimental setting/details

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

Answer: [Yes]

Guidelines:

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

7. Experiment statistical significance

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

Answer: [Yes]

Guidelines:

- The answer NA means that the paper does not include experiments.

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

8. Experiments compute resources

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

Answer: [NA]

Justification: The small scale experiments can be easily ran on all typical personal computers.

Guidelines:

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

9. Code of ethics

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

Answer: [Yes]

Guidelines:

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

10. Broader impacts

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

Answer: [NA]

Justification: The focus of this paper is on theoretical aspects of deep learning. We expect the results to be illuminating for the deep learning theory community. We do not anticipate any negative societal impact.

Guidelines:

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

11. Safeguards

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

Answer: [NA]

Guidelines:

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

12. Licenses for existing assets

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

Answer: [NA]

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

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

Answer: [NA]

Guidelines:

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

14. Crowdsourcing and research with human subjects

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

Answer: [NA]

Guidelines:

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

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

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

Answer: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

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

Answer: [NA]

Guidelines:

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