# Supplementary Materials

Guanzhe Hong, Zhiyuan Mao, Xiaojun Lin, Stanley Chan
School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana USA
{hong288, mao114, linx, stanchan}@purdue.edu

## Contents



# 1 Introduction

This document contains the supplementary materials to the paper "Student-Teacher Learning from Clean Inputs to Noisy Inputs". We shall provide the detailed versions of the theorems in the paper and their proofs. We will also provide some extra experimental results demonstrating the utility of student-teacher learning for the $\ell_1$-regularized linear networks, under the setting of section 5 of the paper.

# 2 Proofs for Theorems in Section 3 of the Paper

In this section, we shall present the proof for the Section 3 of the paper.

## 2.1 Notations and Conventions

Consider input-output training data pairs $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N_s} \subset \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$, where $\boldsymbol{x}_i$ is the $i$-th clean training sample, and $\boldsymbol{y}_i$ is the $i$-th target. $\{\boldsymbol{\epsilon}_i\}_{i=1}^{N_s} \subset \mathbb{R}^{d_x}$ are the noise samples.

We write $\boldsymbol{X} \in \mathbb{R}^{d_x \times N_s}$ as the clean input training data matrix, with its columns beings the $\boldsymbol{x}_i$'s. Similarly, we construct the noisy input matrix $\boldsymbol{X}_{\boldsymbol{\epsilon}} \in \mathbb{R}^{d_x \times N_s}$ and target matrix $\boldsymbol{Y} \in \mathbb{R}^{d_y \times N_s}$.

Given matrix $\boldsymbol{M}$, we use row($\boldsymbol{M}$) and col($\boldsymbol{M}$) to denote the row and column spaces of matrix $\boldsymbol{M}$. We use rank($\boldsymbol{M}$) to denote the rank of the matrix. We use $\boldsymbol{P_M}$ to denote the orthogonal projection matrix onto col($\boldsymbol{M}$), and $\boldsymbol{P}_{\boldsymbol{M}}^{\perp}$ for projecting onto col($\boldsymbol{M}$)$^{\perp}$, the orthogonal complement of col($\boldsymbol{M}$). We use $[\boldsymbol{M}]_{i,j}$ to denote the $(i,j)$ entry in $\boldsymbol{M}$. If $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ is symmetric, for its eigen-decomposition $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$, we assume that $[\boldsymbol{\Lambda}]_{1,1} \geq [\boldsymbol{\Lambda}]_{2,2} \geq ... \geq [\boldsymbol{\Lambda}]_{n,n}$.

We consider a general deep linear network

$$\boldsymbol{W_L} = \boldsymbol{W}_L \boldsymbol{W}_{L-1} ... \boldsymbol{W}_1 \tag{1}$$

where $\boldsymbol{W}_i \in \mathbb{R}^{d_i \times d_{i-1}}$. Set $d_0 \coloneqq d_x$ and $d_L \coloneqq d_y$. We restrict $L \geq 2$.

We denote $p \coloneqq \min_{i\in\{0,...,L\}} d_i$. For any $(\boldsymbol{W}_L, ..., \boldsymbol{W}_1)$, clearly rank($\boldsymbol{W_L}$) $\leq p$. We allow the networks to be wide, hence $\min_{i\in\{0,...,L\}} d_i = \min(d_x, d_y)$ is possible.

For convenience, we will sometimes write $\boldsymbol{W}_{i:j} = \boldsymbol{W}_i \boldsymbol{W}_{i-1} ... \boldsymbol{W}_j$. **Caution**: do not confuse this with the matrix notation $[\boldsymbol{W}]_{i,j}$.

## 2.2 Training Losses

We consider two losses specialized to the deep linear networks.

The base loss (**we assume that it is the MSE loss in this whole section**):

$$\begin{aligned}(\boldsymbol{W}_L^{\text{base}}, ..., \boldsymbol{W}_1^{\text{base}}) &= \underset{\boldsymbol{W}_L,...,\boldsymbol{W}_1}{\operatorname{argmin}} \widehat{\mathcal{L}}_{\text{base}}(\boldsymbol{W}_L, ..., \boldsymbol{W}_1) \\ &= \underset{\boldsymbol{W}_L,...,\boldsymbol{W}_1}{\operatorname{argmin}} \|\boldsymbol{W_L X_\epsilon} - \boldsymbol{Y}\|_F^2\end{aligned} \tag{2}$$

To define the student-teacher loss (ST loss), first pick an $i^* \in \{1, ..., L\}$ (we exclude the trivial case $i^* = 0$), and then define

$$\begin{aligned}(\boldsymbol{W}_L^{\text{st}}, ..., \boldsymbol{W}_1^{\text{st}}) &= \underset{\boldsymbol{W}_L,...,\boldsymbol{W}_1}{\operatorname{argmin}} \widehat{\mathcal{L}}_{\text{st}}(\boldsymbol{W}_L, ..., \boldsymbol{W}_1) \\ &= \underset{\boldsymbol{W}_L,...,\boldsymbol{W}_1}{\operatorname{argmin}} \left( \|\boldsymbol{W_L X_\epsilon} - \boldsymbol{Y}\|_F^2 + \lambda \|\boldsymbol{W}_{i^*:1} \boldsymbol{X_\epsilon} - \widetilde{\boldsymbol{W}}_{i^*:1} \boldsymbol{X}\|_F^2 \right)\end{aligned} \tag{3}$$

where we use the tuple $(\widetilde{\boldsymbol{W}}_L, ..., \widetilde{\boldsymbol{W}}_1)$ to denote the teacher's weight. Recall that *the student and teacher share the same architecture.*

## 2.3 Proof of Theorem 1 from Paper

In this subsection, we prove theorem 1 from the paper, i.e. we focus on the undersampling regime $N_s < d_x$. Moreover, we assume that $L = 2$ for the teacher and student. In this case, the hidden dimension of the networks is just $d_1$.

Recall from the paper that we assume the base loss is MSE.

We first restate the theorem from the paper, with all assumptions precisely described.

**Theorem 2.1** (Theorem 1 from paper, detailed version)**.** *Denote $\boldsymbol{W}_i^{base}(t)$ and $\boldsymbol{W}_i^{st}(t)$ as the weights for the student network during training with the the base loss* (2) *and the student-teacher loss* (3)*, respectively.*

*Let the following assumptions hold:*

1. *The optimizer is gradient flow;*
2. *$N_s < d_x$;*
3. *$L = 2$;*
4. *$\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N_s}$ and $\{\boldsymbol{\epsilon}_i\}_{i=1}^{N_s}$ are all sampled independently, and $\boldsymbol{x}$ and $\boldsymbol{\epsilon}$ are continuous random vectors;*

5. *There exists some $\delta > 0$ such that $\|\boldsymbol{W}_i^{base}(0)\|_F \leq \delta$ and $\|\boldsymbol{W}_i^{st}(0)\|_F \leq \delta$ for all $i$;*

6. *The teacher network $(\widetilde{\boldsymbol{W}}_2, \widetilde{\boldsymbol{W}}_1)$ minimizes the training loss for clean data:*

$$(\widetilde{\boldsymbol{W}}_2, \widetilde{\boldsymbol{W}}_1) = \underset{\boldsymbol{W}_2, \boldsymbol{W}_1}{argmin} \widehat{\mathcal{L}}_{teacher}(\boldsymbol{W}_2, \boldsymbol{W}_1) = \underset{\boldsymbol{W}_2, \boldsymbol{W}_1}{argmin} \|\boldsymbol{W}_{\boldsymbol{L}}\boldsymbol{X} - \boldsymbol{Y}\|_F^2 \tag{4}$$

7. *The $\boldsymbol{W}_i^{base}(0)$'s are initialized with the balanced initialization [1], i.e.*

$$\boldsymbol{W}_2^{base}(0)^T \boldsymbol{W}_2^{base}(0) = \boldsymbol{W}_1^{base}(0) \boldsymbol{W}_1^{base}(0)^T \tag{5}$$

8. *The gradient flow successfully converges to a global minimizer for both the MSE- and student-teacher-trained networks;*

9. *The weights $\boldsymbol{W}_i^{st}(t)$ remain in a compact set for $t \in [0, \infty)$. In particular, denote $\|\boldsymbol{W}_i^{st}(t)\|_F \leq M, t \in [0, \infty)$.*

*When $\delta$ is sufficiently small, the following is true almost surely:*

$$\lim_{t \to \infty} \|\boldsymbol{W}_{\boldsymbol{L}}^{base}(t) - \boldsymbol{W}_{\boldsymbol{L}}^{st}(t)\|_F \leq C\delta \tag{6}$$

*where $C$ is a constant independent of $\delta$.*

*Proof.* By lemma 2.2 and lemma 2.3 below, and applying the triangle inequality, we obtain

$$\lim_{t \to \infty} \|\boldsymbol{W}_{\mathbf{L}}^{\text{base}}(t) - \boldsymbol{W}_{\mathbf{L}}^{\text{st}}(t)\|_F \leq C\delta \tag{7}$$

where $C \in \mathcal{O}(M + p^{1/4} \|\boldsymbol{U}_p \boldsymbol{U}_p^T \boldsymbol{Y} (\boldsymbol{X}_{\boldsymbol{\epsilon}}^T \boldsymbol{X}_{\boldsymbol{\epsilon}})^{-1} \boldsymbol{X}_{\boldsymbol{\epsilon}}^T\|_F^{1/2})$ when $\delta$ is sufficiently small ($\boldsymbol{U}_p$ shall be defined below). □

### 2.3.1 Main Lemmas

We define and elaborate on some terms that will be used frequently throughout this subsection.

First recall that $p \coloneqq \min(d_x, d_1, d_y)$. We define the matrix $\boldsymbol{U}_p \in \mathbb{R}^{d_y \times p}$ as follows. The columns of $\boldsymbol{U}_p$ are the dominant $p$ eigenvectors of the matrix $\boldsymbol{Y}\boldsymbol{Y}^T = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ (assuming that the eigenvalues in all the eigen-decompositions are sorted from largest to smallest). Note that if $\text{rank}(\boldsymbol{Y}) < p$, then one can choose arbitrary unit vectors orthogonal to the dominant $\text{rank}(\boldsymbol{Y})$ eigenvectors of $\boldsymbol{Y}\boldsymbol{Y}^T$ as the last $p - \text{rank}(\boldsymbol{Y})$ columns in $\boldsymbol{U}_p$.

**Lemma 2.2** (Bias of MSE-induced Gradient Flow)**.** *With the assumptions in the main theorem, the following holds almost surely:*

$$\lim_{t \to \infty} \boldsymbol{W}_{\boldsymbol{L}}^{base}(t) = \boldsymbol{U}_p \boldsymbol{U}_p^T \boldsymbol{Y} (\boldsymbol{X}_{\boldsymbol{\epsilon}}^T \boldsymbol{X}_{\boldsymbol{\epsilon}})^{-1} \boldsymbol{X}_{\boldsymbol{\epsilon}}^T + \boldsymbol{W}(\delta) \tag{8}$$

*where $\|\boldsymbol{W}(\delta)\|_F \leq C\delta$, for some $C \in \mathcal{O}(p^{1/4}\gamma^{1/2})$, when $\delta$ is sufficiently small, and $\gamma \coloneqq \|\boldsymbol{U}_p \boldsymbol{U}_p^T \boldsymbol{Y} (\boldsymbol{X}_{\boldsymbol{\epsilon}}^T \boldsymbol{X}_{\boldsymbol{\epsilon}})^{-1} \boldsymbol{X}_{\boldsymbol{\epsilon}}^T\|_F$.*

*Proof.* In this proof, for the sake of readability, we will write $\boldsymbol{W}_i(t) = \boldsymbol{W}_i^{\text{base}}(t)$. We also abuse notation a bit by writing $\boldsymbol{W}_i(\infty)$, with the understanding that they mean $\lim_{t \to \infty} \boldsymbol{W}_i(t)$. These limits do exist, due to our assumption that gradient flow converges to a global minimizer.

The proof has three steps:

1. **Structure of the Solution**.

   We prove that

$$\boldsymbol{W}_2(\infty)\boldsymbol{W}_1(\infty) = \boldsymbol{U}_p \boldsymbol{U}_p^T \boldsymbol{Y} (\boldsymbol{X}_{\boldsymbol{\epsilon}}^T \boldsymbol{X}_{\boldsymbol{\epsilon}})^{-1} \boldsymbol{X}_{\boldsymbol{\epsilon}}^T + \boldsymbol{W}_2(\infty)\boldsymbol{W}_1(0)_\perp \tag{9}$$

   where we orthogonally decomposed the row space $\boldsymbol{W}_1(0) = \boldsymbol{W}_1(0)_\parallel + \boldsymbol{W}_1(0)_\perp$, where $\text{row}(\boldsymbol{W}_1(0)_\parallel) \subseteq \text{col}(\boldsymbol{X}_{\boldsymbol{\epsilon}})$ and $\text{row}(\boldsymbol{W}_1(0)_\perp) \subseteq \text{col}(\boldsymbol{X}_{\boldsymbol{\epsilon}})^\perp$.

We begin by observing the updates made by gradient flow to $\boldsymbol{W}_1$:

$$\frac{\partial \boldsymbol{W}_1}{\partial t} = \eta(\boldsymbol{W}_2(t)^T \boldsymbol{Y} \boldsymbol{X}_{\boldsymbol{\epsilon}}^T - \boldsymbol{W}_2(t)^T \boldsymbol{W}_2(t) \boldsymbol{W}_1(t) \boldsymbol{X}_{\boldsymbol{\epsilon}} \boldsymbol{X}_{\boldsymbol{\epsilon}}^T). \tag{10}$$

Here, $\eta$ is the update step size, and assumed to be close to 0. As explained in [1] section 5, when $\eta^2 \approx 0$, the discrete gradient descent steps translate into the gradient flow differential equation. The right-hand side of this differential equation is simply the derivative of the base MSE loss (2) with respect to $\boldsymbol{W}_1$.

Notice that $\text{row}\left(\frac{\partial \boldsymbol{W}_1}{\partial t}\right) \subseteq \text{col}(\boldsymbol{X}_{\boldsymbol{\epsilon}})$ at all time. We have the following:

$$\text{row}(\boldsymbol{W}_1(\infty) - \boldsymbol{W}_1(0)) = \text{row}\left(\int_{t=0}^{\infty} \frac{\partial \boldsymbol{W}_1}{\partial t} dt\right) \subseteq \text{col}(\boldsymbol{X}_{\boldsymbol{\epsilon}}). \tag{11}$$

The infinite integral is well-defined since we assumed the convergence of gradient flow. The above observation, combined with our definition of $\boldsymbol{W}_1(0)_{\parallel}$ and $\boldsymbol{W}_1(0)_{\perp}$ from before, imply that gradient flow only modifies $\boldsymbol{W}_1(0)_{\parallel}$, and leaves the $\boldsymbol{W}_1(0)_{\perp}$ untouched. In other words, decomposing the row vectors of $\boldsymbol{W}_1(\infty)$ orthogonally w.r.t $\text{col}(\boldsymbol{X}_{\boldsymbol{\epsilon}})$ (identical to what we did with $\boldsymbol{W}_1(0)$), we can write

$$\boldsymbol{W}_1(\infty) = \boldsymbol{W}_1(\infty)_{\parallel} + \boldsymbol{W}_1(\infty)_{\perp} = \boldsymbol{W}_1(\infty)_{\parallel} + \boldsymbol{W}_1(0)_{\perp}. \tag{12}$$

The important point to notice is that,

$$\boldsymbol{W}_2(\infty)\boldsymbol{W}_1(\infty) = \boldsymbol{W}_2(\infty)\boldsymbol{W}_1(\infty)_{\parallel} + \boldsymbol{W}_1(\infty)\boldsymbol{W}_1(0)_{\perp}. \tag{13}$$

Recalling the expression of global minimizers stated in Lemma 2.4, with probability 1 (over the randomness in the training sample matrix $\boldsymbol{X}_{\epsilon}$), all the global minimizers share exactly the same structure as we have for $\boldsymbol{W}_2(\infty)\boldsymbol{W}_1(\infty)$, i.e. these minimizers consist of two terms, first, the minimum-Frobenius-norm solution $\boldsymbol{U}_p \boldsymbol{U}_p^T \boldsymbol{Y} (\boldsymbol{X}_{\boldsymbol{\epsilon}}^T \boldsymbol{X}_{\boldsymbol{\epsilon}})^{-1} \boldsymbol{X}_{\boldsymbol{\epsilon}}^T$ whose row space lies in $\text{col}(\boldsymbol{X}_{\boldsymbol{\epsilon}})$, and second, the "residue matrix" $\boldsymbol{R}$ whose row space lies in $\text{col}(\boldsymbol{X}_{\boldsymbol{\epsilon}})^{\perp}$. It follows that $\boldsymbol{W}_2(\infty)\boldsymbol{W}_1(\infty)_{\parallel} = \boldsymbol{U}_p \boldsymbol{U}_p^T \boldsymbol{Y} (\boldsymbol{X}_{\boldsymbol{\epsilon}}^T \boldsymbol{X}_{\boldsymbol{\epsilon}})^{-1} \boldsymbol{X}_{\boldsymbol{\epsilon}}^T$, which finishes the first step of the overall proof.

2. **Uniform Upper Bound on $\|\boldsymbol{W}_2(\infty)\|_F$ for Small and Balanced Initialization**.

   We relate $\|\boldsymbol{W}_2(\infty)\|_F$ to $\|\boldsymbol{W}_2(\infty)\boldsymbol{W}_1(\infty)\|_F$.

   Let's denote the SVDs of $\boldsymbol{W}_2(\infty) = \boldsymbol{U}^{(2)} \boldsymbol{\Lambda}^{(2)} \boldsymbol{V}^{(2)T}$, and $\boldsymbol{W}_1(\infty) = \boldsymbol{U}^{(1)} \boldsymbol{\Lambda}^{(1)} \boldsymbol{V}^{(1)T}$.

   A deep linear network that is initialized in the balanced fashion remains balanced throughout training [1] (theorem 1), therefore, we have that $\boldsymbol{W}_2^T(\infty)\boldsymbol{W}_2(\infty) = \boldsymbol{W}_1(\infty)\boldsymbol{W}_1^T(\infty)$, which means that

   $$\boldsymbol{V}^{(2)} \boldsymbol{\Lambda}^{(2)T} \boldsymbol{\Lambda}^{(2)} \boldsymbol{V}^{(2)T} = \boldsymbol{U}^{(1)} \boldsymbol{\Lambda}^{(1)} \boldsymbol{\Lambda}^{(1)T} \boldsymbol{U}^{(1)T}. \tag{14}$$

   In other words, $\boldsymbol{\Lambda}^{(2)T} \boldsymbol{\Lambda}^{(2)} = \boldsymbol{\Lambda}^{(1)} \boldsymbol{\Lambda}^{(1)T}$, i.e. $[\boldsymbol{\Lambda}^{(2)}]_{i,i} = [\boldsymbol{\Lambda}^{(1)}]_{i,i}$ for $i \in \{1, ..., d_1\}$, and the orthogonal matrices $\boldsymbol{V}^{(2)}$ and $\boldsymbol{U}^{(1)}$ are equal up to some rotation in the eigenspaces corresponding to each eigenvalue in $\boldsymbol{\Lambda}^{(2)T} \boldsymbol{\Lambda}^{(2)}$ (see the details in [1] Appendix A.1). It also follows that, $\text{rank}(\boldsymbol{\Lambda}^{(1)}) = \text{rank}(\boldsymbol{\Lambda}^{(2)}) \leq p = \min(d_x, d_1, d_y)$. Using equations (23) and (24) (and the equations before these two) from [1], it follows that

   $$\begin{aligned} \|\boldsymbol{W}_2(\infty)\boldsymbol{W}_1(\infty)\|_F &= \|\boldsymbol{\Lambda}^{(2)} \boldsymbol{\Lambda}^{(2)T}\|_F \\ &= \sqrt{\sum_{i=1}^{p} [\boldsymbol{\Lambda}^{(2)}]_{i,i}^4}. \end{aligned} \tag{15}$$

   Recall that, by Hölder's inequality, $\|\boldsymbol{x}\|_1 \leq \sqrt{p}\|\boldsymbol{x}\|_2$ for any $\boldsymbol{x} \in \mathbb{R}^p$. Therefore

   $$\|\boldsymbol{W}_2(\infty)\|_F^2 = \sum_{i=1}^{p} [\boldsymbol{\Lambda}^{(2)}]_{i,i}^2 \leq \sqrt{p} \sqrt{\sum_{i=1}^{p} [\boldsymbol{\Lambda}^{(2)}]_{i,i}^4} = \sqrt{p} \|\boldsymbol{W}_2(\infty)\boldsymbol{W}_1(\infty)\|_F. \tag{16}$$

Let's study the term $\|\boldsymbol{W}_2(\infty)\boldsymbol{W}_1(\infty)\|_F$. By the Pythagorean theorem, $\|\boldsymbol{W}_2(\infty)\boldsymbol{W}_1(\infty)_\|\|_F^2 + \|\boldsymbol{W}_2(\infty)\boldsymbol{W}_1(0)_\perp\|_F^2 = \|\boldsymbol{W}_2(\infty)\boldsymbol{W}_1(\infty)\|_F^2$. Since $\|\boldsymbol{W}_1(0)\|_F = \delta$, and recalling the definition $\gamma \coloneqq \|\boldsymbol{U}_p\boldsymbol{U}_p^T\boldsymbol{Y}(\boldsymbol{X}_{\boldsymbol{\epsilon}}^T\boldsymbol{X}_{\boldsymbol{\epsilon}})^{-1}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T\|_F$, we have

$$\begin{aligned} &\|\boldsymbol{W}_2(\infty)\boldsymbol{W}_1(\infty)\|_F^2 \le \gamma^2 + \delta^2\|\boldsymbol{W}_2(\infty)\|_F^2 \\ \implies &\|\boldsymbol{W}_2(\infty)\boldsymbol{W}_1(\infty)\|_F \le \sqrt{\gamma^2 + \delta^2\|\boldsymbol{W}_2(\infty)\|_F^2} < \gamma + \delta\|\boldsymbol{W}_2(\infty)\|_F. \end{aligned} \tag{17}$$

Therefore,

$$\begin{aligned} &\|\boldsymbol{W}_2(\infty)\|_F^2 < \sqrt{p}\gamma + \sqrt{p}\delta\|\boldsymbol{W}_2(\infty)\|_F \\ \iff &\|\boldsymbol{W}_2(\infty)\|_F^2 - \sqrt{p}\delta\|\boldsymbol{W}_2(\infty)\|_F - \sqrt{p}\gamma < 0 \\ \iff &\frac{\sqrt{p}\delta - \sqrt{p\delta^2 + 4\sqrt{p}\gamma}}{2} < \|\boldsymbol{W}_2(\infty)\|_F < \frac{\sqrt{p}\delta + \sqrt{p\delta^2 + 4\sqrt{p}\gamma}}{2} \\ \implies &\|\boldsymbol{W}_2(\infty)\|_F < \frac{\sqrt{p}\delta + \sqrt{p\delta^2 + 4\sqrt{p}\gamma}}{2} < \sqrt{p}\delta + p^{1/4}\gamma^{1/2}. \end{aligned} \tag{18}$$

The upper bound is clearly $\mathcal{O}(p^{1/4}\gamma^{1/2})$ for $\delta$ sufficiently small.

3. **Conclusion**.

   The desired result now follows by combining 1. and 2., and by applying Cauchy-Schwartz to $\|\boldsymbol{W}_2(\infty)\boldsymbol{W}_1(0)_\perp\|_F$, with $C = \sqrt{p}\delta + p^{1/4}\gamma^{1/2}$.

□

**Lemma 2.3** (Bias of Student-teacher-induced Gradient Flow)**.** *With the assumptions in the main theorem, the following holds:*

$$\lim_{t\to\infty} \boldsymbol{W}_{\boldsymbol{L}}^{st}(t) = \boldsymbol{U}_p\boldsymbol{U}_p^T\boldsymbol{Y}(\boldsymbol{X}_{\boldsymbol{\epsilon}}^T\boldsymbol{X}_{\boldsymbol{\epsilon}})^{-1}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T + \boldsymbol{W}(\delta) \tag{19}$$

*where $\|\boldsymbol{W}(\delta)\|_F \le M\delta$; recall that we assumed $\|\boldsymbol{W}_i^{st}(t)\| \le M$ for all $t \in [0,\infty)$. In other words, small initialization leads to $\lim_{t\to\infty}\boldsymbol{W}_{\boldsymbol{L}}(t) \approx \boldsymbol{U}_p\boldsymbol{U}_p^T\boldsymbol{Y}(\boldsymbol{X}_{\boldsymbol{\epsilon}}^T\boldsymbol{X}_{\boldsymbol{\epsilon}})^{-1}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T$.*

*Proof.* We write $\boldsymbol{W}_i(t) = \boldsymbol{W}_i^{\mathrm{st}}(t)$ for notational simplicity.

First observe that

$$\frac{\partial \boldsymbol{W}_1}{\partial t} = \eta\boldsymbol{W}_2(t)^T(\boldsymbol{Y} - \boldsymbol{W}_2(t)\boldsymbol{W}_1(t)\boldsymbol{X}_{\boldsymbol{\epsilon}})\boldsymbol{X}_{\boldsymbol{\epsilon}}^T + \eta\lambda(\widetilde{\boldsymbol{W}}_1\boldsymbol{X} - \boldsymbol{W}_1(t)\boldsymbol{X}_{\boldsymbol{\epsilon}})\boldsymbol{X}_{\boldsymbol{\epsilon}}^T. \tag{20}$$

It follows that $\mathrm{row}\left(\frac{\partial \boldsymbol{W}_1}{\partial t}\right) \subseteq \mathrm{col}(\boldsymbol{X}_{\boldsymbol{\epsilon}})$. Therefore, arguing similarly to step 1 of the proof of lemma 2.2, we may write $\boldsymbol{W}_1(t) = \boldsymbol{W}_1(t)_\| + \boldsymbol{W}_1(t)_\perp = \boldsymbol{W}_1(t)_\| + \boldsymbol{W}_1(0)_\perp$, where $\mathrm{row}(\boldsymbol{W}_1(t)_\|) \subseteq \mathrm{col}(\boldsymbol{X}_{\boldsymbol{\epsilon}})$, and $\mathrm{row}(\boldsymbol{W}_1(0)_\perp) \subseteq \mathrm{col}(\boldsymbol{X}_{\boldsymbol{\epsilon}})^\perp$.

Knowing the form of global minimizers of the ST loss from Lemma 2.5, we know

$$\boldsymbol{W}_{\mathbf{L}}(\infty) = \boldsymbol{U}_p\boldsymbol{U}_p^T\boldsymbol{Y}(\boldsymbol{X}_{\boldsymbol{\epsilon}}^T\boldsymbol{X}_{\boldsymbol{\epsilon}})^{-1}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T + \boldsymbol{W}_2(\infty)\boldsymbol{W}_1(0)_\perp \tag{21}$$

Therefore $\|\boldsymbol{W}_{\mathbf{L}}(\infty) - \boldsymbol{U}_p\boldsymbol{U}_p^T\boldsymbol{Y}(\boldsymbol{X}_{\boldsymbol{\epsilon}}^T\boldsymbol{X}_{\boldsymbol{\epsilon}})^{-1}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T\|_F \le M\delta$. □

### 2.3.2 Auxilliary Lemmas

**Lemma 2.4** (Global minimizers of MSE loss (2), $N < d_x$)**.** *The set of global minimizers to the MSE loss (2) is the following almost surely (over the randomness of the training samples):*

$$\{\boldsymbol{W}_2, \boldsymbol{W}_1 |\ \boldsymbol{W}_2\boldsymbol{W}_1 = \boldsymbol{U}_p\boldsymbol{U}_p^T\boldsymbol{Y}(\boldsymbol{X}_{\boldsymbol{\epsilon}}^T\boldsymbol{X}_{\boldsymbol{\epsilon}})^{-1}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T + \boldsymbol{R},\ row(\boldsymbol{R}) \subseteq col(\boldsymbol{X}_{\boldsymbol{\epsilon}})^\perp\} \tag{22}$$

*where the columns of $\boldsymbol{U}_p$ are the dominant $p$ eigenvectors of the matrix $\boldsymbol{Y}\boldsymbol{Y}^T = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$. Note that if rank$(\boldsymbol{Y}) < p$, then one can choose arbitrary unit vectors orthogonal to the dominant rank$(\boldsymbol{Y})$ eigenvectors of $\boldsymbol{Y}\boldsymbol{Y}^T$ as the last $p -$ rank$(\boldsymbol{Y})$ columns in $\boldsymbol{U}_p$.*

*Proof.* First of all, note that since $\boldsymbol{x}$ and $\boldsymbol{\epsilon}$ are continuous random vectors, $\boldsymbol{X}_\epsilon$ must be full rank almost surely, so $(\boldsymbol{X}_{\boldsymbol{\epsilon}}^T \boldsymbol{X}_{\boldsymbol{\epsilon}})^{-1}$ exists.

Now, we note that

$$\begin{aligned}&\{\boldsymbol{W} \in \mathbb{R}^{d_y \times d_x} | (\text{rank}(\boldsymbol{W}) \leq p) \ \wedge \ (\boldsymbol{W} \text{ minimizes } \|\boldsymbol{W}'\boldsymbol{X}_\epsilon - \boldsymbol{Y}\|_F^2\} \\ =&\{\boldsymbol{W}_2\boldsymbol{W}_1 | (\boldsymbol{W}_2 \in \mathbb{R}^{d_y \times d_1}) \wedge (\boldsymbol{W}_1 \in \mathbb{R}^{d_1 \times d_x}) \wedge ((\boldsymbol{W}_2, \boldsymbol{W}_1) \text{ minimizes (2)})\}.\end{aligned} \tag{23}$$

To see "$\subseteq$" direction, take any $\boldsymbol{W}$ in the first set, we can decompose it as $\boldsymbol{W} = \boldsymbol{A}_W \boldsymbol{B}_W$ where $\boldsymbol{A}_W \in \mathbb{R}^{d_y \times d_1}$ and $\boldsymbol{B}_W \in \mathbb{R}^{d_1 \times d_x}$, and $(\boldsymbol{A}_W \boldsymbol{B}_W)$ clearly minimizes (2). The other direction can also be easily shown.

It follows that, the set of solutions $\boldsymbol{W}_{\mathbf{L}} = \boldsymbol{W}_2\boldsymbol{W}_1$ that we need to solve for is the same as the set

$$\underset{\text{rank}(\boldsymbol{W}) \leq d_1}{\text{argmin}} \ \|\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{X}_{\boldsymbol{\epsilon}}\|_F^2. \tag{24}$$

This is basically a low-rank approximation problem. By the Eckart-Young-Mirsky theorem [4], the matrix $\boldsymbol{U}_p\boldsymbol{U}_p^T\boldsymbol{Y}$ is the best approximation to the matrix $\boldsymbol{Y}$ under the Frobenius norm with rank no greater than $p$.

To achieve this solution, we need

$$\begin{aligned}&\boldsymbol{W}\boldsymbol{X}_{\boldsymbol{\epsilon}} = \boldsymbol{U}_p\boldsymbol{U}_p^T\boldsymbol{Y} \\ \iff &\boldsymbol{W}\boldsymbol{X}_{\boldsymbol{\epsilon}} = \boldsymbol{U}_p\boldsymbol{U}_p^T\boldsymbol{Y}(\boldsymbol{X}_{\boldsymbol{\epsilon}}^T\boldsymbol{X}_{\boldsymbol{\epsilon}})^{-1}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T\boldsymbol{X}_{\boldsymbol{\epsilon}} \\ \iff &(\boldsymbol{W} - \boldsymbol{U}_p\boldsymbol{U}_p^T\boldsymbol{Y}(\boldsymbol{X}_{\boldsymbol{\epsilon}}^T\boldsymbol{X}_{\boldsymbol{\epsilon}})^{-1}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)\boldsymbol{X}_{\boldsymbol{\epsilon}} = \boldsymbol{0} \\ \iff &\boldsymbol{W} - \boldsymbol{U}_p\boldsymbol{U}_p^T\boldsymbol{Y}(\boldsymbol{X}_{\boldsymbol{\epsilon}}^T\boldsymbol{X}_{\boldsymbol{\epsilon}})^{-1}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T = \boldsymbol{R}, \ \text{ s.t. } \text{row}(\boldsymbol{R}) \subseteq \text{col}(\boldsymbol{X}_{\boldsymbol{\epsilon}})^\perp.\end{aligned} \tag{25}$$

which concludes the proof. □

**Lemma 2.5** (Global Minimizers of ST loss (3), $N < d_x$)**.** *Choose a global minimizer* $(\widetilde{\boldsymbol{W}}_2, \widetilde{\boldsymbol{W}}_1)$ *of* (4)*. Then almost surely, the set of global minimizers to* (3) *is the following:*

$$\begin{aligned}\{\boldsymbol{W}_2 \in \mathbb{R}^{d_y \times d_1}, \boldsymbol{W}_1 \in \mathbb{R}^{d_1 \times d_x} | \ &\boldsymbol{W}_2\boldsymbol{W}_1 = \boldsymbol{U}_p\boldsymbol{U}_p^T\boldsymbol{Y}(\boldsymbol{X}_{\boldsymbol{\epsilon}}^T\boldsymbol{X}_{\boldsymbol{\epsilon}})^{-1}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T + (\widetilde{\boldsymbol{W}}_2 + \boldsymbol{R}_2)\boldsymbol{R}_1, \\ &\boldsymbol{R}_1 \in \mathbb{R}^{d_1 \times d_x} \wedge \textit{row}(\boldsymbol{R}_1) \subseteq \textit{col}(\boldsymbol{X}_{\boldsymbol{\epsilon}})^\perp \wedge \boldsymbol{R}_2 \in \mathbb{R}^{d_x \times d_1} \wedge \textit{row}(\boldsymbol{R}_2) \subseteq \textit{col}(\widetilde{\boldsymbol{W}}_1\boldsymbol{X})^\perp\}.\end{aligned} \tag{26}$$

*Remark.* Notice that this solution set is just a subset of the MSE solution set from lemma 2.4. In the MSE solution set, the "residue matrix" $\boldsymbol{R}$ just satisfies $\text{row}(\boldsymbol{R}) \subseteq \text{col}(\boldsymbol{X}_\epsilon)^\perp$. For this ST loss solution set, the "residue matrix" also satisfies $\text{row}((\widetilde{\boldsymbol{W}}_2 + \boldsymbol{R}_2)\boldsymbol{R}_1) \subseteq \text{col}(\boldsymbol{X}_\epsilon)^\perp$, although it does have more structure.

*Proof.* Like in the last lemma, note that since $\boldsymbol{x}$ and $\boldsymbol{\epsilon}$ are continuous random vectors, $\boldsymbol{X}_\epsilon$ must be full rank almost surely, so $(\boldsymbol{X}_{\boldsymbol{\epsilon}}^T \boldsymbol{X}_{\boldsymbol{\epsilon}})^{-1}$ exists.

The proof relies on the fact that,

$$\begin{aligned}&\min_{\boldsymbol{W}_2, \boldsymbol{W}_1} \left\{\|\boldsymbol{Y} - \boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X}_{\boldsymbol{\epsilon}}\|_F^2 + \lambda\|\widetilde{\boldsymbol{W}}_1\boldsymbol{X} - \boldsymbol{W}_1\boldsymbol{X}_{\boldsymbol{\epsilon}}\|_F^2\right\} \\ \geq &\min_{\boldsymbol{W}_2, \boldsymbol{W}_1} \left\{\|\boldsymbol{Y} - \boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X}_{\boldsymbol{\epsilon}}\|_F^2\right\} + \min_{\boldsymbol{W}_1} \left\{\lambda\|\widetilde{\boldsymbol{W}}_1\boldsymbol{X} - \boldsymbol{W}_1\boldsymbol{X}_{\boldsymbol{\epsilon}}\|_F^2\right\}\end{aligned} \tag{27}$$

We can see that the lower bound is achievable only when we individually minimize the two loss terms in the lower bound, in other words, denoting $l_1(\boldsymbol{W}_2, \boldsymbol{W}_1) = \|\boldsymbol{Y} - \boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X}_{\boldsymbol{\epsilon}}\|_F^2$, and $l_2(\boldsymbol{W}_1) = \|\widetilde{\boldsymbol{W}}_1\boldsymbol{X} - \boldsymbol{W}_1\boldsymbol{X}_{\boldsymbol{\epsilon}}\|_F^2$, equality is true only for $(\boldsymbol{W}_2, \boldsymbol{W}_1)$ that lies in the following intersection

$$\{\boldsymbol{W}_2, \boldsymbol{W}_1 | \boldsymbol{W}_2, \boldsymbol{W}_1 \text{ minimizes } l_1\} \cap \{\boldsymbol{W}_1 | \boldsymbol{W}_1 \text{ minimizes } l_2\} \tag{28}$$

We proceed to minimize the two terms individually.

First notice that the regularizer can be made 0 with the following set of expressions for $\boldsymbol{W}_1$

$$\{\boldsymbol{W}_1 |\ \boldsymbol{W}_1 = \widetilde{\boldsymbol{W}}_1 \boldsymbol{X}(\boldsymbol{X}_{\boldsymbol{\epsilon}}^T \boldsymbol{X}_{\boldsymbol{\epsilon}})^{-1} \boldsymbol{X}_{\boldsymbol{\epsilon}}^T + \boldsymbol{R}_1,\ \boldsymbol{R}_1 \in \mathbb{R}^{d_1 \times d_x} \wedge \text{row}(\boldsymbol{R}_1) \subseteq \text{col}(\boldsymbol{X}_{\boldsymbol{\epsilon}})^{\perp}\} \tag{29}$$

Its proof is very similar to (25).

We already know the set of minimizers of the MSE loss from the previous section. We take the intersection of the two sets. First note that $\boldsymbol{W}_1 \boldsymbol{X}_{\boldsymbol{\epsilon}} = \widetilde{\boldsymbol{W}}_1 \boldsymbol{X}$, therefore, for $\boldsymbol{W}_2, \boldsymbol{W}_1$ to minimize the MSE loss $l_1$, we need $\boldsymbol{W}_2 \widetilde{\boldsymbol{W}}_1 \boldsymbol{X} = \boldsymbol{U}_p \boldsymbol{U}_p^T \boldsymbol{Y}$. But since $(\widetilde{\boldsymbol{W}}_2, \widetilde{\boldsymbol{W}}_1)$ minimizes (4), $\widetilde{\boldsymbol{W}}_2 \widetilde{\boldsymbol{W}}_1 \boldsymbol{X} = \boldsymbol{U}_p \boldsymbol{U}_p^T \boldsymbol{Y}$ has to be true (can be proven using essentially the same argument as in the previous lemma). In other words,

$$\boldsymbol{W}_2 \widetilde{\boldsymbol{W}}_1 \boldsymbol{X} = \widetilde{\boldsymbol{W}}_2 \widetilde{\boldsymbol{W}}_1 \boldsymbol{X} \iff (\boldsymbol{W}_2 - \widetilde{\boldsymbol{W}}_2) \widetilde{\boldsymbol{W}}_1 \boldsymbol{X} = \boldsymbol{0} \tag{30}$$

The rest of the proof follows directly from here. □

## 2.4 Proof of Theorem 2 from Paper

Similar to the proof for theorem 1 of the paper, we restate the theorem itself precisely first, and then present its proof and the relevant lemmas.

Again, recall that the base loss is MSE.

**Theorem 2.6** (Theorem 2 from paper, detailed version)**.** *Assume the following:*

1. $N \geq d_x$, *and* $\boldsymbol{X}_{\epsilon}$ *is full rank;*
2. $L \geq 2$ *(a general deep linear network);*
3. $p \coloneqq \min_{i \in \{0,\ldots,L\}} d_i \geq$ *rank*$(\boldsymbol{Y}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T(\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)^{-1})$
4. $\widetilde{\boldsymbol{W}}_{\boldsymbol{L}} \boldsymbol{X} = \boldsymbol{Y}$

*Then the global minimizers*

$$\boldsymbol{W}_{\boldsymbol{L}}^{base} = \boldsymbol{W}_{\boldsymbol{L}}^{st} = \boldsymbol{Y}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T(\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)^{-1} \tag{31}$$

*Remark.* Let us interpret the assumptions.

- Assumption 3. ensures that $\boldsymbol{W}_{\mathbf{L}}^{\text{base}} = \boldsymbol{Y}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T(\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)^{-1}$ can be true.

  Intuitively speaking, this assumption is requiring the student to be sufficiently "complex" for the task that it needs to solve.

- Assumption 4. enforces that the teacher perfectly interpolates the clean input-output training pairs.

  This assumption can be satisfied by enforcing, for instance, that $\text{rank}(\boldsymbol{X}) \leq d_x$, and the maximum possible rank of $\widetilde{\boldsymbol{W}}_{\mathbf{L}}$ is no less than rank $\left(\boldsymbol{Y}\left(\overline{\boldsymbol{X}}^T\overline{\boldsymbol{X}}\right)^{-1}\overline{\boldsymbol{X}}^T\right)$, where $\overline{\boldsymbol{X}}$ is constructed by removing every linearly dependent column of $\boldsymbol{X}$.

  Intuitively speaking, we are requiring that the task which the teacher needs to solve is sufficiently "simple".

- If a slightly stronger condition was added, the argument in our proof can in fact handle the situation that the teacher and the student networks have different architectures, and the only requirements on the teacher are that, its hidden feature's dimension matches $d_{i^*}$, and the teacher can perfectly interpolate the clean training samples.

*Proof.* This proof is divided into two parts. We study the MSE and student-teacher solutions respectively.

1. We first study the global minimizers of the MSE loss. First of all, the following is true:

$$\begin{aligned}&\{\boldsymbol{W_L}|(\boldsymbol{W}_L,...,\boldsymbol{W}_1)\text{ minimizes (2)})\}\\=&\left\{\boldsymbol{W}\middle|\boldsymbol{W}=\underset{\text{rank}(\boldsymbol{W})\le p}{\text{argmin}}\|\boldsymbol{W}\boldsymbol{X_\epsilon}-\boldsymbol{Y}\|_F^2\right\}\end{aligned}\tag{32}$$

To see the "$\subseteq$" direction, take any $(\boldsymbol{W}_L,...,\boldsymbol{W}_1)$ that minimizes the MSE loss (2), clearly $\text{rank}(\boldsymbol{W_L})\le p$. Furthermore, $\boldsymbol{W_L}$ must minimize the single-layer-network rank-restricted MSE loss, since if it was not true, then there exists some $\boldsymbol{W}^*$ with $\text{rank}(\boldsymbol{W}^*)\le p$ such that

$$\|\boldsymbol{W}^*\boldsymbol{X_\epsilon}-\boldsymbol{Y}\|_F^2<\|\boldsymbol{W_L}\boldsymbol{X_\epsilon}-\boldsymbol{Y}\|_F^2\tag{33}$$

But clearly one can find a tuple $(\boldsymbol{W}_L^*,...,\boldsymbol{W}_1^*)$ that decomposes $\boldsymbol{W}^*$, which contradicts the minimality of $(\boldsymbol{W}_L,...,\boldsymbol{W}_1)$. The "$\supseteq$" direction can be proven in a similar way.

Therefore, it suffices to study the set of global minimizers of the rank-restricted MSE problem

$$\overline{\boldsymbol{W}}=\underset{\text{rank}(\boldsymbol{W})\le p}{\text{argmin}}\|\boldsymbol{W}\boldsymbol{X_\epsilon}-\boldsymbol{Y}\|_F^2\tag{34}$$

With our assumption that $p\ge\text{rank}(\boldsymbol{Y}\boldsymbol{X_\epsilon}^T(\boldsymbol{X_\epsilon}\boldsymbol{X_\epsilon}^T)^{-1})$, clearly $\overline{\boldsymbol{W}}$ is unique and $\overline{\boldsymbol{W}}=\boldsymbol{Y}\boldsymbol{X_\epsilon}^T(\boldsymbol{X_\epsilon}\boldsymbol{X_\epsilon}^T)^{-1}$. It follows that

$$\boldsymbol{W_L}^{\text{base}}=\boldsymbol{Y}\boldsymbol{X_\epsilon}^T(\boldsymbol{X_\epsilon}\boldsymbol{X_\epsilon}^T)^{-1}\tag{35}$$

2. We now study the student-teacher loss (3).

Note the inequality

$$\min_{(\boldsymbol{W}_L,...,\boldsymbol{W}_1)}\widehat{\mathcal{L}}_{\text{st}}(\boldsymbol{W}_L,...,\boldsymbol{W}_1)\ge\min_{(\boldsymbol{W}_L,...,\boldsymbol{W}_1)}\widehat{\mathcal{L}}_{\text{base}}(\boldsymbol{W}_L,...,\boldsymbol{W}_1)+\lambda\min_{(\boldsymbol{W}_L,...,\boldsymbol{W}_1)}\|\boldsymbol{W}_{i^*:1}\boldsymbol{X_\epsilon}-\widehat{\boldsymbol{E}}\boldsymbol{X}\|_F^2\tag{36}$$

The equality can only be achieved by solution(s) of the following form:

$$\begin{aligned}(\boldsymbol{W}_L,...,\boldsymbol{W}_1)\in&\{(\boldsymbol{W}_L,...,\boldsymbol{W}_1)|(\boldsymbol{W}_L,...,\boldsymbol{W}_1)\text{ minimizes }\widehat{\mathcal{L}}_{\text{base}}(\boldsymbol{W}_L,...,\boldsymbol{W}_1)\}\cap\\&\{(\boldsymbol{W}_L,...,\boldsymbol{W}_1)|(\boldsymbol{W}_L,...,\boldsymbol{W}_1)\text{ minimizes }\|\boldsymbol{W}_{i^*:1}\boldsymbol{X_\epsilon}-\widetilde{\boldsymbol{W}}_{i^*:1}\boldsymbol{X}\|_F^2\}\end{aligned}\tag{37}$$

Eventually we will show that this intersection is nonempty, and the solutions take on a specific form.

Let's start by examining the second set in the intersection above. For $(\boldsymbol{W}_L,...,\boldsymbol{W}_1)$ to belong to the second set, we only have one unique choice for the product of the matrices in the tuple $(\boldsymbol{W}_{i^*},...,\boldsymbol{W}_1)$:

$$\widehat{\boldsymbol{W}}_{i^*:1}\coloneqq\widetilde{\boldsymbol{W}}_{i^*:1}\boldsymbol{X}\boldsymbol{X_\epsilon}^T(\boldsymbol{X_\epsilon}\boldsymbol{X_\epsilon}^T)^{-1}\tag{38}$$

This is just the global minimizer of the loss $\|\boldsymbol{W}_{i^*:1}\boldsymbol{X_\epsilon}-\widetilde{\boldsymbol{W}}_{i^*:1}\boldsymbol{X}\|_F^2$ with rank constraint no less than $\text{rank}(\widetilde{\boldsymbol{W}}_{i^*:1}\boldsymbol{X}\boldsymbol{X_\epsilon}^T(\boldsymbol{X_\epsilon}\boldsymbol{X_\epsilon}^T)^{-1})$. $\boldsymbol{W}_{i^*:1}$ can indeed take on this value, since $\text{rank}(\widetilde{\boldsymbol{W}}_{i^*:1}\boldsymbol{X}\boldsymbol{X_\epsilon}^T(\boldsymbol{X_\epsilon}\boldsymbol{X_\epsilon}^T)^{-1})\le p$, while $p$ is the maximum rank $\boldsymbol{W}_{i^*:1}$ can take on.

We now need to minimize $\widehat{\mathcal{L}}_{\text{base}}$, assuming that $\boldsymbol{W}_{i^*:1}=\widehat{\boldsymbol{W}}_{i^*:1}$:

$$\begin{aligned}&\underset{(\boldsymbol{W}_L,...,\boldsymbol{W}_{i^*+1})}{\text{argmin}}\|\boldsymbol{W}_{L:i^*+1}\widehat{\boldsymbol{W}}_{i^*:1}\boldsymbol{X_\epsilon}-\boldsymbol{Y}\|_F^2\\=&\underset{(\boldsymbol{W}_L,...,\boldsymbol{W}_{i^*+1})}{\text{argmin}}\|\boldsymbol{W}_{L:i^*+1}\widetilde{\boldsymbol{W}}_{i^*:1}\boldsymbol{X}\boldsymbol{X_\epsilon}^T(\boldsymbol{X_\epsilon}\boldsymbol{X_\epsilon}^T)^{-1}\boldsymbol{X_\epsilon}-\boldsymbol{Y}\|_F^2\end{aligned}\tag{39}$$

We may simplify the above loss as follows:

$$\begin{aligned}&\|\boldsymbol{W}_{L:i^*+1}\widetilde{\boldsymbol{W}}_{i^*:1}\boldsymbol{X}\boldsymbol{X_\epsilon}^T(\boldsymbol{X_\epsilon}\boldsymbol{X_\epsilon}^T)^{-1}\boldsymbol{X_\epsilon}-\boldsymbol{Y}\|_F^2\\=&\|\boldsymbol{W}_{L:i^*+1}\widetilde{\boldsymbol{W}}_{i^*:1}\boldsymbol{X}\boldsymbol{X_\epsilon}^T(\boldsymbol{X_\epsilon}\boldsymbol{X_\epsilon}^T)^{-1}\boldsymbol{X_\epsilon}-\boldsymbol{Y}\boldsymbol{X_\epsilon}^T(\boldsymbol{X_\epsilon}\boldsymbol{X_\epsilon}^T)^{-1}\boldsymbol{X_\epsilon}+\boldsymbol{Y}\boldsymbol{X_\epsilon}^T(\boldsymbol{X_\epsilon}\boldsymbol{X_\epsilon}^T)^{-1}\boldsymbol{X_\epsilon}-\boldsymbol{Y}\|_F^2\\=&\|\boldsymbol{W}_{L:i^*+1}\widetilde{\boldsymbol{W}}_{i^*:1}\boldsymbol{X}\boldsymbol{X_\epsilon}^T(\boldsymbol{X_\epsilon}\boldsymbol{X_\epsilon}^T)^{-1}\boldsymbol{X_\epsilon}-\boldsymbol{Y}\boldsymbol{X_\epsilon}^T(\boldsymbol{X_\epsilon}\boldsymbol{X_\epsilon}^T)^{-1}\boldsymbol{X_\epsilon}\|_F^2+\|\boldsymbol{Y}\boldsymbol{X_\epsilon}^T(\boldsymbol{X_\epsilon}\boldsymbol{X_\epsilon}^T)^{-1}\boldsymbol{X_\epsilon}-\boldsymbol{Y}\|_F^2\end{aligned}\tag{40}$$

The last equality comes from the Pythagorean theorem, the fact that $\boldsymbol{W}_{sol} = \boldsymbol{Y}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T(\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)^{-1}$ is the solution to the MSE problem $\|\boldsymbol{W}\boldsymbol{X}_\epsilon - \boldsymbol{Y}\|_F^2$, and the following equilibrium identity

$$(\boldsymbol{W}_{sol}\boldsymbol{X}_{\boldsymbol{\epsilon}} - \boldsymbol{Y})\boldsymbol{X}_{\boldsymbol{\epsilon}}^T = \boldsymbol{0} \implies \text{row}(\boldsymbol{W}_{sol}\boldsymbol{X}_{\boldsymbol{\epsilon}} - \boldsymbol{Y}) \perp \text{row}(\boldsymbol{X}_{\boldsymbol{\epsilon}}) \tag{41}$$

Since the second term $\|\boldsymbol{Y}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T(\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)^{-1}\boldsymbol{X}_{\boldsymbol{\epsilon}} - \boldsymbol{Y}\|_F^2$ in (40) is independent of $\boldsymbol{W}_j$ for all $j$, we may discard it in the minimization problem. Therefore, we are left to solve

$$\underset{(\boldsymbol{W}_L,...,\boldsymbol{W}_{i^*+1})}{\text{argmin}} \|\boldsymbol{W}_{L:i^*+1}\widetilde{\boldsymbol{W}}_{i^*:1}\boldsymbol{X}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T(\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)^{-1}\boldsymbol{X}_{\boldsymbol{\epsilon}} - \boldsymbol{Y}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T(\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)^{-1}\boldsymbol{X}_{\boldsymbol{\epsilon}}\|_F^2 \tag{42}$$

If the set of $(\boldsymbol{W}_L, ..., \boldsymbol{W}_{i^*+1})$ that can make the above loss vanish is nonempty, then they are clearly the only set of minimizers of the loss.

But what exactly does making (42) zero mean? Since $\widetilde{\boldsymbol{W}}_{\mathbf{L}}\boldsymbol{X} = \boldsymbol{Y}$, the following rearrangement is true:

$$\begin{aligned}
&\boldsymbol{W}_{L:i^*+1}\widetilde{\boldsymbol{W}}_{i^*:1}\boldsymbol{X}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T(\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)^{-1}\boldsymbol{X}_{\boldsymbol{\epsilon}} - \boldsymbol{Y}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T(\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)^{-1}\boldsymbol{X}_{\boldsymbol{\epsilon}} \\
=&\boldsymbol{W}_{L:i^*+1}\widetilde{\boldsymbol{W}}_{i^*:1}\boldsymbol{X}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T(\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)^{-1}\boldsymbol{X}_{\boldsymbol{\epsilon}} - \widetilde{\boldsymbol{W}}_{L:i^*+1}\widetilde{\boldsymbol{W}}_{i^*:1}\boldsymbol{X}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T(\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)^{-1}\boldsymbol{X}_{\boldsymbol{\epsilon}} \\
=&(\boldsymbol{W}_{L:i^*+1} - \widetilde{\boldsymbol{W}}_{L:i^*+1})\widetilde{\boldsymbol{W}}_{i^*:1}\boldsymbol{X}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T(\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)^{-1}\boldsymbol{X}_{\boldsymbol{\epsilon}}
\end{aligned} \tag{43}$$

But notice that:

$$\begin{aligned}
&(\boldsymbol{W}_{L:i^*+1} - \widetilde{\boldsymbol{W}}_{L:i^*+1})\widetilde{\boldsymbol{W}}_{i^*:1}\boldsymbol{X}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T(\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)^{-1}\boldsymbol{X}_{\boldsymbol{\epsilon}} = \boldsymbol{0} \\
\iff &(\boldsymbol{W}_{L:i^*+1} - \widetilde{\boldsymbol{W}}_{L:i^*+1})\widetilde{\boldsymbol{W}}_{i^*:1}\boldsymbol{X}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T(\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)^{-1}\boldsymbol{x}_{test} = \boldsymbol{0}, \forall \boldsymbol{x}_{test} \in \mathbb{R}^{d_x}
\end{aligned} \tag{44}$$

To see " $\implies$ ", notice that since $\boldsymbol{X}_{\boldsymbol{\epsilon}}$ is of full column rank, for any $\boldsymbol{x}_{test} \in \mathbb{R}^{d_x}$, $\boldsymbol{x}_{test} = \boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{\alpha}$ for some $\boldsymbol{\alpha} \in \mathbb{R}^N$. So

$$\begin{aligned}
&(\boldsymbol{W}_{L:i^*+1} - \widetilde{\boldsymbol{W}}_{L:i^*+1})\widetilde{\boldsymbol{W}}_{i^*:1}\boldsymbol{X}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T(\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)^{-1}\boldsymbol{x}_{test} \\
=&(\boldsymbol{W}_{L:i^*+1} - \widetilde{\boldsymbol{W}}_{L:i^*+1})\widetilde{\boldsymbol{W}}_{i^*:1}\boldsymbol{X}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T(\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)^{-1}\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{\alpha} \\
=&\boldsymbol{0}\boldsymbol{\alpha} \\
=&\boldsymbol{0}
\end{aligned} \tag{45}$$

The " $\impliedby$ " direction is obvious.

The condition on $\boldsymbol{W}_{L:i^*+1}$ in (44) is clearly equivalent to the following:

$$(\boldsymbol{W}_{L:i^*+1} - \widetilde{\boldsymbol{W}}_{L:i^*+1})\widetilde{\boldsymbol{W}}_{i^*:1}\boldsymbol{X}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T(\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)^{-1} = \boldsymbol{0} \tag{46}$$

Therefore, driving (42) to zero is equivalent to the condition (46). Now, what is the set of $(\boldsymbol{W}_L, ..., \boldsymbol{W}_{i^*+1})$ that satisfies this condition, and more importantly, is this set even nonempty? We shall prove in the next paragraph that this set is indeed nonempty.

By assumption 2. in the theorem statement, $\text{rank}(\boldsymbol{Y}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T(\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)^{-1}) \leq p$, and since $\widetilde{\boldsymbol{W}}_{\mathbf{L}}\boldsymbol{X} = \boldsymbol{Y}$,

$$\text{rank}(\widetilde{\boldsymbol{W}}_{L:i^*+1}\boldsymbol{P}_{\widetilde{\boldsymbol{W}}_{i^*:1}\boldsymbol{X}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T(\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)^{-1}}) \leq \text{rank}(\widetilde{\boldsymbol{W}}_{L:i^*+1}\widetilde{\boldsymbol{W}}_{i^*:1}\boldsymbol{X}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T(\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)^{-1}) \leq p \tag{47}$$

must be true. The first inequality needs some justification, which we will discuss below. But assuming that it is true, we now know that there indeed exists a (set of) $\boldsymbol{W}_{L:i^*+1}$ such that (46) is true, in fact, $\boldsymbol{W}_{L:i^*+1} = \widetilde{\boldsymbol{W}}_{L:i^*+1}\boldsymbol{P}_{\widetilde{\boldsymbol{W}}_{i^*:1}\boldsymbol{X}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T(\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)^{-1}}$ is an example.

Going back to the first inequality, it holds because for any $\boldsymbol{A}, \boldsymbol{B}$ for which their product $\boldsymbol{A}\boldsymbol{B}$ makes sense, $\text{rank}(\boldsymbol{A}\boldsymbol{B}) \geq \text{rank}(\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{B}})$. To see this, consider the following situations. Case 1: $\boldsymbol{B}$ has linearly independent columns. Then $\text{rank}(\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{B}}) = \text{rank}(\boldsymbol{A}\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{B}^T) \leq \text{rank}(\boldsymbol{A}\boldsymbol{B})$. Case 2: $\boldsymbol{B}$ does not have linearly independent columns. Construct $\overline{\boldsymbol{B}}$ from $\boldsymbol{B}$ by removing the linearly dependent columns of $\boldsymbol{B}$. Notice that $\text{rank}(\boldsymbol{A}\boldsymbol{B}) = \text{rank}(\boldsymbol{A}\overline{\boldsymbol{B}})$. But

$$\text{rank}(\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{B}}) = \text{rank}(\boldsymbol{A}\boldsymbol{P}_{\overline{\boldsymbol{B}}}) = \text{rank}(\boldsymbol{A}\overline{\boldsymbol{B}}(\overline{\boldsymbol{B}}^T\overline{\boldsymbol{B}})^{-1}\overline{\boldsymbol{B}}^T) \leq \text{rank}(\boldsymbol{A}\overline{\boldsymbol{B}}) = \text{rank}(\boldsymbol{A}\boldsymbol{B}) \tag{48}$$

We now arrive at the fact that there does exist (a set of) $(\boldsymbol{W}_L, ..., \boldsymbol{W}_{i^*+1})$ that satisfies (46), therefore, they form the set of minimizers of (42).

But clearly the identity (46) which characterizes this set of minimizers is equivalent to

$$\boldsymbol{W}_{L:i^*+1}\widetilde{\boldsymbol{W}}_{i^*:1}\boldsymbol{X}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T(\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)^{-1} = \widetilde{\boldsymbol{W}}_{L:i^*+1}\widetilde{\boldsymbol{W}}_{i^*:1}\boldsymbol{X}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T(\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)^{-1} \tag{49}$$

and because $\widehat{\boldsymbol{W}}_{i^*:1} = \widetilde{\boldsymbol{W}}_{i^*:1}\boldsymbol{X}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T(\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)^{-1}$ and $\widetilde{\boldsymbol{W}}_{\mathbf{L}}\boldsymbol{X} = \boldsymbol{Y}$, the above equality is equivalent to

$$\boldsymbol{W}_{L:i^*+1}\widehat{\boldsymbol{W}}_{i^*:1} = \boldsymbol{Y}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T(\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)^{-1} \tag{50}$$

We have now arrived at the point to say that, the following set from (37) is indeed nonempty

$$\begin{aligned}&\{(\boldsymbol{W}_L, ..., \boldsymbol{W}_1) | (\boldsymbol{W}_L, ..., \boldsymbol{W}_1) \text{ minimizes } \widehat{\mathcal{L}}_{\text{base}}(\boldsymbol{W}_L, ..., \boldsymbol{W}_1)\} \cap \\ &\{(\boldsymbol{W}_L, ..., \boldsymbol{W}_1) | (\boldsymbol{W}_L, ..., \boldsymbol{W}_1) \text{ minimizes } \|\boldsymbol{W}_{i^*:1}\boldsymbol{X}_{\boldsymbol{\epsilon}} - \widetilde{\boldsymbol{W}}_{i^*:1}\boldsymbol{X}\|_F^2\}\end{aligned} \tag{51}$$

and any $(\boldsymbol{W}_L, ..., \boldsymbol{W}_1)$ belonging to this intersection must satisfy the property

$$\boldsymbol{W}_{i^*:1} = \widehat{\boldsymbol{W}}_{i^*:1}, \text{ and } \boldsymbol{W}_{L:i^*+1}\widehat{\boldsymbol{W}}_{i^*:1} = \boldsymbol{Y}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T(\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)^{-1} \tag{52}$$

Finally, we can conclude that, due to the nonemptiness of the intersection of the two sets from (37), the equality in (36) is indeed achievable, and every solution $(\boldsymbol{W}_L^{\text{st}}, ..., \boldsymbol{W}_1^{\text{st}})$ achieving the equality satisfies

$$\boldsymbol{W}_{\mathbf{L}}^{\text{st}} = \boldsymbol{Y}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T(\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)^{-1} = \boldsymbol{W}_{\mathbf{L}}^{\text{base}} \tag{53}$$

The proof is complete.

□

**Corollary 2.6.1.** *If $N \geq d_x$, $\widetilde{\boldsymbol{W}}_{\boldsymbol{L}}\boldsymbol{X} = \boldsymbol{Y}$, and $p = \min(d_x, d_y)$ (wide networks), then the global minimizers of MSE and student-teacher are identical.*

*Proof.* The inequality $\text{rank}(\boldsymbol{Y}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T(\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)^{-1/2}) \leq \min(d_y, d_x) = p$ must be true, so the application of the above theorem is legal. □

## 2.5 Nonlinear-Teacher-Network Results

**Theorem 2.7** (Nonlinear teacher, $N < d_x$)**.** *Denote $\boldsymbol{W}_i^{base}(t)$ and $\boldsymbol{W}_i^{st}(t)$ as the weights for the student network trained with the the base loss* (2)*, and the student network trained with the student-teacher loss* (3)*, respectively.*

*Let the following assumptions hold:*

1. *Gradient flow is the optimizer;*
2. *$N_s < d_x$;*
3. *$L = 2$;*
4. *$\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N_s}$ and $\{\boldsymbol{\epsilon}_i\}_{i=1}^{N_s}$ are all sampled independently, and $\boldsymbol{x}$ and $\boldsymbol{\epsilon}$ are continuous random vectors;*
5. *There exists some $\delta > 0$ such that $\|\boldsymbol{W}_i^{base}(0)\|_F \leq \delta$ and $\|\boldsymbol{W}_i^{st}(0)\|_F \leq \delta$ for all $i$;*
6. *The teacher network takes the form $\widetilde{\boldsymbol{W}}_2\sigma(\widetilde{\boldsymbol{W}}_1\boldsymbol{x})$, with $\sigma(\cdot)$ being a (nonlinear) entry-wise activation function. Furthermore, assume that $\widetilde{\boldsymbol{W}}_2\sigma(\widetilde{\boldsymbol{W}}_1\boldsymbol{X}) = \boldsymbol{Y}$, i.e. the teacher network can perfectly solve the clean training problem.*
7. *The $\boldsymbol{W}_i^{base}(0)$'s are initialized with the balanced initialization;*
8. *Gradient flow successfully converges to a global minimizer for both the MSE- and ST-trained networks;*

9. *The weights $\boldsymbol{W}_i^{st}(t)$ remain in a compact set for $t \in [0,\infty)$. In particular, denote $\|\boldsymbol{W}_i^{st}(t)\|_F \le M, t \in [0,\infty)$.*

*Then the following is true almost surely:*

$$\lim_{t\to\infty} \|\boldsymbol{W}_{\boldsymbol{L}}^{base}(t) - \boldsymbol{W}_{\boldsymbol{L}}^{st}(t)\|_F \le C\delta \tag{54}$$

*for some $C$ that is bounded as $\delta$ tends to 0.*

*Proof.* Note that the only difference in assumption between this theorem and the linear-teacher-network theorem is that, we assume the teacher network has nonlinear activation now, and $\widetilde{\boldsymbol{W}}_2\sigma(\widetilde{\boldsymbol{W}}_1\boldsymbol{X}) = \boldsymbol{Y}$. Consider the following two points.

- Notice that even though the activation function of the teacher is now nonlinear, $\widetilde{\boldsymbol{W}}_2\sigma(\widetilde{\boldsymbol{W}}_1\boldsymbol{X})$ still is the product of two matrices, $\widetilde{\boldsymbol{W}}_2 \in \mathbb{R}^{d_y\times d_1}$ and $\sigma(\widetilde{\boldsymbol{W}}_1\boldsymbol{X}) \in \mathbb{R}^{d_1\times d_x}$, therefore, $\boldsymbol{Y}$ has at most rank $p$. It also follows that $\boldsymbol{U}_p\boldsymbol{U}_p^T\boldsymbol{Y} = \boldsymbol{Y}$. Noting that $\boldsymbol{X}_\epsilon$ is full-rank almost surely, it is indeed possible to find $(\boldsymbol{W}_2, \boldsymbol{W}_1)$ such that $\boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X}_\epsilon = \boldsymbol{Y}$. In fact, the base-loss solution set is now

$$\{\boldsymbol{W}_2, \boldsymbol{W}_1 |\ \boldsymbol{W}_2\boldsymbol{W}_1 = \boldsymbol{Y}(\boldsymbol{X}_{\boldsymbol{\epsilon}}^T\boldsymbol{X}_{\boldsymbol{\epsilon}})^{-1}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T + \boldsymbol{R},\ \mathrm{row}(\boldsymbol{R}) \subseteq \mathrm{col}(\boldsymbol{X}_{\boldsymbol{\epsilon}})^\perp\} \tag{55}$$

  Therefore, using exactly the same argument as in the proof for theorem 1 of the paper, we can show that, $\boldsymbol{W}_2^{\mathrm{base}}(t)\boldsymbol{W}_1^{\mathrm{base}}(t)$ tends to $\boldsymbol{Y}(\boldsymbol{X}_\epsilon^T\boldsymbol{X}_\epsilon)^{-1}\boldsymbol{X}_\epsilon^T + \boldsymbol{W}^{\mathrm{base}}(\delta)$ as $t\to\infty$, with $\|\boldsymbol{W}^{\mathrm{base}}(\delta)\|_F \in \mathcal{O}(p^{1/4}\|\boldsymbol{Y}(\boldsymbol{X}_{\boldsymbol{\epsilon}}^T\boldsymbol{X}_{\boldsymbol{\epsilon}})^{-1}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T\|_F^{1/2})$ when $\delta$ is sufficiently small.

- For $\boldsymbol{W}_2^{\mathrm{st}}(t)\boldsymbol{W}_1^{\mathrm{st}}(t)$, we note that the global minimizers of the student-teacher loss is that set

$$\{\boldsymbol{W}_2 \in \mathbb{R}^{d_y\times d_1}, \boldsymbol{W}_1 \in \mathbb{R}^{d_1\times d_x} |\ \boldsymbol{W}_2\boldsymbol{W}_1 = \boldsymbol{Y}(\boldsymbol{X}_{\boldsymbol{\epsilon}}^T\boldsymbol{X}_{\boldsymbol{\epsilon}})^{-1}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T + (\widetilde{\boldsymbol{W}}_2 + \boldsymbol{R}_2)\boldsymbol{R}_1,$$
$$\boldsymbol{R}_1 \in \mathbb{R}^{d_1\times d_x} \wedge \mathrm{row}(\boldsymbol{R}_1) \subseteq \mathrm{col}(\boldsymbol{X}_{\boldsymbol{\epsilon}})^\perp \wedge \boldsymbol{R}_2 \in \mathbb{R}^{d_x\times d_1} \wedge \mathrm{row}(\boldsymbol{R}_2) \subseteq \mathrm{col}(\sigma(\widetilde{\boldsymbol{W}}_1\boldsymbol{X}))^\perp\} \tag{56}$$

  The "residue matrix" $(\widetilde{\boldsymbol{W}}_2 + \boldsymbol{R}_2)\boldsymbol{R}_1$ still satisfies the property that $\mathrm{row}((\widetilde{\boldsymbol{W}}_2 + \boldsymbol{R}_2)\boldsymbol{R}_1) \subseteq \mathrm{col}(\boldsymbol{X}_\epsilon^\perp)$. Therefore, the gradient-flow argument for $\boldsymbol{W}_2^{\mathrm{st}}(t)\boldsymbol{W}_1^{\mathrm{st}}(t)$ still holds, so $\boldsymbol{W}_2^{\mathrm{st}}(t)\boldsymbol{W}_1^{\mathrm{st}}(t)$ must also tend to $\boldsymbol{Y}(\boldsymbol{X}_\epsilon^T\boldsymbol{X}_\epsilon)^{-1}\boldsymbol{X}_\epsilon^T + \boldsymbol{W}^{\mathrm{st}}(\delta)$ as $t\to\infty$, with $\|\boldsymbol{W}^{\mathrm{st}}(\delta)\|_F \le M\delta$.

Combining the above two results finishes our proof. □

**Theorem 2.8** (Nonlinear teacher, $N \ge d_x$)**.** *Assume the following:*

1. *$N \ge d_x$, and $\boldsymbol{X}_\epsilon$ is full rank;*
2. *$L \ge 2$ (a general deep linear network);*
3. *$p \coloneqq \min_{i\in\{0,\dots,L\}} d_i \ge$ rank$(\boldsymbol{Y}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T(\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)^{-1})$*
4. *Assume that the teacher takes the form $\widetilde{\boldsymbol{W}}\boldsymbol{F}(\boldsymbol{x})$, with the output dimension of $\boldsymbol{F}(\cdot)$ equal to $d_{i^*}$. Also denote $\boldsymbol{F}(\boldsymbol{X}) \in \mathbb{R}^{d_{i^*}\times N_s}$ as the features the teacher provides to the student, i.e. the student-teacher loss has the form*

$$\underset{\boldsymbol{W}_L,\dots,\boldsymbol{W}_1}{argmin} \left(\|\boldsymbol{W}_{\boldsymbol{L}}\boldsymbol{X}_{\boldsymbol{\epsilon}} - \boldsymbol{Y}\|_F^2 + \lambda\|\boldsymbol{W}_{i^*:1}\boldsymbol{X}_{\boldsymbol{\epsilon}} - \boldsymbol{F}(\boldsymbol{X})\|_F^2\right) \tag{57}$$

   *Furthermore, assume that the teacher satisfies $\widetilde{\boldsymbol{W}}\boldsymbol{F}(\boldsymbol{X}) = \boldsymbol{Y}$.*
5. *$\min_{i\in\{0,\dots,i^*\}} d_i \ge$ rank$(\boldsymbol{F}(\boldsymbol{X})\boldsymbol{X}_{\boldsymbol{\epsilon}}^T(\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)^{-1})$.*

*Then the global minimizers*

$$\boldsymbol{W}_{\boldsymbol{L}}^{base} = \boldsymbol{W}_{\boldsymbol{L}}^{st} = \boldsymbol{Y}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T(\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)^{-1} \tag{58}$$

*Proof.* We prove the theorem in two main steps.

1. Since $p \coloneqq \min_{i\in\{0,...,L\}} d_i \geq \text{rank}(\boldsymbol{Y}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T(\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)^{-1})$ is still true, the solution for the base loss does not change from before:
$$\boldsymbol{W}_{\mathbf{L}}^{\text{base}} = \boldsymbol{Y}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T(\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)^{-1} \tag{59}$$

2. For the student-teacher loss, we argue in almost the same way as the linear-teacher case. We still prove that
$$\begin{aligned}\varnothing \neq &\{(\boldsymbol{W}_L,...,\boldsymbol{W}_1)|(\boldsymbol{W}_L,...,\boldsymbol{W}_1) \text{ minimizes } \widehat{\mathcal{L}}_{\text{base}}(\boldsymbol{W}_L,...,\boldsymbol{W}_1)\} \cap \\ &\{(\boldsymbol{W}_L,...,\boldsymbol{W}_1)|(\boldsymbol{W}_L,...,\boldsymbol{W}_1) \text{ minimizes } \|\boldsymbol{W}_{i^*:1}\boldsymbol{X}_{\boldsymbol{\epsilon}} - \boldsymbol{F}(\boldsymbol{X})\|_F^2\}\end{aligned} \tag{60}$$
Like before, we focus on the second set first. To minimize $\|\boldsymbol{W}_{i^*:1}\boldsymbol{X}_{\boldsymbol{\epsilon}} - \boldsymbol{F}(\boldsymbol{X})\|_F^2$, due to assumption 5., only one solution exists:
$$\boldsymbol{W}_{i^*:1} = \widehat{\boldsymbol{W}}_{i^*:1} \coloneqq \boldsymbol{F}(\boldsymbol{X})\boldsymbol{X}_{\boldsymbol{\epsilon}}^T(\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)^{-1} \tag{61}$$
Now, to obtain the solutions in the intersection of the two sets, we assume $\boldsymbol{W}_{i^*:1} = \widehat{\boldsymbol{W}}_{i^*:1}$ and check what value $\boldsymbol{W}_{L:i^*+1}$ can take on. One particular choice is simply $\boldsymbol{W}_{L:i^*+1} = \widetilde{\boldsymbol{W}}$, in which case we obtain $\boldsymbol{W}_{\mathbf{L}} = \boldsymbol{Y}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T(\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)^{-1}$, which indeed minimizes $\widehat{\mathcal{L}}_{\text{base}}$. It follows that the above intersection is nonempty, so arguing similarly to the linear-teacher case, we may conclude that
$$\boldsymbol{W}_{\mathbf{L}}^{\text{base}} = \boldsymbol{W}_{\mathbf{L}}^{\text{st}} = \boldsymbol{Y}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T(\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{X}_{\boldsymbol{\epsilon}}^T)^{-1} \tag{62}$$

□

# 3 Proofs for Theorem 3 in Section 5 of the Paper

In this section, we shall present the proof for theorem 3 of the paper.

## 3.1 Notations, Conventions and Assumptions

Most of the notations and conventions we use are the same as the ones we use for the previous section. We only emphasize the differences here.

Denote $\boldsymbol{X} \in \mathbb{R}^{N \times d_x}$ as the clean design matrix, defined by $[\boldsymbol{X}]_{i,:} = \boldsymbol{x}_i^T$. Similarly, $\boldsymbol{Z} \in \mathbb{R}^{N \times d_x}$ is the noise matrix, defined by $[\boldsymbol{Z}]_{i,:} = \boldsymbol{\epsilon}_i^T$. $\boldsymbol{X}_\epsilon = \boldsymbol{X} + \boldsymbol{Z}$ is the noisy training input matrix. The target vector is $\boldsymbol{y} \in \mathbb{R}^{d_y}$. Recall that the individual target samples are one-dimensional as we are focusing on linear regression in this section.

Given some index set $S \subseteq \{1, ..., n\}$ and vector $\boldsymbol{\beta} \in \mathbb{R}^n$, we use $\boldsymbol{\beta}_S \in \mathbb{R}^{|S|}$ to denote the sub-vector created by extracting the entries in $\boldsymbol{\beta}$ with indices contained in $S$, e.g. given $\boldsymbol{\beta} = (1, 5, 2, 10)$ and $S = \{2, 4\}$, then $\boldsymbol{\beta}_S = (5, 10)$. Similarly, given a matrix $\boldsymbol{M} \in \mathbb{R}^{m \times n}$, we denote $\boldsymbol{M}_S \in \mathbb{R}^{m \times |S|}$ to be the sub-matrix of $\boldsymbol{M}$, created by extracting the columns in $\boldsymbol{M}$ with indices contained in $S$.

We restate the basic assumptions made in the paper, in addition to a few that we specify on the input data:

1. The learning problem is linear regression. The ground truth is a linear model $\boldsymbol{\beta}^* \in \mathbb{R}^d$, with sparsity level $s$, i.e. only $s$ entries in it are nonzero.
2. The student and teacher networks are both shallow networks.
3. We set $m = s$, i.e. the hidden dimension of the networks (i.e. the output dimension of $\boldsymbol{W}_1$) is equal to $s$.
4. We use $\ell_1$ regularization during training.
5. The student's architecture is $\widetilde{\boldsymbol{W}}_2 \boldsymbol{P} \boldsymbol{W}_1(\boldsymbol{x}+\boldsymbol{\epsilon})$, and teacher's architecture is $\widetilde{\boldsymbol{W}}_2 \boldsymbol{P} \widetilde{\boldsymbol{W}}_1 \boldsymbol{x}$. $\widetilde{\boldsymbol{W}}_2 \in \mathbb{R}^{1 \times m/g}$, and $\boldsymbol{W}_1, \widetilde{\boldsymbol{W}}_1 \in \mathbb{R}^{m \times d_x}$. Moreover, $\boldsymbol{P} \in \mathbb{R}^{(m/g) \times m}$, $g \in \mathbb{N}$ is a divisor of $m$, and $\boldsymbol{P}_{i,j} = 1$ if $j \in \{ig, ..., (i+1)g\}$, and zero everywhere else. Multiplication with $\boldsymbol{P}$ essentially *sums every $g$ neurons' output*, similar to how average pooling works in convolutional neural networks.
   - In Theorem 3, the weights of the teacher satisfy $[\widetilde{\boldsymbol{W}}_2]_i = 1$ for all $i = 1, ..., s/g$; $[\widetilde{\boldsymbol{W}}_1]_{i,i} = \beta_i^*$ for $i = 1, ..., s$, and the remaining entries are all zeros.
   - Figure 1 illustrates $\boldsymbol{P}\widetilde{\boldsymbol{W}}_1 \boldsymbol{x}$, for a simple case of $d_x = 4$, $g = 2$, and the ground truth is $\boldsymbol{\beta}^* = (\beta_1^*, ..., \beta_4^*)$. This figure visualizes how the teacher's hidden features are pooled, and presented to the student.
6. Throughout this whole section, we shall assume that $\boldsymbol{x}$ comes from a distribution whose covariance matrix is the identity. The noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma_\epsilon^2 \boldsymbol{I}_{d_x \times d_x})$, and $\sigma_\epsilon < 1$.

## 3.2 Simplifying the Problem

We now introduce an equivalent, but more succinct, formulation of the student-teacher learning procedure in Section 5.2, so that we can present the proofs more easily. As we will explain soon, the student-teacher learning setting in section 5.2 of the paper can be decomposed into $s/g$ subproblems of the following form: for $i \in \{1, ..., s/g\}$,

$$\boldsymbol{w}^i = \underset{\boldsymbol{w}}{\operatorname{argmin}} \|\boldsymbol{X}_\epsilon \boldsymbol{w} - \boldsymbol{X}\boldsymbol{\beta}^{*i}\|_2^2 / N_s + \lambda_i \|\boldsymbol{w}\|_1 \tag{63}$$

where $\boldsymbol{\beta}_j^{*i} = \beta_j^*$ for $j \in \{ig, ..., (i+1)g\}$ and zero everywhere else (it has a sparsity level of $g$). **We denote the support set $\operatorname{supp}(\boldsymbol{\beta}^{*i})$ to be** $S_i$ (i.e. it is the set of indices on which $\boldsymbol{\beta}^{*i}$ is nonzero).

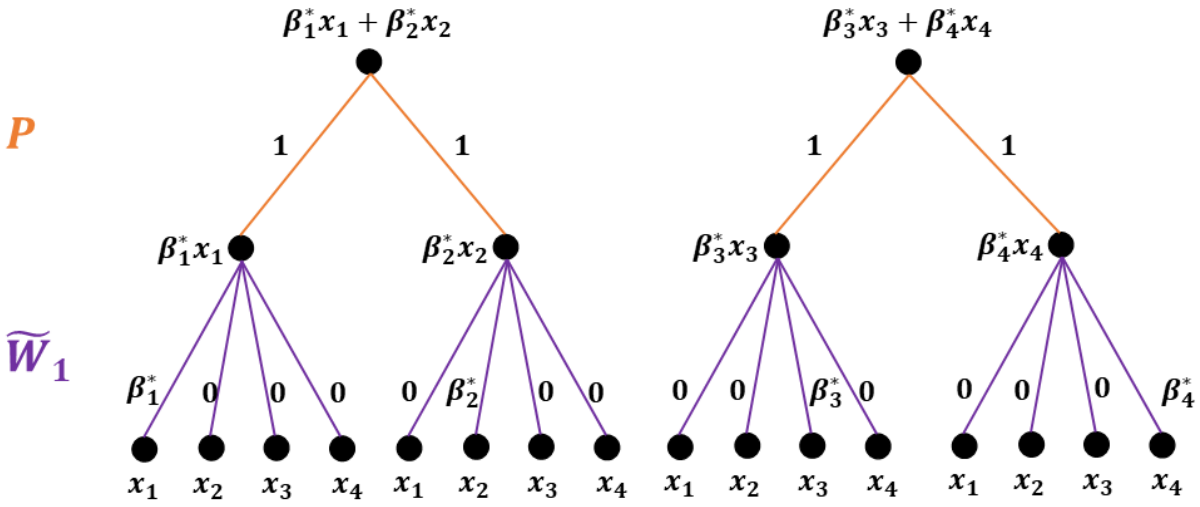


Figure 1: An example of the operation $\boldsymbol{P}\widetilde{\boldsymbol{W}}_1\boldsymbol{x}$. In this example, $d_x = 4$, $g = 2$. Note that each hidden neuron of the teacher $[\widetilde{\boldsymbol{W}}_1]_{i,:}$ only encodes one entry from $\boldsymbol{\beta}^*$. The hidden features $[\widetilde{\boldsymbol{W}}_1]_{i,:}^T\boldsymbol{x}$ are then pooled by the matrix $\boldsymbol{P}$. Therefore, in the feature difference loss $\|\boldsymbol{P}\boldsymbol{W}_1(\boldsymbol{x}+\boldsymbol{\epsilon}) - \boldsymbol{P}\widetilde{\boldsymbol{W}}_1\boldsymbol{x}\|_2^2$, the first group of student neurons sees $\beta_1^*x_1 + \beta_2^*x_2$, while the second group sees $\beta_3^*x_3 + \beta_4^*x_4$. Consequently, each group of the student's neurons sees the action of a 2-sparse subset of $\boldsymbol{\beta}^*$ on the clean input signal $\boldsymbol{x}$.

To see why the above simplified training problem is equivalent to the paper's one, recall that the problem stated in the paper is the following (the feature difference loss itself, without the $\ell_1$ regularization)

$$\frac{1}{N_s}\sum_{i=1}^{N_s}\left\|\boldsymbol{P}\left[\boldsymbol{W}_1(\boldsymbol{x}_i+\boldsymbol{\epsilon}_i) - \widetilde{\boldsymbol{W}}_1\boldsymbol{x}_i\right]\right\|_2^2 \tag{64}$$

But since $\boldsymbol{P}_{i,j} = 1$ if $j \in \{ig, ..., (i+1)g\}$, and zero everywhere else, the above loss can be written as a collection of losses independent from each other (enumerated by $i \in \{1, ..., s/g\}$):

$$\frac{1}{N_s}\left\|\boldsymbol{X}_\epsilon\left(\sum_{j=ig}^{(i+1)g}[\boldsymbol{W}_1]_{j,:}^T\right) - \boldsymbol{X}\left(\sum_{j=ig}^{(i+1)g}[\widetilde{\boldsymbol{W}}_1]_{j,:}^T\right)\right\|_2^2 \tag{65}$$

For the term $\sum_{j=ig}^{(i+1)g}[\widetilde{\boldsymbol{W}}_1]_{j,:}^T$, since in the theorem we assume that $[\widetilde{\boldsymbol{W}}_1]_{i,i} = \beta_i^*$ for $i = 1, ..., s$, and the remaining entries are all zeros, the vector $\sum_{j=ig}^{(i+1)g}[\widetilde{\boldsymbol{W}}_1]_{j,:}^T$'s $ig$-th to $(i+1)g$-th entries are equal to those of $\boldsymbol{\beta}^*$ at the same indices, and zero everywhere else. This is where the $\boldsymbol{\beta}^{*i}$ came from.

Let's now add in the $\ell_1$ regularization. For every $i \in \{1, ..., s/g\}$, we have the loss

$$\frac{1}{N_s}\left\|\boldsymbol{X}_\epsilon\left(\sum_{j=ig}^{(i+1)g}[\boldsymbol{W}_1]_{j,:}^T\right) - \boldsymbol{X}\boldsymbol{\beta}^{*i}\right\|_2^2 + \lambda_i\left\|\sum_{j=ig}^{(i+1)g}[\boldsymbol{W}_1]_{j,:}^T\right\|_1 \tag{66}$$

Note that we regularize every group of hidden neurons in the student network. One can verify that the minimizer(s) $\sum_{j=ig}^{(i+1)g}[\boldsymbol{W}_1]_{j,:}^T$ of the above loss is the same as the minimizer(s) $\boldsymbol{w}^i$ of the following loss:

$$\|\boldsymbol{X}_\epsilon\boldsymbol{w}^i - \boldsymbol{X}\boldsymbol{\beta}^{*i}\|_2^2/N_s + \lambda_i\|\boldsymbol{w}^i\|_1 \tag{67}$$

By noting that $\widetilde{\boldsymbol{W}}_2$'s entries are all 1's, the simplification of the testing loss $\mathbb{E}\left[\left(\widetilde{\boldsymbol{W}}_2\boldsymbol{P}\boldsymbol{W}_1(\boldsymbol{x}+\boldsymbol{\epsilon})-\boldsymbol{\beta}^{*T}\boldsymbol{x}\right)^2\right]$ can be argued in a similar way as above, and it simplifies to

$$\mathbb{E}\left[\left(\sum_{i=1}^{s/g}\boldsymbol{w}^{iT}(\boldsymbol{x}+\boldsymbol{\epsilon})-\boldsymbol{\beta}^{*T}\boldsymbol{x}\right)^2\right]. \tag{68}$$

### 3.3 Optimal Test Error

Before going into the theorem and its proof, let us try to understand what the optimal testing error of this regression problem is.

Since we assumed that $\boldsymbol{x}$ comes from a distribution with identity covariance, and $\boldsymbol{\epsilon}\sim\mathcal{N}(\boldsymbol{0},\sigma_\epsilon^2\boldsymbol{I}_{d_x\times d_x})$, the following is true:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x},\boldsymbol{\epsilon}}[(\boldsymbol{\beta}^T(\boldsymbol{x}+\boldsymbol{\epsilon})-\boldsymbol{\beta}^{*T}\boldsymbol{x})^2] &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{\epsilon}}[((\boldsymbol{\beta}-\boldsymbol{\beta}^*)^T\boldsymbol{x}+\boldsymbol{\beta}^T\boldsymbol{\epsilon})^2]\\ &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{\epsilon}}[((\boldsymbol{\beta}-\boldsymbol{\beta}^*)^T\boldsymbol{x})^2]+2\mathbb{E}_{\boldsymbol{x},\boldsymbol{\epsilon}}[(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^T\boldsymbol{x})(\boldsymbol{\beta}^T\boldsymbol{\epsilon})]+\mathbb{E}_{\boldsymbol{x},\boldsymbol{\epsilon}}[(\boldsymbol{\beta}^T\boldsymbol{\epsilon})^2]\\ &= \|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2^2+\sigma_\epsilon^2\|\boldsymbol{\beta}\|_2^2.\end{aligned} \tag{69}$$

It is then easy to show that the optimal linear model that minimizes the above testing error is as follows:

$$\boldsymbol{\beta}^*_{noise}=\frac{1}{1+\sigma_\epsilon^2}\boldsymbol{\beta}^* \tag{70}$$

Furthermore, the optimal testing error is:

$$\begin{aligned}\|\boldsymbol{\beta}^*_{noise}-\boldsymbol{\beta}^*\|_2^2+\sigma_\epsilon^2\|\boldsymbol{\beta}^*_{noise}\|_2^2 &= \frac{\sigma_\epsilon^4}{(1+\sigma_\epsilon^2)^2}\|\boldsymbol{\beta}^*\|_2^2+\frac{\sigma_\epsilon^2}{(1+\sigma_\epsilon^2)^2}\|\boldsymbol{\beta}^*\|_2^2\\ &= \frac{\sigma_\epsilon^2\|\boldsymbol{\beta}^*\|_2^2}{1+\sigma_\epsilon^2}\end{aligned} \tag{71}$$

### 3.4 Theorem 3 and Its Proof

**Theorem 3.1** (Theorem 3 from paper, detailed version)**.** *Let the following assumptions hold:*

1. *Assumptions in subsection 3.1 hold;*
2. *The number of samples satisfies $N_s\in\Omega\left(g^2\log(d_x)\right)$.*
3. *$\boldsymbol{X}$ is fixed, randomness only comes from the noise $\boldsymbol{Z}$.*
4. *The columns of $\boldsymbol{X}$ satisfy $N_s^{-1}\|\boldsymbol{X}_i\|_2^2\le K_x$ for all $i$, with $K_x\in\mathcal{O}(1)$.*
5. *Let $\boldsymbol{X}$ satisfy the property that, with high probability (over the randomness of $\boldsymbol{Z}$), $\boldsymbol{X}_\epsilon$ has the mutual incoherence condition for some $\gamma\in(0,1)$: for every $i\in\{1,...,s/g\}$, for all $j\notin S_i$,*

$$\max_{j\notin S_i}\|\boldsymbol{X}_{\epsilon,j}^T\boldsymbol{X}_{\epsilon,S_i}(\boldsymbol{X}_{\epsilon,S_i}^T\boldsymbol{X}_{\epsilon,S_i})^{-1}\|_1\le(1-\gamma) \tag{72}$$

6. *With high probability, for every $S_i$, $\boldsymbol{X}_{\epsilon,S_i}^T\boldsymbol{X}_{\epsilon,S_i}/N_s$ is invertible, and denote its minimum eigenvalue as $\Lambda_{min}^i$. Furthermore, define $\Lambda_{min}=\min_{i\in\{1,...,s/g\}}\Lambda_{min}^i$.*

*Then there exists a choice of $\lambda_i$ for each problem $i$ in* (63)*, such that with high probability, the overall test error satisfies*

$$\mathbb{E}\left[\left(\widetilde{\boldsymbol{W}}_2\boldsymbol{P}\boldsymbol{W}_1(\boldsymbol{x}+\boldsymbol{\epsilon})-\boldsymbol{\beta}^{*T}\boldsymbol{x}\right)^2\right]\in\mathcal{O}\left(\frac{1}{\gamma^2\Lambda_{min}^2}\frac{\sigma_\epsilon^2\|\boldsymbol{\beta}^*\|_2^2}{1+\sigma_\epsilon^2}\right) \tag{73}$$

*Remark.* Let us interpret the result and assumptions.

1. Condition 4. can be easily satisfied by many types of random matrices, e.g. it would be satisfied with high probability if $\boldsymbol{X}$'s entries are sampled independently from the standard Gaussian distribution.
2. The invertibility condition is almost trivially true, since if we fix $\boldsymbol{X}$ and only allow randomness in $\boldsymbol{Z}$, then the columns $\boldsymbol{X}_{\boldsymbol{\epsilon}}$ are continuous random vectors, and must be independent from each other almost surely. Therefore, $\boldsymbol{X}_{\epsilon,S_i}$ will be full-rank almost surely.
3. The mutual incoherence condition is a common assumption used in the LASSO literature to ensure basis recovery (ours is modified from the standard one, since unlike the traditional case, we have noise in the input). The types of matrices that satisfy mutual incoherence is discussed in [3] (section 2 and 4) and [5] (see proposition 24). For instance, if $\boldsymbol{X}$'s entries were sampled independently from the standard Gaussian, then with $N_s \in \Omega(g^2 \log(d_x))$, $\boldsymbol{X}_{\boldsymbol{\epsilon}}$ must satisfy mutual incoherence with high probability in high dimensions (over the randomness of $\boldsymbol{X}$ and $\boldsymbol{Z}$). Note that there are some subtleties with general iid random matrices that have finite exponential moments, as $\log(d_x) \leq o(N_s^c)$ for some $c > 0$ could be needed. The general treatment on this condition is beyond the scope of our work.
4. Note that the sample complexity $N_s \in \Omega(g^2 \log(d_x))$ indicates the "bare minimum" to ensure reasonable performance of student-teacher learning. As mentioned in the previous point, *more samples are always better*, e.g. if we instead pick $g^2 \log(d_x) \leq o(N_s^c)$ for some small $c > 0$, then we could get better testing error in the end.

*Proof.* The proof of this theorem follows directly from lemma 3.3. During testing, by equations (67) and (69), we just need to compute

$$\left\| \sum_{i=1}^{s/g} \boldsymbol{w}^i - \boldsymbol{\beta}^* \right\|_2^2 + \sigma_\epsilon^2 \left\| \sum_{i=1}^{s/g} \boldsymbol{w}^i \right\|_2^2 \tag{74}$$

By lemma 3.3, with high probability, for all $i \in \{1, ..., s/g\}$, $\text{supp}(\boldsymbol{w}^i) \subseteq S_i$, therefore, the above loss can be written as

$$\left\| \sum_{i=1}^{s/g} (\boldsymbol{w}^i - \boldsymbol{\beta}^{*i}) \right\|_2^2 + \sigma_\epsilon^2 \left\| \sum_{i=1}^{s/g} \boldsymbol{w}^i \right\|_2^2 \tag{75}$$

Furthermore, since $S_i \cap S_j = \varnothing$ for every $i \neq j$, the above can be written as

$$\sum_{i=1}^{s/g} \left( \left\| \boldsymbol{w}^i - \boldsymbol{\beta}^{*i} \right\|_2^2 + \sigma_\epsilon^2 \left\| \boldsymbol{w}^i \right\|_2^2 \right) \tag{76}$$

Again, by lemma 3.3, the above can be bounded with

$$\sum_{i=1}^{s/g} \left( \left\| \boldsymbol{w}^i - \boldsymbol{\beta}^{*i} \right\|_2^2 + \sigma_\epsilon^2 \left\| \boldsymbol{w}^i \right\|_2^2 \right) \in \mathcal{O}\left( \frac{1}{\gamma^2 \Lambda_{min}^2} \frac{\sigma_\epsilon^2 \|\boldsymbol{\beta}^*\|_2^2}{1 + \sigma_\epsilon^2} \right) \tag{77}$$

□

## 3.5 Main Lemmas

**We denote $H_{\text{thm}}$ as the intersection of the high-probability events (over the randomness of $\boldsymbol{Z}$) from the theorem's assumptions.** In other words, $H_{thm}$ contains the events described in assumptions 5. and 6. in the theorem's statement.

Now, according to the assumptions of Theorem 3, $H_{\text{thm}}$ is assumed to happen with high probability. In the following, we will show that with high probability the solutions $\{\boldsymbol{w}^i\}_{i=1}^{s/g}$ will also exhibit certain desirable properties. We will establish these results by showing that the intersection of $H_{\text{thm}}$ and the event that these properties hold has a probability very close to $\mathbb{P}(H_{\text{thm}})$.

**Lemma 3.2.** *Let the assumptions in the theorem hold. For every $i \in \{1, ..., s/g\}$, choose the $\lambda_i$ in problem (63) as follows*

$$\lambda_i = \frac{20}{\gamma}\sqrt{\frac{\log(d_x)\sigma_\epsilon^2\|\boldsymbol{\beta}^{*i}\|_2^2 K_x}{N_s}}. \tag{78}$$

*Then with probability at least $\mathbb{P}(H_{thm}) - 5\exp(-c\log(d_x))$ with $c \geq 1$, the following are both true: (i) event $H_{thm}$ happens; (ii) for all $i$, the solution $\boldsymbol{w}^i$ is unique, and $supp(\boldsymbol{w}^i) \subseteq S_i$ is true.*

*Proof.* We adopt the approach of the primal dual witness method in [7]. Notice that our optimization problem (63) can be rewritten as (since $\boldsymbol{X}_\epsilon = \boldsymbol{X} + \boldsymbol{Z}$)

$$\|\boldsymbol{X}_\epsilon(\boldsymbol{w} - \boldsymbol{\beta}^{*i}) + \boldsymbol{Z}\boldsymbol{\beta}^{*i}\|_2^2 + \lambda_i\|\boldsymbol{w}\|_1. \tag{79}$$

It has the same form as theirs (the only difference is that the noise term for us is $-\boldsymbol{Z}\boldsymbol{\beta}^{*i}$ , while for them it is a noise vector that is independent from the design matrix). Hence, we may directly apply lemmas 2(a) and 3(a) in [7]. Therefore, it suffices for us to prove that, with high probability, for every $i \in \{1, ..., s/g\}$, $(\boldsymbol{X}_{\epsilon,S_i}^T\boldsymbol{X}_{\epsilon,S_i})^{-1}$ exists, and the following is true:

$$\max_{j\notin S_i}\left|\boldsymbol{X}_{\epsilon,j}^T\left[\boldsymbol{X}_{\epsilon,S_i}(\boldsymbol{X}_{\epsilon,S_i}^T\boldsymbol{X}_{\epsilon,S_i})^{-1}\boldsymbol{h}_{S_i} + \boldsymbol{P}^\perp_{\boldsymbol{X}_{\epsilon,S_i}}\left(\frac{-\boldsymbol{Z}\boldsymbol{\beta}^{*i}}{\lambda_i N_s}\right)\right]\right| < 1 \tag{80}$$

where $\boldsymbol{h}_{S_i}$ is a subgradient vector for the $\ell_1$ norm coming from the primal dual witness construction ([7] equation (10)), so $\|\boldsymbol{h}_{S_i}\|_\infty \leq 1$. Note that $(\boldsymbol{X}_{\epsilon,S_i}^T\boldsymbol{X}_{\epsilon,S_i})^{-1}$ exists as long as $H_{\text{thm}}$ happens. Additionally, recall that $\boldsymbol{P}^\perp_{\boldsymbol{X}_{\epsilon,S_i}}$ denotes the projection onto the orthogonal complement of the column space of $\boldsymbol{X}_{\epsilon,S_i}$.

Apply the triangle inequality to the term on the left of the above inequality. We obtain the upper bound

$$\text{LHS of (80)} \leq \max_{j\notin S_i}|\boldsymbol{X}_{\epsilon,j}^T\boldsymbol{X}_{\epsilon,S_i}(\boldsymbol{X}_{\epsilon,S_i}^T\boldsymbol{X}_{\epsilon,S_i})^{-1}\boldsymbol{h}_{S_i}| + \max_{j\notin S_i}\left|\boldsymbol{X}_{\epsilon,j}^T\boldsymbol{P}^\perp_{\boldsymbol{X}_{\epsilon,S_i}}\left(\frac{\boldsymbol{Z}\boldsymbol{\beta}^{*i}}{\lambda_i N_s}\right)\right|. \tag{81}$$

If $H_{\text{thm}}$ happens, we can apply the mutual incoherence condition and Hölder's inequality to obtain:

$$\max_{j\notin S_i}|\boldsymbol{X}_{\epsilon,j}^T\boldsymbol{X}_{\epsilon,S_i}(\boldsymbol{X}_{\epsilon,S_i}^T\boldsymbol{X}_{\epsilon,S_i})^{-1}\boldsymbol{h}_{S_i}| \leq \max_{j\notin S_i}\|\boldsymbol{X}_{\epsilon,j}^T\boldsymbol{X}_{\epsilon,S_i}(\boldsymbol{X}_{\epsilon,S_i}^T\boldsymbol{X}_{\epsilon,S_i})^{-1}\|_1\|\boldsymbol{h}_{S_i}\|_\infty \leq 1 - \gamma. \tag{82}$$

Then, to show (80), it only remains to show

$$\max_{j\notin S_i}\left|\boldsymbol{X}_{\epsilon,j}^T\boldsymbol{P}^\perp_{\boldsymbol{X}_{\epsilon,S_i}}\left(\frac{\boldsymbol{Z}\boldsymbol{\beta}^{*i}}{\lambda_i N_s}\right)\right| \leq \frac{\gamma}{2} \tag{83}$$

holds for all $i$ with probability at least $1 - 5\exp(-c\log(d_x))$.

First note that

$$\left|\boldsymbol{X}_{\epsilon,j}^T\boldsymbol{P}^\perp_{\boldsymbol{X}_{\epsilon,S_i}}\left(\frac{\boldsymbol{Z}\boldsymbol{\beta}^{*i}}{\lambda_i N_s}\right)\right| \leq \left|\boldsymbol{X}_j^T\boldsymbol{P}^\perp_{\boldsymbol{X}_{\epsilon,S_i}}\left(\frac{\boldsymbol{Z}\boldsymbol{\beta}^{*i}}{\lambda_i N_s}\right)\right| + \left|\boldsymbol{Z}_j^T\boldsymbol{P}^\perp_{\boldsymbol{X}_{\epsilon,S_i}}\left(\frac{\boldsymbol{Z}\boldsymbol{\beta}^{*i}}{\lambda_i N_s}\right)\right|. \tag{84}$$

The first term on the right hand side (inside the absolute value) is zero-mean sub-Gaussian with parameter at most (by lemma 3.8)

$$(1/\lambda_i^2 N_s^2)\sigma_\epsilon^2\|\boldsymbol{\beta}^{*i}\|_2^2\|\boldsymbol{P}^\perp_{\boldsymbol{X}_{\epsilon,S_i}}\boldsymbol{X}_j\|_2^2 \leq (1/\lambda_i^2 N_s)\sigma_\epsilon^2\|\boldsymbol{\beta}^{*i}\|_2^2 K_x \tag{85}$$

where we recall from the theorem's assumption that, the columns of $\boldsymbol{X}$ satisfy $N_s^{-1}\|\boldsymbol{X}_i\|_2^2 \leq K_x$ for all $i$, with $K_x \in \mathcal{O}(1)$. We also made use of the fact that the spectral norm of projection matrices is 1.

Therefore, the following is true in general:

$$\mathbb{P}\left(\left|\boldsymbol{X}_j^T \boldsymbol{P}^{\perp}_{\boldsymbol{X}_{\epsilon,S_i}}\left(\frac{\boldsymbol{Z}\boldsymbol{\beta}^{*i}}{\lambda_i N_s}\right)\right| > \gamma/4\right) \leq 2\exp\left(-\frac{\lambda_i^2 N_s}{\sigma_\epsilon^2 \|\boldsymbol{\beta}^{*i}\|_2^2 K_x}\frac{\gamma^2}{32}\right). \tag{86}$$

To ensure the inequality over all $j \notin S_i$, we apply the union bound and obtain:

$$\mathbb{P}\left(\max_{j\notin S_i}\left|\boldsymbol{X}_j^T \boldsymbol{P}^{\perp}_{\boldsymbol{X}_{\epsilon,S_i}}\left(\frac{\boldsymbol{Z}\boldsymbol{\beta}^{*i}}{\lambda_i N_s}\right)\right| > \gamma/4\right) \leq 2\exp\left(-\frac{\lambda_i^2 N_s}{\sigma_\epsilon^2 \|\boldsymbol{\beta}^{*i}\|_2^2 K_x}\frac{\gamma^2}{32} + \log(d_x - g)\right). \tag{87}$$

Our choice of $\lambda_i$ ensures that the above probability is upper bounded by $2\exp(-12\log(d_x))$.

Now we deal with the second term on the right-hand side of (84):

$$\left|\boldsymbol{Z}_j^T \boldsymbol{P}^{\perp}_{\boldsymbol{X}_{\epsilon,S_i}}\left(\frac{\boldsymbol{Z}\boldsymbol{\beta}^{*i}}{\lambda_i N_s}\right)\right| = \left|\boldsymbol{Z}_j^T \boldsymbol{P}^{\perp}_{\boldsymbol{X}_{\epsilon,S_i}}\left(\frac{\sum_{k\in S_i}\boldsymbol{Z}_k\beta_k^*}{\lambda_i N_s}\right)\right|. \tag{88}$$

Note that $\boldsymbol{Z}_j$ for $j \notin S_i$ is independent from $\boldsymbol{P}^{\perp}_{\boldsymbol{X}_{\epsilon,S_i}}\left(\frac{\sum_{k\in S_i}\boldsymbol{Z}_k\beta_k^*}{\lambda_i N_s}\right)$ (the only random terms in it are the $\boldsymbol{Z}_k$'s with $k \in S_i$). Therefore, this second term also is zero mean, and in fact has a Gaussian-type tail bound. In particular, denoting $\boldsymbol{v} = \boldsymbol{P}^{\perp}_{\boldsymbol{X}_{\epsilon,S_i}}\left(\frac{\sum_{k\in S_i}\boldsymbol{Z}_k\beta_k^*}{\sqrt{N_s}}\right)$, we can write

$$\boldsymbol{Z}_j^T \boldsymbol{P}^{\perp}_{\boldsymbol{X}_{\epsilon,S_i}}\left(\frac{\boldsymbol{Z}\boldsymbol{\beta}^{*i}}{\lambda_i N_s}\right) = \boldsymbol{Z}_j^T\boldsymbol{v} = \left(\frac{1}{\sqrt{N_s}\lambda_i}\boldsymbol{Z}_j^T\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_2}\right)\|\boldsymbol{v}\|_2. \tag{89}$$

Notice that due to the rotational invariance of $\boldsymbol{Z}_j$, the inner product now produces a Normal random variable regardless of what $\boldsymbol{v}$ is. Furthermore,

$$\begin{aligned}\mathbb{P}\left(\left|\boldsymbol{Z}_j^T \boldsymbol{P}^{\perp}_{\boldsymbol{X}_{\epsilon,S_i}}\left(\frac{\sum_{j\in S_i}\boldsymbol{Z}_j\beta_j^*}{\lambda_i N_s}\right)\right| > \frac{\gamma}{4}\right) \leq& \mathbb{P}\left(\left|\frac{1}{\sqrt{N_s}\lambda_i}\boldsymbol{Z}_j^T\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_2}\right| > \frac{\gamma}{4}\frac{1}{2\sigma_\epsilon\|\boldsymbol{\beta}^{*i}\|_2}\right) \\ &+ \mathbb{P}\left(\|\boldsymbol{v}\|_2 \geq 2\sigma_\epsilon\|\boldsymbol{\beta}^{*i}\|_2\right).\end{aligned} \tag{90}$$

Let's bound the first probability. Since $\frac{\boldsymbol{Z}_j^T\boldsymbol{v}}{\|\boldsymbol{v}\|_2\sqrt{N_s}\lambda_i}$ is zero-mean sub-Gaussian with parameter at most $\sigma_\epsilon^2/(\lambda_i^2 N_s)$, by lemma 3.8 and union bound we have

$$\mathbb{P}\left(\max_{j\notin S_i}\left|\frac{1}{\sqrt{N_s}\lambda_i}\boldsymbol{Z}_j^T\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_2}\right| > \frac{\gamma}{4}\frac{1}{2\sigma_\epsilon\|\boldsymbol{\beta}^{*i}\|_2}\right) \leq 2\exp\left(-\frac{\gamma^2}{128}\frac{N_s\lambda_i^2}{\sigma_\epsilon^4\|\boldsymbol{\beta}^{*i}\|_2^2} + \log(d_x - g)\right). \tag{91}$$

With our choice of $\lambda_i$, we can upper bound the above probability by $2\exp(-2\log(d_x))$.

The second probability can be bounded with

$$\mathbb{P}\left(\|\boldsymbol{v}\|_2 \geq 2\sigma_\epsilon\|\boldsymbol{\beta}^{*i}\|_2\right) \leq \mathbb{P}\left(\|\boldsymbol{Z}\boldsymbol{\beta}^*\|_2^2/N_s \geq 2\sigma_\epsilon^2\|\boldsymbol{\beta}^{*i}\|_2^2\right) \tag{92}$$

since $\left\|\boldsymbol{P}^{\perp}_{\boldsymbol{X}_{\epsilon,S_i}}\left(\frac{\boldsymbol{Z}\boldsymbol{\beta}^{*i}}{\sqrt{N_s}}\right)\right\|_2 \leq \|\boldsymbol{Z}\boldsymbol{\beta}^{*i}\|_2/\sqrt{N_s}$. The upper bound on this probability then follows from lemma 3.6, and is at most $\exp(-N_s/16)$. With an appropriate choice of $N_s \in \Omega(g^2\log(d_x))$ (sufficiently large constant to multiply with $g^2\log(d_x)$), $\exp(-N_s/16)$ is dominated by $\exp(-2\log(d_x))$.

From the above bounds, we now know that, the following holds with probability at least $1-5\exp(-c'\log(d_x))$ with $c' \geq 2$ (in high dimensions):

$$\max_{j\notin S_i}\left|\boldsymbol{X}_{\epsilon,j}^T \boldsymbol{P}^{\perp}_{\boldsymbol{X}_{\epsilon,S_i}}\left(\frac{\boldsymbol{Z}\boldsymbol{\beta}^{*i}}{\lambda_i N_s}\right)\right| \leq \gamma/2. \tag{93}$$

To ensure that the above inequality holds for all $i$, we take a union bound, and end up with the above inequality holding for all $i$ with probability at least $1 - \exp(-c\log(d_x))$ with $c \geq 1$. Combining this with (82), with probability at least $\mathbb{P}(H_{\text{thm}}) - 5\exp(-c\log(d_x))$ with $c \geq 1$, the event $H_{\text{thm}}$ is true, and the following holds for all $i$ (which completes the proof)

$$\max_{j \notin S_i} \left| \boldsymbol{X}_{\epsilon,j}^T \left[ \boldsymbol{X}_{\epsilon,S_i} (\boldsymbol{X}_{\epsilon,S_i}^T \boldsymbol{X}_{\epsilon,S_i})^{-1} \boldsymbol{h}_{S_i} + \boldsymbol{P}^{\perp}_{\boldsymbol{X}_{\epsilon,S_i}} \left( \frac{\boldsymbol{Z}\boldsymbol{\beta}^{*i}}{\lambda_i N_s} \right) \right] \right| < 1 - \frac{\gamma}{2} < 1. \tag{94}$$

□

**Lemma 3.3.** *Assume the conditions in the theorem hold. Choose*

$$\lambda_i = \frac{20}{\gamma} \sqrt{\frac{\log(d_x)\sigma_\epsilon^2 \|\boldsymbol{\beta}^{*i}\|_2^2 K_x}{N_s}} \tag{95}$$

*(same as the last lemma). Then, with probability at least $\mathbb{P}(H_{thm}) - 5\exp(-c\log(d_x)) - 3\exp(-\log(d_x))$, the following are both true: (i) $H_{thm}$ holds; (ii) for all $i \in \{1, ..., s/g\}$, the solution $\boldsymbol{w}^i$ is unique and supp$(\boldsymbol{w}^i) \subseteq$ supp$(\boldsymbol{\beta}^{*i})$ is true, and the following is true:*

$$\mathbb{E}\left[ ((\boldsymbol{x} + \boldsymbol{\epsilon})^T \boldsymbol{w}^i - \boldsymbol{x}^T \boldsymbol{\beta}^{*i})^2 \right] \leq \mathcal{O}\left( \frac{1}{\gamma^2 \Lambda_{min}^2} \frac{\sigma_\epsilon^2 \|\boldsymbol{\beta}^{*i}\|_2^2}{1 + \sigma_\epsilon^2} \right). \tag{96}$$

*Proof.* Recall from the previous lemma that, with probability at least $\mathbb{P}(H_{\text{thm}}) - 5\exp(-c\log(d_x))$ (some $c \geq 1$), $H_{\text{thm}}$ holds, *and* for all $i$, $\boldsymbol{w}^i$ is unique and $\text{supp}(\boldsymbol{w}^i) \subseteq \text{supp}(\boldsymbol{\beta}^{*i})$. Let's call this overall event $H_{\text{good}}$.

Assuming $H_{\text{good}}$, the following inequality is true, since $\boldsymbol{w}^i$ is the unique solution to the problem (63):

$$\|\boldsymbol{X}_\epsilon(\boldsymbol{w}^i - \boldsymbol{\beta}^{*i}) + \boldsymbol{Z}\boldsymbol{\beta}^{*i}\|_2^2/N_s + \lambda_i \|\boldsymbol{w}^i\|_1 \leq \|\boldsymbol{X}_\epsilon(\boldsymbol{\beta}^{*i} - \boldsymbol{\beta}^{*i}) + \boldsymbol{Z}\boldsymbol{\beta}^{*i}\|_2^2/N_s + \lambda_i \|\boldsymbol{\beta}^{*i}\|_1 \tag{97}$$

$$= \|\boldsymbol{Z}\boldsymbol{\beta}^{*i}\|_2^2/N_s + \lambda_i \|\boldsymbol{\beta}^{*i}\|_1. \tag{98}$$

By expanding the first square and cancelling out the $\|\boldsymbol{Z}\boldsymbol{\beta}^{*i}\|_2^2$, we have:

$$\|\boldsymbol{X}_\epsilon(\boldsymbol{w}^i - \boldsymbol{\beta}^{*i})\|_2^2/N_s + \lambda_i \|\boldsymbol{w}^i\|_1 \leq \lambda_i \|\boldsymbol{\beta}^{*i}\|_1 + 2\boldsymbol{\beta}^{*iT} \boldsymbol{Z}^T (\boldsymbol{X} + \boldsymbol{Z})(\boldsymbol{\beta}^{*i} - \boldsymbol{w}^i)/N_s. \tag{99}$$

Now, note that $\|\boldsymbol{w}^i\|_1 = \|\boldsymbol{w}^i - \boldsymbol{\beta}^{*i} + \boldsymbol{\beta}^{*i}\|_1 \geq \|\boldsymbol{\beta}^{*i}\|_1 - \|\boldsymbol{w}^i - \boldsymbol{\beta}^*\|_1$. Therefore, the above inequality leads to

$$\|\boldsymbol{X}_\epsilon(\boldsymbol{w}^i - \boldsymbol{\beta}^{*i})\|_2^2/N_s \leq \lambda_i \|\boldsymbol{w}^i - \boldsymbol{\beta}^{*i}\|_1 + 2\boldsymbol{\beta}^{*iT} \boldsymbol{Z}^T (\boldsymbol{X} + \boldsymbol{Z})(\boldsymbol{\beta}^{*i} - \boldsymbol{w}^i)/N_s. \tag{100}$$

The rest of the proof relies on two main claims. We prove each of them next.

**Claim 1**: Assuming that $H_{\text{good}}$ happens, with constant $\widehat{C}_1 \in \mathcal{O}(1)$, for all $i$,

$$\lambda_i \|\boldsymbol{w}^i - \boldsymbol{\beta}^{*i}\|_1 \leq \frac{\widehat{C}_1}{\gamma} \sigma_\epsilon \|\boldsymbol{\beta}^{*i}\|_2 \|\boldsymbol{w}^i - \boldsymbol{\beta}^{*i}\|_2. \tag{101}$$

**Proof of Claim 1**: Recall that we require $\lambda_i = \frac{20}{\gamma}\sqrt{\frac{\log(d_x)\sigma_\epsilon^2 \|\boldsymbol{\beta}^{*i}\|_2^2 K_x}{N_s}}$. Therefore,

$$\lambda_i \|\boldsymbol{w}^i - \boldsymbol{\beta}^{*i}\|_1 = \left( \frac{20}{\gamma} \sqrt{\frac{g\log(d_x)K_x}{N_s}} \right) \sigma_\epsilon \|\boldsymbol{\beta}^{*i}\|_2 \|\boldsymbol{w}^i - \boldsymbol{\beta}^{*i}\|_2 \tag{102}$$

where we used the fact that $\|\cdot\|_1 \leq \sqrt{g}\|\cdot\|_2$ for $g$-dimensional vectors.

By making an appropriate choice of $N_s \in \Omega(g^2 \log(d_x))$ (choosing a sufficiently large constant to multiply with $g^2\log(d_x)$), we have that the following is true:

$$\lambda_i \|\boldsymbol{w}^i - \boldsymbol{\beta}^{*i}\|_1 \leq \frac{\widehat{C}_1}{\gamma} \sigma_\epsilon \|\boldsymbol{\beta}^{*i}\|_2 \|\boldsymbol{w}^i - \boldsymbol{\beta}^{*i}\|_2. \tag{103}$$

□

**Claim 2**: With probability at least $\mathbb{P}(H_{\text{thm}}) - \exp(-c\log(d_x)) - 3\exp(-\log(d_x))$ and constant $\widehat{C}_2 \in \mathcal{O}(1)$, $H_{\text{good}}$ holds, and for all $i$,

$$2\boldsymbol{\beta}^{*iT}\boldsymbol{Z}^T(\boldsymbol{X}+\boldsymbol{Z})(\boldsymbol{\beta}^{*i}-\boldsymbol{w}^i)/N_s \le \widehat{C}_2\sigma_\epsilon^2\|\boldsymbol{\beta}^{*i}\|_2\|\boldsymbol{\beta}^{*i}-\boldsymbol{w}^i\|_2. \tag{104}$$

**Proof of Claim 2**: We first note that, if $H_{\text{good}}$ happens, then :

$$\begin{aligned}
&2\boldsymbol{\beta}^{*iT}\boldsymbol{Z}^T(\boldsymbol{X}+\boldsymbol{Z})(\boldsymbol{\beta}^{*i}-\boldsymbol{w}^i)/N_s \\
=&2\boldsymbol{\beta}_{S_i}^{*iT}\boldsymbol{Z}_{S_i}^T(\boldsymbol{X}_{S_i}+\boldsymbol{Z}_{S_i})(\boldsymbol{\beta}_{S_i}^{*i}-\boldsymbol{w}_{S_i}^i)/N_s \\
\le&2|\boldsymbol{\beta}_{S_i}^{*iT}\boldsymbol{Z}_{S_i}^T\boldsymbol{X}_{S_i}(\boldsymbol{\beta}_{S_i}^{*i}-\boldsymbol{w}_{S_i}^i)|/N_s + 2|\boldsymbol{\beta}_{S_i}^{*iT}\boldsymbol{Z}_{S_i}^T\boldsymbol{Z}_{S_i}(\boldsymbol{\beta}_{S_i}^{*i}-\boldsymbol{w}_{S_i}^i)|/N_s.
\end{aligned} \tag{105}$$

The rest of the proof has two main steps, and the claimed inequality is a direct consequence of the results from the two steps and the above inequality:

1. For the first term on the right of the above inequality, consider the following basic upper bound: suppose $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^p$, and $\boldsymbol{M} \in \mathbb{R}^{p\times p}$, then

$$\begin{aligned}
|\boldsymbol{x}_1^T\boldsymbol{M}\boldsymbol{x}_2| =&|\sum_{i=1}^p\sum_{j=1}^p M_{i,j}x_{1,i}x_{2,j}| \\
\le&\sum_{i=1}^p\sum_{j=1}^p |M_{i,j}x_{1,i}x_{2,j}| \\
\le&\max_{1\le i',j'\le p}|M_{i',j'}|\sum_{i=1}^p\sum_{j=1}^p|x_{1,i}||x_{2,j}| \\
=&\max_{1\le i',j'\le p}|M_{i',j'}|\|\boldsymbol{x}_1\|_1\|\boldsymbol{x}_2\|_1.
\end{aligned} \tag{106}$$

   So we have

$$|\boldsymbol{\beta}_{S_i}^{*iT}\boldsymbol{Z}_{S_i}^T\boldsymbol{X}_{S_i}(\boldsymbol{\beta}_{S_i}^{*i}-\boldsymbol{w}_{S_i}^i)|/N_s \le \max_{1\le i',j'\le g}\left|[\boldsymbol{Z}_{S_i}^T\boldsymbol{X}_{S_i}]_{i',j'}\right|\|\boldsymbol{\beta}_{S_i}\|_1\|\boldsymbol{\beta}^{*i}-\boldsymbol{w}^i\|_1/N_s. \tag{107}$$

   But note that, by lemma 3.4, in general (not assuming $H_{\text{good}}$), with probability at least $1-\exp(-2\log(d_x))$,

$$\max_{1\le i,j\le g}|[\boldsymbol{Z}_{S_i}^T\boldsymbol{X}_{S_i}]_{i,j}|/N_s \le \sigma_\epsilon\sqrt{\frac{K_x(4\log(d_x)+4\log(g))}{N_s}}. \tag{108}$$

   Therefore, if $H_{\text{good}}$ *and* the event of lemma 3.4 happen, we have

$$\begin{aligned}
&\boldsymbol{\beta}^{*iT}\boldsymbol{Z}^T\boldsymbol{X}(\boldsymbol{\beta}^{*i}-\boldsymbol{w}^i)/N_s \\
\le&|\boldsymbol{\beta}_{S_i}^{*iT}\boldsymbol{Z}_{S_i}^T\boldsymbol{X}_{S_i}(\boldsymbol{\beta}_{S_i}^{*i}-\boldsymbol{w}_{S_i}^i)|/N_s \\
\le&\sigma_\epsilon\|\boldsymbol{\beta}^{*i}\|_1\times\sqrt{\frac{K_x(4\log(d_x)+4\log(g))}{N_s}}\|\boldsymbol{\beta}^{*i}-\boldsymbol{w}^i\|_1 \\
\le&\sigma_\epsilon\|\boldsymbol{\beta}^{*i}\|_2\times g\sqrt{\frac{K_x(4\log(d_x)+4\log(g))}{N_s}}\|\boldsymbol{\beta}^{*i}-\boldsymbol{w}^i\|_2.
\end{aligned} \tag{109}$$

   In the last inequality, we used the fact that $\|\cdot\|_1 \le \sqrt{g}\|\cdot\|_2$ for $g$-dimensional real vectors.

   With a proper choice of $N_s \in \Omega(g^2\log(d_x))$ (choosing a large enough constant for multiplying with $g^2\log(d_x)$), the above inequality leads to

$$2\boldsymbol{\beta}^{*iT}\boldsymbol{Z}^T\boldsymbol{X}(\boldsymbol{\beta}^{*i}-\boldsymbol{w}^i)/N_s \le C_1\sigma_\epsilon\|\boldsymbol{\beta}^{*i}\|_2\|\boldsymbol{\beta}^{*i}-\boldsymbol{w}^i\|_2 \tag{110}$$

for constant $C_1 \in \mathcal{O}(1)$.

Taking a union bound over all $i$, in general, the event from lemma 3.4 is true for all $i$ with probability at least $1-\exp(-\log(d_x))$. Therefore, in general, the above inequality is true for all $i$ with probability at least $\mathbb{P}(H_{\text{thm}}) - 5\exp(-c\log(d_x)) - \exp(-\log(d_x))$, since we need $H_{\text{good}}$ and the union of events (over all $i$) of lemma 3.4 to both hold.

2. Now we upper bound the second inner product term, $|\boldsymbol{\beta}^{*iT}_{S_i}\boldsymbol{Z}^T_{S_i}\boldsymbol{Z}_{S_i}(\boldsymbol{\beta}^{*i}_{S_i} - \boldsymbol{w}^i_{S_i})|/N_s$. By lemma 3.7, denoting $\boldsymbol{E}_i = \boldsymbol{Z}^T_{S_i}\boldsymbol{Z}_{S_i}/N_s - \sigma^2_\epsilon \boldsymbol{I}_{g\times g}$, we have that in general (not assuming $H_{\text{good}}$), for some universal constant $C_2$, with probability at least $1-2\exp(-2\log(d_x))$,

$$\|\boldsymbol{E}_i\|_2 \leq C_2\sigma_\epsilon\sqrt{\frac{g+2\log(d_x)}{N_s}} \tag{111}$$

where $\|\cdot\|_2$ represents the spectral norm for square matrices. Note that the above is true *for all* $i$ with probability at least $1-2\exp(-\log(d_x))$. With appropriate choice of $N_s \in \Omega(g^2\log(d_x))$ (sufficiently large constant to multiply with $g^2\log(d_x)$), the above expression simplifies to $\|\boldsymbol{E}_i\|_2 \leq C_2\sigma_\epsilon$, for some $C_2 \in \mathcal{O}(1)$.

Now, if $H_{\text{good}}$ and the union of events (over all $i$) from lemma 3.7 happen we may write, for all $i$,

$$|\boldsymbol{\beta}^{*iT}_{S_i}\boldsymbol{Z}^T_{S_i}\boldsymbol{Z}_{S_i}(\boldsymbol{\beta}^{*i}_{S_i} - \boldsymbol{w}^i_{S_i})|/N_s = \sigma_\epsilon|\boldsymbol{\beta}^{*iT}_{S_i}(\boldsymbol{\beta}^{*i}_{S_i} - \boldsymbol{w}^i_{S_i})| + |\boldsymbol{\beta}^{*iT}_{S_i}\boldsymbol{E}_i(\boldsymbol{\beta}^{*i}_{S_i} - \boldsymbol{w}^i_{S_i})| \tag{112}$$

$$\leq \sigma_\epsilon\|\boldsymbol{\beta}^{*i}\|_2\|\boldsymbol{\beta}^{*i} - \boldsymbol{w}^i\|_2 + C_2\sigma_\epsilon\|\boldsymbol{\beta}^{*i}\|_2\|\boldsymbol{\beta}^{*i} - \boldsymbol{w}^i\|_2 \tag{113}$$

where in the last step we have used $\sigma_\epsilon|\boldsymbol{\beta}^{*iT}_{S_i}(\boldsymbol{\beta}^{*i}_{S_i} - \boldsymbol{w}^i_{S_i})| \leq \sigma_\epsilon\|\boldsymbol{\beta}^{*i}\|_2\|\boldsymbol{\beta}^{*i} - \boldsymbol{w}^i\|_2$ thanks to Cauchy-Schwartz, and $|\boldsymbol{\beta}^{*iT}_{S_i}\boldsymbol{E}_i(\boldsymbol{\beta}^{*i}_{S_i} - \boldsymbol{w}^i_{S_i})| \leq C_2\sigma_\epsilon\|\boldsymbol{\beta}^{*i}\|_2\|\boldsymbol{\beta}^{*i} - \boldsymbol{w}^i\|_2$ which comes from the following basic inequality. Suppose we have $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^g$, and $\boldsymbol{M} \in \mathbb{R}^{g\times g}$ being symmetric, then $\boldsymbol{M}$ has an eigen-decomposition, and we write it as $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$. Then the following holds:

$$\begin{aligned}
\boldsymbol{x}^T\boldsymbol{M}\boldsymbol{y} &= \boldsymbol{x}^T\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T\boldsymbol{y} \\
&= \boldsymbol{x}^T\boldsymbol{U}\boldsymbol{\Lambda}^{1/2}\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^T\boldsymbol{y} \qquad (\text{Define } [\boldsymbol{\Lambda}^{1/2}]_{i,j} = [\boldsymbol{\Lambda}]^{1/2}_{i,j}) \\
&= (\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^T\boldsymbol{x})^T(\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^T\boldsymbol{y}) \\
&\leq \|\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^T\boldsymbol{x}\|_2\|\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^T\boldsymbol{y}\|_2 \qquad (\text{Cauchy Schwartz}) \\
&\leq \|\boldsymbol{M}\|_2\|\boldsymbol{U}^T\boldsymbol{x}\|_2\|\boldsymbol{U}^T\boldsymbol{y}\|_2 \\
&\leq \|\boldsymbol{M}\|_2\|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2 \qquad (\text{Orthogonal matrices preserve } \ell_2 \text{ norm})
\end{aligned} \tag{114}$$

Using (113), the inequality below is true with constant $\widetilde{C}_2 \in \mathcal{O}(1)$:

$$2|\boldsymbol{\beta}^{*iT}_{S_i}\boldsymbol{Z}^T_{S_i}\boldsymbol{Z}_{S_i}(\boldsymbol{\beta}^{*i}_{S_i} - \boldsymbol{w}^i_{S_i})|/N_s \leq \widetilde{C}_2\sigma_\epsilon\|\boldsymbol{\beta}^{*i}\|_2\|\boldsymbol{\beta}^{*i} - \boldsymbol{w}^i\|_2. \tag{115}$$

Now, let us summarize the probabilities calculated so far. From the previous point we need $H_{\text{good}}$ and the union of events of lemma 3.4 to be true. Now we also need the union of events of lemma 3.7 to be true, so we end up with a probability at least $\mathbb{P}(H_{\text{thm}}) - 5\exp(-c\log(d_x)) - \exp(-\log(d_x)) - 2\exp(-\log(d_x))$.

□

**Proof of Lemma 3.3 continued**: With Claim 1 and Claim 2 in hand, we arrive at the fact that, with constant $C \in \mathcal{O}(1)$, $c \geq 1$ and probability at least $\mathbb{P}(H_{\text{thm}}) - 5\exp(-c\log(d_x)) - 3\exp(-\log(d_x))$, $H_{\text{good}}$ holds, and for all $i \in \{1, ..., s/g\}$, the following is true

$$\|\boldsymbol{X}_{\epsilon,S_i}(\boldsymbol{w}^i_{S_i} - \boldsymbol{\beta}^{*i}_{S_i})\|_2^2/N_s \leq \frac{C}{\gamma}\sigma_\epsilon\|\boldsymbol{\beta}^{*i}\|_2\|\boldsymbol{w}^i - \boldsymbol{\beta}^{*i}\|_2. \tag{116}$$

All that is left is some algebraic manipulations.

Recalling that the minimum eigenvalue of $\boldsymbol{X}_{\epsilon,S_i}^T \boldsymbol{X}_{\epsilon,S_i}/N_s$ is $\Lambda_{min}^i > 0$, we have the inequality

$$\|\boldsymbol{X}_{\epsilon,S_i}(\boldsymbol{w}_{S_i}^i - \boldsymbol{\beta}_{S_i}^{*i})\|_2^2/N_s \geq \Lambda_{min}^i \|\boldsymbol{w}_{S_i}^i - \boldsymbol{\beta}_{S_i}^{*i}\|_2^2 \geq \Lambda_{min} \|\boldsymbol{w}_{S_i}^i - \boldsymbol{\beta}_{S_i}^{*i}\|_2^2. \tag{117}$$

It follows that,

$$\Lambda_{min}\|\boldsymbol{w}^i - \boldsymbol{\beta}^{*i}\|_2^2 \leq \frac{C}{\gamma}\sigma_\epsilon \|\boldsymbol{\beta}^{*i}\|_2 \|\boldsymbol{w}^i - \boldsymbol{\beta}^{*i}\|_2 \tag{118}$$

$$\implies \|\boldsymbol{w}^i - \boldsymbol{\beta}^{*i}\|_2 \leq \frac{C}{\gamma\Lambda_{min}}\sigma_\epsilon \|\boldsymbol{\beta}^{*i}\|_2. \tag{119}$$

Furthermore, by noting that $\sigma_\epsilon\|\boldsymbol{w}^i\|_2 \leq \sigma_\epsilon\|\boldsymbol{w}^i - \boldsymbol{\beta}^{*i}\|_2 + \sigma_\epsilon\|\boldsymbol{\beta}^{*i}\|_2$, $1/\sqrt{1+\sigma_\epsilon^2} \in \mathcal{O}(1)$, and using $C$ to absorb $\mathcal{O}(1)$ constants, we arrive at

$$\|\boldsymbol{w}^i - \boldsymbol{\beta}^{*i}\|_2 + \sigma_\epsilon\|\boldsymbol{w}^i\|_2 \leq \frac{C}{\gamma\Lambda_{min}} \frac{\sigma_\epsilon}{\sqrt{1+\sigma_\epsilon^2}} \|\boldsymbol{\beta}^{*i}\|_2. \tag{120}$$

Now, consider the following basic identity:

$$\sqrt{\|\boldsymbol{w}^i - \boldsymbol{\beta}^{*i}\|_2^2 + \sigma_\epsilon^2\|\boldsymbol{w}^i\|_2^2} \leq \|\boldsymbol{w}^i - \boldsymbol{\beta}^{*i}\|_2 + \sigma_\epsilon\|\boldsymbol{w}^i\|_2. \tag{121}$$

By noting that $\mathbb{E}\left[((\boldsymbol{x}+\boldsymbol{\epsilon})^T\boldsymbol{w}^i - \boldsymbol{x}^T\boldsymbol{\beta}^{*i})^2\right] = \|\boldsymbol{w}^i - \boldsymbol{\beta}^{*i}\|_2^2 + \sigma_\epsilon^2\|\boldsymbol{w}^i\|_2^2$ from section 3.3, and by combining (120) and (121), we obtain the desired expression in the lemma.

□

## 3.6 Probability Lemmas

**Lemma 3.4.** *Let $\boldsymbol{Z}$'s entries be sampled from $\mathcal{N}(0,\sigma_\epsilon^2)$ independently, and the columns of $\boldsymbol{X}$ satisfy $N_s^{-1}\|\boldsymbol{X}_i\|_2^2 \leq K_x$ for all $i$. $S_i \subset \{1,...,d_x\}$ is an index set of size $g$. Only $\boldsymbol{Z}$ is random, $\boldsymbol{X}$ is fixed.*

*For any $t > 0$, with probability at least $1 - \exp(-t^2/2)$,*

$$\max_{1\leq i,j\leq g} |[\boldsymbol{Z}_{S_i}^T\boldsymbol{X}_{S_i}]_{i,j}|/N_s \leq \sigma_\epsilon\sqrt{\frac{K_x(t^2+4\log(g))}{N_s}}. \tag{122}$$

*Proof.* Recall that $[\boldsymbol{Z}_{S_i}]_{:,i} \sim \mathcal{N}(\boldsymbol{0}, \sigma_\epsilon^2\boldsymbol{I}_{N\times N})$, and since $\boldsymbol{X}_{S_i}$ is deterministic, for each $i,j$, $[\boldsymbol{Z}_{S_i}^T\boldsymbol{X}_{S_i}]_{i,j} = [\boldsymbol{Z}_{S_i}]_{:,i}^T[\boldsymbol{X}_{S_i}]_{:,j} \sim \mathcal{N}(0, \sigma_\epsilon^2\|[\boldsymbol{X}]_{:,j}\|_2^2)$. Therefore, $[\boldsymbol{Z}_{S_i}^T\boldsymbol{X}_{S_i}]_{i,j}/(\sqrt{N_sK_x}\sigma_\epsilon)$ is zero-mean sub-Gaussian random variable with its parameter no greater than 1 for all $i,j$. Now we may apply the union bound and the tail bound for sub-Gaussian random variables (from lemma 3.8), and arrive at the following result: for any $t > 0$, the following holds:

$$\begin{aligned}
&\mathbb{P}\left(\max_{1\leq i,j\leq g} |[\boldsymbol{Z}_{S_i}^T\boldsymbol{X}_{S_i}]_{i,j}|/(\sqrt{N_sK_x}\sigma_\epsilon) \geq \sqrt{t^2+2\log(g^2)}\right) \\
\leq& 2g^2\exp\left\{-\frac{t^2+2\log(g^2)}{2}\right\} \\
=& 2\exp\left\{-\frac{t^2}{2}\right\}.
\end{aligned} \tag{123}$$

But clearly

$$\mathbb{P}\left(\max_{1\leq i,j\leq g} |[\boldsymbol{Z}_{S_i}^T\boldsymbol{X}_{S_i}]_{i,j}|/(\sqrt{N_sK_x}\sigma_\epsilon) \geq \sqrt{t^2+2\log(g^2)}\right) \tag{124}$$

$$=\mathbb{P}\left(\max_{1\leq i,j\leq g} |[\boldsymbol{Z}_{S_i}^T\boldsymbol{X}_{S_i}]_{i,j}|/\sqrt{N_s} \geq \sigma_\epsilon\sqrt{K_x(t^2+4\log(g))}\right). \tag{125}$$

The proof is complete. □

**Lemma 3.5.** *Let $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{d\times d})$. Then for any $\delta \geq 0$, the following is true:*

$$\mathbb{P}(\|\boldsymbol{x}\|_2^2 \geq d + \delta) \leq \left(\frac{d}{d+\delta}\right)^{-d/2} \exp(-\delta/2). \tag{126}$$

*Proof.* This is a relatively standard concentration bound for the $\ell_2$ norm of random Gaussian vectors. We provide its proof for the sake of completeness.

Denote $f_x$ as the probability density function of $\boldsymbol{x}$.

Choose $\lambda = \delta/(d+\delta)$. The following is true:

$$\|\boldsymbol{x}\|_2^2 \geq d + \delta \implies \exp(\lambda\|\boldsymbol{x}\|_2^2/2) \geq \exp(\lambda(d+\delta)/2). \tag{127}$$

Moreover,

$$\int_{\mathbb{R}^d} \exp(\lambda\|\boldsymbol{x}\|_2^2/2) f_x(\boldsymbol{x}) d\boldsymbol{x} \geq \mathbb{P}(\|\boldsymbol{x}\|_2^2 \geq d+\delta) \exp(\lambda(d+\delta)/2). \tag{128}$$

Therefore

$$\mathbb{P}(\|\boldsymbol{x}\|_2^2 \geq d + \delta) \leq \exp(-\lambda(d+\delta)/2) \int_{\mathbb{R}^d} \exp(\lambda\|\boldsymbol{x}\|_2^2/2) f_x(\boldsymbol{x}) d\boldsymbol{x}. \tag{129}$$

Explicitly computing the integral on the right-hand-side yields

$$\mathbb{P}(\|\boldsymbol{x}\|_2^2 \geq d + \delta) \leq (1-\lambda)^{-d/2} \exp(-\lambda(d+\delta)/2). \tag{130}$$

Substituting $\lambda = \delta/(d+\delta)$ into the expression completes the proof. □

**Corollary 3.5.1.** *Let $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{d\times d})$. Then for any $\epsilon \in (0,1)$, the following is true:*

$$\mathbb{P}(\|\boldsymbol{x}\|_2^2 \geq (1-\epsilon)^{-1} d) \leq \exp(-\epsilon^2 d/4). \tag{131}$$

*Proof.* In the result of lemma 3.5, choose $\delta = d\epsilon/(1-\epsilon)$.

Then $d + \delta = d/(1-\epsilon)$, and we obtain:

$$\mathbb{P}(\|\boldsymbol{x}\|_2^2 \geq d/(1-\epsilon)) \leq (1-\epsilon)^{-d/2} \exp\left(-\frac{d}{2}\frac{\epsilon}{1-\epsilon}\right) \leq \exp\left(-\frac{d}{2}\left(\frac{\epsilon}{1-\epsilon} + \log(1-\epsilon)\right)\right). \tag{132}$$

We obtain the desired expression by noting that

$$\frac{\epsilon}{1-\epsilon} + \log(1-\epsilon) \geq \epsilon^2/2. \tag{133}$$

□

**Lemma 3.6.** *Let the entries of $\boldsymbol{W} \in \mathbb{R}^{N\times d}$ be independent and have the random distribution $\mathcal{N}(0, \sigma_\epsilon^2)$, and $\boldsymbol{\beta}^* \in \mathbb{R}^d$.*

*For any $\delta \in (0,1)$, with probability at least $1 - \exp(-\delta^2 N/4)$,*

$$\|\boldsymbol{W}\boldsymbol{\beta}^*\|_2^2/N \leq (1-\delta)^{-1}\sigma_\epsilon^2\|\boldsymbol{\beta}^*\|_2^2. \tag{134}$$

*Proof.* Note that $\boldsymbol{W}\boldsymbol{\beta}^*/(\sigma_\epsilon\|\boldsymbol{\beta}^*\|_2) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{N\times N})$. We now invoke the concentration inequality for standard Gaussian random vector from corollary 3.5.1 to obtain

$$\mathbb{P}\left(\frac{\|\boldsymbol{W}\boldsymbol{\beta}^*\|_2^2}{N} \geq (1-\delta)^{-1}\sigma_\epsilon^2\|\boldsymbol{\beta}^*\|_2^2\right) = \mathbb{P}\left(\frac{\|\boldsymbol{W}\boldsymbol{\beta}^*\|_2^2}{\sigma_\epsilon^2\|\boldsymbol{\beta}^*\|_2^2} \geq (1-\delta)^{-1}N\right) \leq \exp(-\delta^2 N/4). \tag{135}$$

□

**Lemma 3.7** (Exercise 4.7.3 in [6], specialized to iid sub-Guassian vectors)**.** *Let $\boldsymbol{z}$ be a zero-mean sub-Gaussian random vector in $\mathbb{R}^g$ with independent and identically distributed entries, with each entry having the same sub-Gaussian random distribution. Moreover, define $\Sigma_Z = \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^T]$.*

*Given $\{\boldsymbol{z}_i\}_{i=1}^N$, then there exists universal constant $C$ such that, for any $u \geq 0$, the following is true with probability at least $1 - 2\exp(-u)$:*

$$\left\| \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{z}_i \boldsymbol{z}_i^T - \boldsymbol{\Sigma}_Z \right\|_2 \leq C \left( \sqrt{\frac{g+u}{N}} + \frac{g+u}{N} \right) \|\boldsymbol{\Sigma}_Z\|_2 \tag{136}$$

*where for matrices, $\|\cdot\|_2$ represents the spectral norm.*

**Lemma 3.8.** *Recall that a zero-mean random variable $X$ is sub-Gaussian if there exists some $\sigma > 0$ such that for all $t \in \mathbb{R}$*

$$\mathbb{E}\left[\exp(tX)\right] \leq \exp(\sigma^2 t^2/2). \tag{137}$$

*Moreover, $X$ must satisfy (from [7] Appendix A)*

$$\mathbb{P}(|X| > x) \leq 2\exp\left(-\frac{x^2}{2\sigma^2}\right). \tag{138}$$

*An additional useful result is that, if $X_1, ..., X_n$ are independent and zero-mean sub-Gaussian random variables with parameters $\sigma_1^2, ..., \sigma_n^2$, then $\sum_{i=1}^n X_i$ is sub-Gaussian with parameter $\sum_{i=1}^n \sigma_i^2$ (from [2] lemma 1.7).*

## 3.7 Experimental Result

We carry out the following simple experiment to further support the utility of student-teacher learning over target-based learning. We use the Lasso and LassoLars methods from the scikit-learn library to numerically solve the LASSO problems described below.

In this experiment, we focus on student-teacher learning and target-based LASSO learning. For student-teacher learning, we let $g = 1$. For target-based LASSO, we simply solve the following problem:

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^{d_x}}{\operatorname{argmin}} \|\boldsymbol{X}_\epsilon \boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1. \tag{139}$$

The exact experimental parameters are set as follows. We choose the list $D_x = \{500, 1000, 2000, 4000\}$. For every $d_x \in D_x$, we set the corresponding $\boldsymbol{\beta}^*$ with $\beta_j^* = 1.0$ for $j \in \{1, ..., d_x/20\}$, and 0 everywhere else. So for each $d_x$, $\boldsymbol{\beta}^*$ has a sparsity level $s = d_x/20$. The sample size $N_s = 5\log(d_x)$ for every $d_x$. The noise variance $\sigma_\epsilon^2 = 0.1$. To solve the base learning problem and the student-teacher learning problem, we run parameter sweep over $\lambda$, and report only the best testing error out of all the $\lambda$'s chosen.

Figure 2 reports the testing error of the network trained with student-teacher loss and the target-based loss. We also draw the optimal test error curve for comparison. The horizontal axis is $d_x$, and the vertical axis is the testing error. As $d_x$ increases, the testing error of the network trained with target-based LASSO diverges very quickly to infinity, while the testing error of the network trained with the student-teacher loss stays very close to the optimal one.

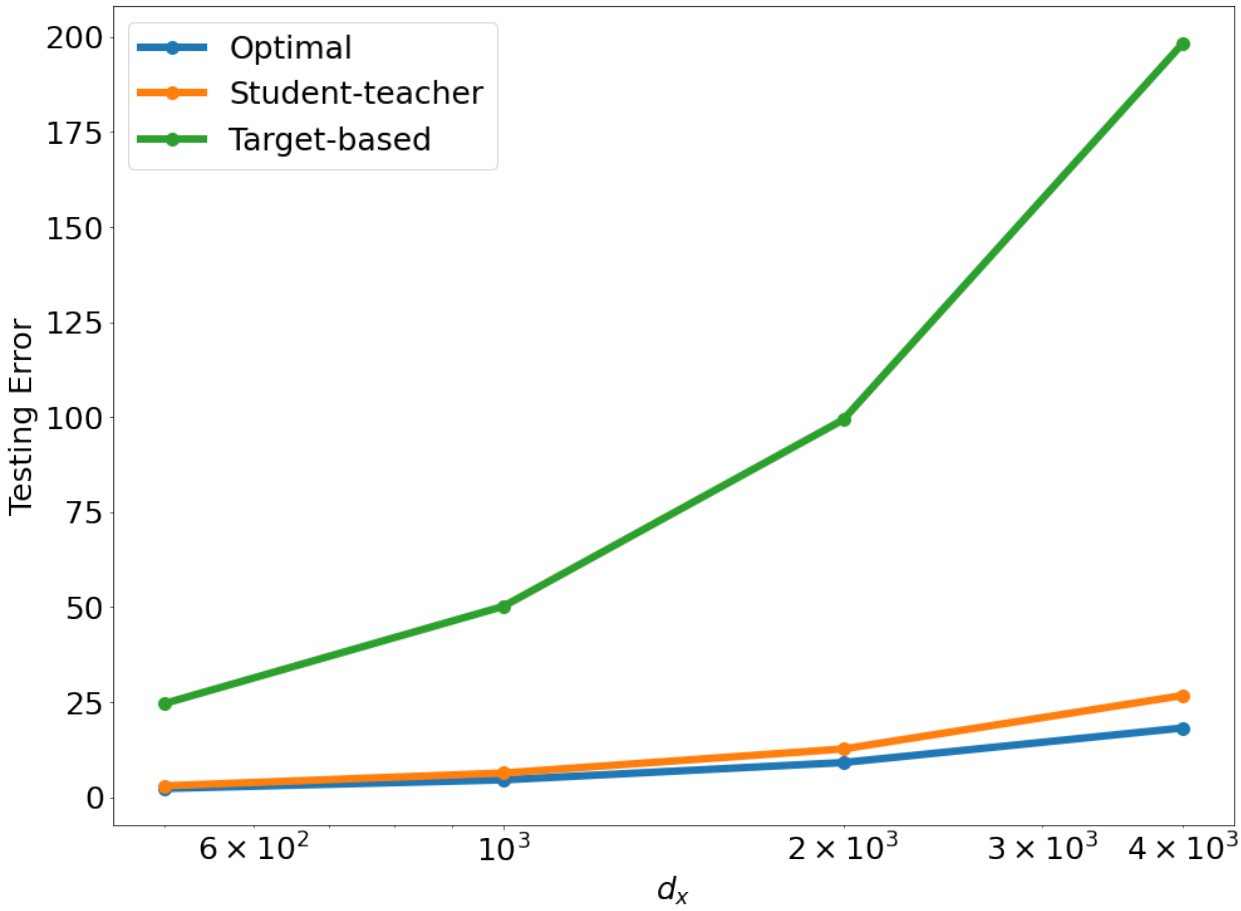


Figure 2: The testing error of the network trained with student-teacher loss and the target-based loss. Optimal testing error is also drawn for comparison. The horizontal axis $d_x$ indicates the data vector dimension, and the vertical axis indicates the testing error of the network. At each $d_x$, we set $s = d_x/20$, $N_s = 5\log(d_x)$. For student-teacher, $g = 1$. The noise variance $\sigma_\epsilon^2 = 0.1$. We carry out parameter sweep over $\lambda$ for both the target-based and student-teacher problem, and only report the best testing error.

# References


[1] Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *International Conference on Machine Learning (ICML)*, page 244–253, 2018.

[2] Valery V. Buldygin and Kozachenko Yu V. *Metric Characterization of Random Variables and Random Processes*. American Mathematical Society, 2000.

[3] T. Tony Cai and Tiefeng Jiang. Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices. *Annals of Statistics*, 39(3):1496–1525, 06 2011.

[4] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–218, 1936.

[5] Joel A. Tropp. Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 52(3):1030–1051, 2006.

[6] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[7] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using $\ell_1$ -constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.