Implicit Regularization Leads to Benign Overfitting for Sparse Linear Regression

Mo Zhou
Duke University
mozhou@cs.duke.edu

Rong Ge
Duke University
rongge@cs.duke.edu

February 2, 2023

Abstract

In deep learning, often the training process finds an interpolator (a solution with 0 training loss), but the test loss is still low. This phenomenon, known as *benign overfitting*, is a major mystery that received a lot of recent attention. One common mechanism for benign overfitting is *implicit regularization*, where the training process leads to additional properties for the interpolator, often characterized by minimizing certain norms. However, even for a simple sparse linear regression problem $y = \boldsymbol{\beta}^{*\top}\boldsymbol{x} + \xi$ with sparse $\boldsymbol{\beta}^*$, neither minimum ℓ_1 or ℓ_2 norm interpolator gives the optimal test loss. In this work, we give a different parametrization of the model which leads to a new implicit regularization effect that combines the benefit of ℓ_1 and ℓ_2 interpolators. We show that training our new model via gradient descent leads to an interpolator with near-optimal test loss. Our result is based on careful analysis of the training dynamics and provides another example of implicit regularization effect that goes beyond norm minimization.

1 Introduction

Benign overfitting – the phenomenon that the training loss becomes 0, but the test loss remains low – is a major mystery in deep learning. Recently, a long line of works (Belkin et al., 2019; Bartlett et al., 2020; Belkin et al., 2020; Hastie et al., 2022; Advani et al., 2020; Koehler et al., 2021) tried to explain why interpolators (solutions with 0 training loss) can still enjoy good test loss for various models. This phenomenon is interesting and was studied extensively even for simple models of linear regression (see e.g., Bartlett et al. (2020); Tsigler and Bartlett (2020); Hastie et al. (2022)), where data $(\boldsymbol{x}, y)$ is generated as

$$y = \boldsymbol{\beta}^{*\top}\boldsymbol{x} + \xi.$$

Here, $\boldsymbol{\beta}^*$ is an unknown vector that we hope to learn, $\boldsymbol{x}$ is generated from a data distribution and ξ represents the noise.

One of the major explanations for benign overfitting is *implicit regularization*, which suggests that the training process promotes additional properties for the interpolator that it finds. In the context of the simple linear regression, it was known that fitting the model $y = \boldsymbol{\beta}^\top\boldsymbol{x}$ directly by gradient descent gives the $\boldsymbol{\beta}$ with minimum ℓ_2 norm; while parametrizing $\boldsymbol{\beta}$ as $\boldsymbol{\beta} = \boldsymbol{w}^{\odot 2} - \boldsymbol{u}^{\odot 2}$ (here $\odot 2$ represents entry-wise square) gives the $\boldsymbol{\beta}$ with minimum ℓ_1 norm.

However, for sparse linear regression, implicit regularization in the form of ℓ_1 or ℓ_2 norm minimization does not lead to benign overfitting. More precisely, if $\boldsymbol{\beta}^* \in \mathbb{R}^d$ is an s-sparse vector, given n samples $(\boldsymbol{x}_i, y_i)$ for $i = 1, 2, ..., n$ where $n \ll d$, $\boldsymbol{x}_i \sim N(\boldsymbol{0}, \boldsymbol{I})$ and $\xi_i \sim N(0, \sigma^2)$, one can still hope to find a parameter $\boldsymbol{\beta}$ such that the test loss $\frac{1}{2}\mathbb{E}[(y - \boldsymbol{\beta}^\top\boldsymbol{x})^2]$ is on the order of $\sigma^2 s \log(d/s)/n$. Neither minimum ℓ_1 or ℓ_2 interpolator achieves anything near this guarantee: the best ℓ_2 norm interpolator achieves a test loss of $\Omega(\|\boldsymbol{\beta}^*\|_2^2)$ (Bartlett et al., 2020; Hastie et al., 2022) while the best ℓ_1 norm interpolator achieves a test loss of $\Omega(\sigma^2/\log(d/n))$ (Chatterji and Long, 2022; Wang et al., 2022).

In the sparse regression setting, Muthukumar et al. (2020) showed that when the model is significantly overparametrized ($d \gg n$), it is still possible to find an interpolator with near-optimal test loss. The interpolator in Muthukumar et al. (2020) has to be constructed *explicitly* through a 2-stage process which combines ℓ_1 and ℓ_2 norm minimization. In this paper, we ask whether such an interpolator can be found via *implicit regularization* – by directly minimizing the loss using a new parametrization.

1.1 Our result and technique

We show that implicit regularization can indeed give near-optimal interpolators (up to polylog factors) and therefore achieve benign overfitting in the sparse regression setting:

Theorem 1 (Informal). *In the sparse linear regression setting with unknown s-sparse target $\boldsymbol{\beta}^*$, suppose we parametrize linear function $\boldsymbol{\beta}^\top \boldsymbol{x}$ as*

$$\boldsymbol{\beta} = \boldsymbol{v} + \lambda(\boldsymbol{w}^{\odot 2} - \boldsymbol{u}^{\odot 2}).$$

If $\tilde{\Omega}(s^4) \leq n \leq \widetilde{O}(\sqrt{d})$, with proper choice of parameters, gradient descent converges to a solution $\boldsymbol{\beta}$ with 0 training loss and test loss

$$\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 = O\left(\sigma\sqrt{\frac{s\log^5(d)}{n}}\right).$$

More formal versions of this theorem appear as Theorem 3 and Corollary 4. Note that the test loss is within polylog factor to the minimax rate.

The model we use is similar to a 2-layer scalar network (which gives the $\boldsymbol{w}^{\odot 2} - \boldsymbol{u}^{\odot 2}$ term) with a skip-through connection (the $\boldsymbol{v}$ term) like in the ResNet (He et al., 2016). Intuitively, the term $\lambda(\boldsymbol{w}^{\odot 2} - \boldsymbol{u}^{\odot 2})$ promotes minimum ℓ_1 norm properties and can be used to fit the sparse signal $\boldsymbol{\beta}^*$, while the term $\boldsymbol{v}$ promotes minimum ℓ_2 norm properties and can be used to fit the noise.

Of course, showing that training this model via gradient descent leads to the correct trade-off between fitting the signal and noise is still challenging. We rely on dynamics analysis and show that the term $\lambda(\boldsymbol{w}^{\odot 2} - \boldsymbol{u}^{\odot 2})$ first grows fast to recover the sparse signal and then the term $\boldsymbol{v}$ grows to fit the noise. Interactions between all parameters $\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{u}$ makes it difficult to directly derive an accurate dynamics analysis. To address this issue, we introduce a new way to decompose $\boldsymbol{v}$ that allows us to separate the effect of learning signal and fitting noise and leads to a better characterization of the training dynamics. See details in Section 4 and Section 5.

1.2 Related works

There is a long line of work trying to understand implicit regularization effect, we refer the readers to some surveys for more complete discussions (Bartlett et al., 2021; Dar et al., 2021; Vardi, 2022). Here, we first summarize implicit regularization effect for interpolating linear models and their variants in regression setting. We then discuss related works for implicit regularization that are more related to training dynamic analysis instead of norm-minimizing.

Min-ℓ_2-norm interpolator When using linear model $\boldsymbol{\beta}^\top \boldsymbol{x}$ for regression $\frac{1}{2n}\sum_i(\boldsymbol{\beta}^\top \boldsymbol{x}_i - y_i)^2$, it is known that gradient flow/descent with 0 initialization will converge to the solution that minimizes its ℓ_2 norm (e.g., Gunasekar et al. (2018)). Recently, many papers have studied the generalization error of such min-ℓ_2-norm interpolator in the overparametrized regime where the dimension is much larger than then number of samples (Hastie et al., 2022; Bartlett et al., 2020; Tsigler and Bartlett, 2020; Belkin et al., 2020; Zhou et al., 2020; Negrea et al., 2020; Mitra, 2019; Koehler et al., 2021). In particular, these results suggest that min-ℓ_2-norm interpolator can achieve benign overfitting when the spectrum of input data covariance matrix has certain structure. On the other hand, it suffers from large test loss with isotropic features (identity covariance matrix for $\boldsymbol{x}$).

Min-ℓ_1-norm interpolator Going beyond the simplest linear model $\boldsymbol{\beta}^\top \boldsymbol{x}$, when the underlying signal is known to be sparse, one could reparametrize $\boldsymbol{\beta}$ by $\boldsymbol{\beta}(\boldsymbol{w}, \boldsymbol{u}) = \boldsymbol{w}^{\odot L} - \boldsymbol{u}^{\odot L}$, where $\odot L$ represents element-wise L-th power for integer $L \geq 2$. Woodworth et al. (2020); Azulay et al. (2021); Yun et al. (2021) showed that gradient flow with such parametrization converges to min-ℓ_1-norm solution when using small initialization and min-ℓ_2-norm solution when using large initialization. Researchers have studied the test loss of the min-ℓ_1-norm interpolator in the sparse noisy linear regression (Mitra, 2019; Ju et al., 2020; Li and Wei, 2021; Chinot et al., 2022; Koehler et al., 2021).

Lower bounds are also shown in Chatterji and Long (2022); Wang et al. (2022), which suggests that min-ℓ_1-norm interpolator does not have good generalization performance due to its sparsity. Vaskevicius et al. (2019); Li et al. (2021a) showed that gradient descent with early stopping can still achieve near-optimal test loss, but these results do not give interpolating models.

Hybrid Model Muthukumar et al. (2020) proposed an interpolation scheme called hybrid interpolation (Definition 5 in their paper) to achieve optimal test loss. Specifically, the hybrid interpolation is a 2-step procedure to achieve benign overfitting: (1) use any estimator to recover signal (e.g., Lasso (Bickel et al., 2009)); (2) use min-ℓ_2-norm interpolator to memorize the remaining noise. Such two-step procedure shares similarity with the learning dynamics in our analysis: our model will first recover the signal using the second-order term and then fit the noise using the linear term. Different from the hybrid interpolation scheme that requires a 2-step process, in our setup such learning dynamics arise naturally just by running gradient descent.

Beyond norm-minimization for implicit regularization Many of the earlier works for implicit regularization shows that the training process minimizes a certain norm (or maximizes margin with respect to a norm). The first example of implicit regularization that goes beyond norm minimization works in the setting of matrices. Arora et al. (2019) observed that for low-rank matrix problems the solution found does not always minimize the nuclear norm. Similar idea has also been exploited in the full-observation matrix sensing (Gidel et al., 2019; Gissin et al., 2020). Later Li et al. (2021b) was able to characterize the implicit regularization effect in matrix sensing problems via a greedy-low-rank-learning dynamics. Such implicit rank regularization and dynamics analysis are also studied in tensor problems (Razin et al., 2021, 2022; Ge et al., 2021) and neural networks (Timor et al., 2022; Frei et al., 2022). Our result shows that dynamics analysis can be important even in the simpler sparse regression model.

2 Preliminary

In this section we first introduce basic notations. Then we define the precise sparse recovery problem we are solving, and the learner model/algorithms we use. Finally we state several useful properties for the data that we will use throughout our analysis.

Notation Denote $[n] = \{1, 2, \ldots, n\}$. We use bold symbols to represent vectors and matrices. For vector $\boldsymbol{\beta} \in \mathbb{R}^d$, given any set $A \subseteq [d]$, let $\boldsymbol{\beta}_A := \sum_{i \in A} \beta_i \boldsymbol{e}_i$ be the same as $\boldsymbol{\beta}$ for the entries in set A and 0 for other entries, where $\{\boldsymbol{e}_i\}$ is the standard basis. We use standard big-O notations Ω, O to hide constants and $\widetilde{\Omega}, \widetilde{O}$ to hide constants and all logarithmic factors including $\log(d), \log(n), \log(1/\sigma)$. We will drop the time sub/superscripts when the context is clear.

Target function and data Suppose the ground-truth function is

$$f_*(\boldsymbol{x}) = \boldsymbol{\beta}^{*\top} \boldsymbol{x},$$

where $\boldsymbol{\beta}^* \in \mathbb{R}^d$ is s-sparse. Without loss of generality, we assume $|\beta_1^*| \geq \ldots \geq |\beta_s^*| > 0$ and $\beta_{s+1}^* = \ldots = \beta_d^* = 0$. Denote $S_+ := \{i : \beta_i^* > 0\}$ be the set of positive signal entries, $S_- := \{i : \beta_i^* < 0\}$ be the set of negative signal entries, and $S := S_+ \cup S_- = [s]$ be the set of all signal entries. We use $\boldsymbol{\beta}_S := \sum_{i:\beta_i^* \neq 0} \beta_i \boldsymbol{e}_i$

to be the vector that is same as $\boldsymbol{\beta}$ for the signal entries in S and 0 for other entries, and $\boldsymbol{\beta}_e := \sum_{i:\beta_i^*=0} \beta_i \boldsymbol{e}_i$ to be the vector that is the same as $\boldsymbol{\beta}$ for the non-signal entries that are not in S and 0 for other entries. We similarly define $\boldsymbol{\beta}_{S_+}, \boldsymbol{\beta}_{S_-}, \boldsymbol{\beta}_{e_+}, \boldsymbol{\beta}_{e_-}$. Let $\beta_{\max} := |\beta_1^*|$ be the maximum absolute value entry of $\boldsymbol{\beta}_S^*$ and $\beta_{\min} := |\beta_s^*|$ be the minimum absolute value entry of $\boldsymbol{\beta}_S^*$. We assume $\beta_{\min}, \beta_{\max} = \Theta(1)$ for simplicity. Our results can generalize to arbitrary $\beta_{\max}, \beta_{\min}$ with the cost of an additional polynomial dependency on them.

We generate n training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ by

$$y = f_*(\boldsymbol{x}) + \xi,$$

where $\boldsymbol{x}$ is the input data, $\xi \sim N(0, \sigma^2)$ is the label noise and y is the target. Denote the $n \times d$ matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n]^\top$ as the input data matrix, $\boldsymbol{y} = (y_1, \dots, y_n)^\top$ as the target vector and $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)^\top$ as the noise vector.

Learner model, loss and algorithm To learn the target function $f_*(\boldsymbol{x})$, we use the following model

$$f_{\boldsymbol{u},\boldsymbol{w},\boldsymbol{v}}(\boldsymbol{x}) = (\boldsymbol{v} + \lambda \boldsymbol{w}^{\odot 2} - \lambda \boldsymbol{u}^{\odot 2})^\top \boldsymbol{x}. \tag{1}$$

Here $\boldsymbol{w}^{\odot 2} := \boldsymbol{w} \odot \boldsymbol{w}$ and $\boldsymbol{u}^{\odot 2} := \boldsymbol{u} \odot \boldsymbol{u}$ is the element-wise square of $\boldsymbol{w}$ and $\boldsymbol{u}$. In general we use $\boldsymbol{u} \odot \boldsymbol{v}$ to denote the element-wise product of $\boldsymbol{u}$ and $\boldsymbol{v}$. Our model can be viewed as a linear model $\boldsymbol{\beta}^\top \boldsymbol{x}$ with reparametrization $\boldsymbol{\beta} = \boldsymbol{v} + \lambda \boldsymbol{w}^{\odot 2} - \lambda \boldsymbol{u}^{\odot 2}$. Such element-wise product reparametrization $\boldsymbol{w}^{\odot 2} - \boldsymbol{u}^{\odot 2}$ is common in the implicit bias literature (Woodworth et al., 2020; Azulay et al., 2021; Yun et al., 2021). In the view of neural networks, the learner model can also be viewed as a 2-layer diagonal linear network with a shortcut connection (He et al., 2016). For simplicity of notation, denote $\boldsymbol{\beta} = \boldsymbol{v} + \lambda \boldsymbol{w}^{\odot 2} - \lambda \boldsymbol{u}^{\odot 2}$. We are particular interested in the overparametrized regime $n \ll d$, where the model has the ability to overfit the data without learning the target $\boldsymbol{\beta}^*$.

Denote residual $r_i := f_{\boldsymbol{u},\boldsymbol{w},\boldsymbol{v}}(\boldsymbol{x}_i) - y_i$ for $i \in [n]$ and $\boldsymbol{r} := (r_1, \dots, r_n)^\top$. We will use gradient descent to minimize mean-square loss, that is

$$L(\boldsymbol{u}, \boldsymbol{w}, \boldsymbol{v}) := \frac{1}{2n} \sum_{i=1}^n \left(f_{\boldsymbol{u},\boldsymbol{w},\boldsymbol{v}}(\boldsymbol{x}_i) - y_i\right)^2.$$

The gradient for this loss is given below:

$$\begin{cases} \boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} - \eta \nabla_{\boldsymbol{w}} L(\boldsymbol{u}^{(t)}, \boldsymbol{w}^{(t)}, \boldsymbol{v}^{(t)}) = \boldsymbol{w}^{(t)} - \eta \left(\frac{1}{n} \boldsymbol{X}^\top \boldsymbol{r}^{(t)}\right) \odot (2\lambda \boldsymbol{w}^{(t)}) \\ \boldsymbol{u}^{(t+1)} = \boldsymbol{u}^{(t)} - \eta \nabla_{\boldsymbol{u}} L(\boldsymbol{u}^{(t)}, \boldsymbol{w}^{(t)}, \boldsymbol{v}^{(t)}) = \boldsymbol{u}^{(t)} + \eta \left(\frac{1}{n} \boldsymbol{X}^\top \boldsymbol{r}^{(t)}\right) \odot (2\lambda \boldsymbol{u}^{(t)}) \\ \boldsymbol{v}^{(t+1)} = \boldsymbol{v}^{(t)} - \eta \nabla_{\boldsymbol{v}} L(\boldsymbol{u}^{(t)}, \boldsymbol{w}^{(t)}, \boldsymbol{v}^{(t)}) = \boldsymbol{v}^{(t)} - \eta \frac{1}{n} \boldsymbol{X}^\top \boldsymbol{r}^{(t)}. \end{cases} \tag{2}$$

Properties the input data We use several key properties of the input data matrix $\boldsymbol{X}$ and noise $\boldsymbol{\xi}$. First is the classic notion of Restricted Isometry Property (RIP).

Definition 1 ((k, δ)-RIP)**.** *A $n \times d$ matrix $\boldsymbol{X}/\sqrt{n}$ is said to be (k, δ)-RIP if for any k-sparse vector $\boldsymbol{\beta}$ we have*

$$(1 - \delta) \|\boldsymbol{\beta}\|_2^2 \le \left\|\boldsymbol{X}\boldsymbol{\beta}/\sqrt{n}\right\|_2^2 \le (1 + \delta) \|\boldsymbol{\beta}\|_2^2.$$

We will assume data matrix $\boldsymbol{X}/\sqrt{n}$ satisfies $(s+1, \delta)$-RIP with $\delta = \widetilde{O}(1/(1 + n/\sqrt{d})s^{3/2})$ and some regularity conditions on $\boldsymbol{X}, \boldsymbol{\xi}$, as summarized in the Assumption 1 below. These conditions can be easily satisfied under some choice of $\boldsymbol{X}, \boldsymbol{\xi}$, as shown later in Lemma 2.

Assumption 1. *Input data matrix* $\boldsymbol{X}/\sqrt{n}$ *satisfies* $(s+1,\delta)$*-RIP with* $\delta \le c_\delta/(1+n/\sqrt{d\log d})s^{3/2}\log^3(d)$ *where* c_δ *is a small enough constant, and* $\boldsymbol{X},\boldsymbol{\xi}$ *satisfy the following regularity conditions:*

$$\|\boldsymbol{\xi}\|_2 = O(\sigma\sqrt{n}),$$
$$\left\|\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\xi}\right\|_\infty \le B_\xi := O\left(\sigma\sqrt{\frac{\log d}{n}}\right),$$
$$\left\|\boldsymbol{X}^\top\boldsymbol{\xi}\right\|_2 = O\left(\sigma\sqrt{dn}\right),$$
$$\left\|\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\beta}\right\|_\infty = O\left(\frac{\|\boldsymbol{\beta}\|_2}{\sqrt{n}}\right) \text{ for any vector } \boldsymbol{\beta},$$
$$(1-O(\sqrt{n/d}))d \le \lambda_{\min}(\boldsymbol{X}\boldsymbol{X}^\top) \le \lambda_{\max}(\boldsymbol{X}\boldsymbol{X}^\top) \le (1+O(\sqrt{n/d}))d.$$

Note that the notation $B_\xi = O(\sigma\sqrt{\log(d)/n})$ not only is for notation simplicity, but also intuitively stands for the best error in ℓ_∞ that one could hope with Gaussian noise. Indeed, Lounici et al. (2011) showed that the minimax optimal ℓ_∞ test error is $\Omega(\sigma\sqrt{\log(d/s)/n})$. Later in our analysis, we show the test loss is closely related with B_ξ.

When each entry of data matrix $\boldsymbol{X}$ is i.i.d. Gaussian and noise $\boldsymbol{\xi}$ is i.i.d. sampled from $N(\mathbf{0},\sigma^2\boldsymbol{I})$, all the conditions above are satisfied as long as $\widetilde{\Omega}(s^4) \le n \le \widetilde{O}(d/s^4)$. See Appendix A.1 for details.

Lemma 2. *Suppose* $\boldsymbol{X}$ *is a Gaussian random matrix and* $\boldsymbol{\xi} \sim N(\mathbf{0},\sigma^2\boldsymbol{I})$*. Then if* $\widetilde{\Omega}(s^4) \le n \le \widetilde{O}(d/s^4)$*, we have Assumption 1 is satisfied with probability at least* $1-1/d$*.*

3 Main Result

Our main result, formalized in the theorem below, shows that gradient descent on the learner model (1) achieves benign overfitting.

Theorem 3 (Main result)**.** *Under Assumption 1, suppose there exists constant* C *such that* $\sigma \le C$*. We train model* (1) *with initialization* $\boldsymbol{v}^{(0)} = \mathbf{0}$, $\boldsymbol{w}^{(0)} = \boldsymbol{u}^{(0)} = \alpha\mathbf{1}$ *and follow the gradient descent update* (2)*. If* $\widetilde{\Omega}(s) \le n \le \widetilde{O}\left(\min\{d/s, d^{2/3}\}\right)$ *and we choose* $\lambda = \Theta\left(d/\sigma n(\sqrt{\log(d)/n} + \sqrt{n/d})\log(n)\right)$, $\alpha = 1/\mathrm{poly}(d)$, $\eta \le O(\sqrt{n/sd}/\lambda^3)$*, then for every* $t \ge T = O(\log(n/\alpha\varepsilon)n/\eta d)$ *with any given* $\varepsilon > 0$ *we have training loss* $L(\boldsymbol{u}^{(t)},\boldsymbol{w}^{(t)},\boldsymbol{v}^{(t)}) \le \varepsilon$ *and test loss*

$$\left\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\right\|_2 = O\left(\sqrt{s}\log^2(d)\left(\sigma\sqrt{\frac{\log(d)}{n}} + \sigma\sqrt{\frac{n}{d}}\right)\right).$$

Note that the final test error depends on $\log(1/\alpha)$. Since we choose $\alpha = 1/\mathrm{poly}(d)$, it appears as $\log(d)$ in the final error bound. Also, the test loss does not depend on $1/\varepsilon$, so it remains small when ε is very close to 0.

For any interpolator $\boldsymbol{\beta}$, its test loss has lower bound $\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2 = \Omega(\sigma\sqrt{s\log(d/s)/n} + \sigma\sqrt{n/d})$ (Muthukumar et al., 2020), where $\sigma\sqrt{n/d}$ comes from the min-ℓ_2-norm interpolator that fits the noise. Thus, the above test loss is optimal up to $\mathrm{poly}(\log d, s)$ factors. The additional $\log d$, s dependencies in our result (and the fact that n cannot be larger than $d^{2/3}$) are due to technical difficulties in analyzing the dynamics. When $n = O(\sqrt{d\log d})$, the first term dominates the second term, and the above test loss becomes $O(\sigma\sqrt{s\log^5(d)/n})$. This is close to the minimax optimal rate $\Omega(\sigma\sqrt{s\log(d/s)/n})$ up to $\mathrm{polylog}(d)$ factors (Raskutti et al., 2011).

For the Gaussian data case ($\boldsymbol{x} \sim N(\mathbf{0},\boldsymbol{I})$), by Lemma 2 we in addition need $n = \tilde{\Omega}(s^4)$ to satisfy Assumption 1. This leads to the following corollary:

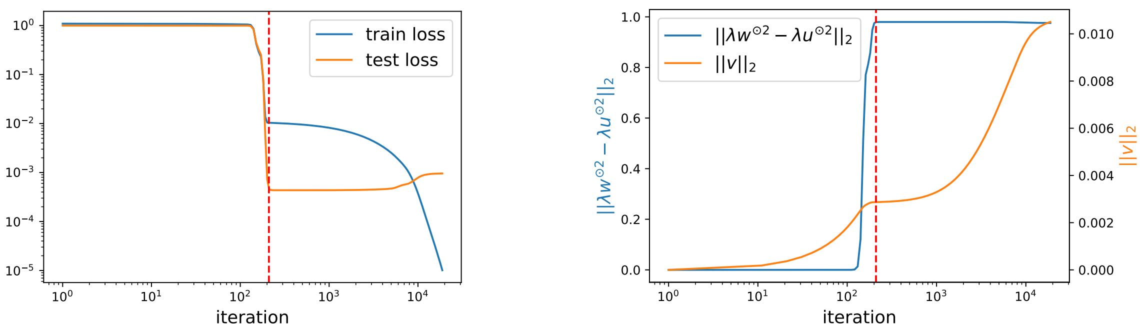

Figure 1: Training dynamics of model (1) following gradient descent update (2) under $d = 5 \times 10^4$, $n = 3\sqrt{d}$, $\sigma = 0.1$, $\boldsymbol{\beta}^* = (1/\sqrt{3}, -1/\sqrt{3}, 1/\sqrt{3}, 0, \ldots, 0)^\top$ and Gaussian data $\boldsymbol{x}_i \sim N(\mathbf{0}, \boldsymbol{I})$. We set $\lambda = 100d/\sigma n \log(n)(\sqrt{\log(d)/n} + \sqrt{n/d})$ and run gradient descent with $\eta = 10^{-6}$ from initialization $\alpha\mathbf{1}$ with $\alpha = 10^{-10}$ until training loss reaches 10^{-5}. Red vertical line stands for the transition between Stage 1 and Stage 2. **Left:** training loss L goes to 0 and test loss $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2$ remain small at the end. **Right:** norm of second-order term $\lambda(\boldsymbol{w}^{\odot 2} - \boldsymbol{u}^{\odot 2})$ grows large to recover the signal in Stage 1 and linear term $\boldsymbol{v}$ remain small during the training. Both x-axis are in log scale as Stage 1 is significantly shorter than Stage 2.

Corollary 4 (Near minimax rate)**.** *Under the setting of Theorem 3 and the choice of λ, α, η, suppose input data $\boldsymbol{X}$ is Gaussian matrix and noise $\boldsymbol{\xi} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$. If $\widetilde{\Omega}(s^4) \le n \le \widetilde{O}(\sqrt{d})$, then for every $t \ge T = O(\log(n/\alpha\varepsilon)n/\eta d)$ with any given $\varepsilon > 0$ we have training loss $L(\boldsymbol{u}^{(t)}, \boldsymbol{w}^{(t)}, \boldsymbol{v}^{(t)}) \le \varepsilon$ and test loss*

$$\left\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\right\|_2 = O\left(\sigma\sqrt{\frac{s\log^5(d)}{n}}\right),$$

which is near-optimal up to polylog(d) factors.

4 Intuitions for the Training Dynamics

Consider the training of our model (1) using gradient descent. Ideally, one would hope the training process to combine the advantages of min-ℓ_1-norm and min-ℓ_2-norm interpolator as done explicitly in Muthukumar et al. (2020): first use $\boldsymbol{w}^{\odot 2} - \boldsymbol{u}^{\odot 2}$ to learn the sparse target $\boldsymbol{\beta}^*$ and then use $\boldsymbol{v}$ to memorize the noise with small ℓ_2 norm. This would require us to fix $\boldsymbol{v} = 0$ when learning the signal and fix $\boldsymbol{w}^{\odot 2} - \boldsymbol{u}^{\odot 2}$ when fitting the noise. However, since training is done on all parameters simultaneously, it's unclear why it follows this ideal dynamics.

Stages of training At a higher level, we show that the actual training dynamics of parameters $\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{u}$ approximately follow the above ideal dynamics in 2 stages (Figure 1):

- In Stage 1, the linear term $\boldsymbol{v}$ remains small so that essentially the second-order term $\boldsymbol{w}^{\odot 2} - \boldsymbol{u}^{\odot 2}$ learns the signal using its bias towards sparse solution.
- In Stage 2, $\boldsymbol{v}$ moves to memorize the noise while $\boldsymbol{w}^{\odot 2} - \boldsymbol{u}^{\odot 2}$ roughly stays the same. Since $\boldsymbol{v}$ is biased towards small ℓ_2 norm, the final test loss remain small after interpolating the data.

However, things are not as simple when we examine the dynamics carefully. It turns out that even though $\boldsymbol{v}$ does not grow to be too large in Stage 1, it still becomes large enough so that existing analysis on $\boldsymbol{w}$ and $\boldsymbol{u}$ will no longer apply. To address this problem, we keep track of the dynamics of $\boldsymbol{v}$ very carefully throughout the training process. This is done through introducing the following decompositions of $\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{v}/n$ and $\boldsymbol{v}$.

Decompositions of $\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{v}/n$ and $\boldsymbol{v}$ To keep track of the dynamics of $\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{v}/n$ and $\boldsymbol{v}$, we first consider the *ideal* dynamics for $\boldsymbol{v}$. We hope $\boldsymbol{v}$ to fit the noise. If we were actually given the noise, we can use the loss function $\|\boldsymbol{X}\boldsymbol{v}-\boldsymbol{\xi}\|_2^2/2n$. Running gradient descent on this function gives a trajectory for $\boldsymbol{v}$, which can be computed explicitly. Our decomposition tries to highlight that the true trajectory of $\boldsymbol{v}$ is close to this ideal trajectory.

There are a few more issues that we need to work with. First, for simplicity, in the ideal trajectory we approximate $\boldsymbol{X}\boldsymbol{X}^\top$ by $d\boldsymbol{I}$ (which is accurate as long as $d \gg n$). Second, because of the signal, the entries of $\boldsymbol{v}$ in S may deviate significantly and in fact contribute a little bit to the fitting of the signal.

Based on these observations, we decompose both $\boldsymbol{v}$ and $\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{v}$ into three terms – a signal term, a noise-fitting term and an approximation error term. They are defined in the following equations:

$$\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{v}^{(t)} := \frac{d}{n}\boldsymbol{v}_S^{(t)} + b_t(\boldsymbol{X}^\top\boldsymbol{\xi})_e + \boldsymbol{\Gamma}_t, \tag{3}$$

$$\boldsymbol{v}^{(t)} := \boldsymbol{v}_S^{(t)} + a_t\boldsymbol{X}^\top\boldsymbol{\xi} + \boldsymbol{\Delta}_v^{(t)}, \tag{4}$$

where

$$b_{t+1} := b_t - \frac{\eta d}{n}\left(b_t - \frac{1}{n}\right),$$

$$a_{t+1} := a_t - \eta\left(b_t - \frac{1}{n}\right),$$

Here $\left\|\boldsymbol{\Gamma}^{(t)}\right\|_\infty \le \gamma_t$, $\left\|\boldsymbol{\Delta}_v^{(t)}\right\|_\infty \le \zeta_t$ give ℓ_∞-norm bounds on the approximation error. Also recall the notation $\boldsymbol{\beta}_S = \sum_{i:\beta_i^*\neq 0}\beta_i\boldsymbol{e}_i$, $\boldsymbol{\beta}_e = \sum_{i:\beta_i^*=0}\beta_i\boldsymbol{e}_i$. Intuitively in the decomposition of $\boldsymbol{v}$, $\boldsymbol{v}_S$ part tries to fit the signal, $\boldsymbol{X}^\top\boldsymbol{\xi}$ part tries to fit the noise and the remaining term is approximation error (the decomposition of $\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}$ has the same structure). We will show in our analysis that $\boldsymbol{v}_S$ contributes little for learning the signal while $\boldsymbol{X}^\top\boldsymbol{\xi}$ fits all the noise and approximation errors remain small.

The recursions of a_t and b_t are exactly the dynamics of $\boldsymbol{v}$ in the ideal setting, where we fit $\|\boldsymbol{X}\boldsymbol{v}-\boldsymbol{\xi}\|_2^2/2n$ (and approximate $\boldsymbol{X}\boldsymbol{X}^\top$ by $d\boldsymbol{I}$).

Finally, the d/n factor appears in front of $\boldsymbol{v}_S$ in the decomposition of $\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}/n$. This is because in the ideal setting (approximate $\boldsymbol{X}\boldsymbol{X}^\top$ by $d\boldsymbol{I}$) the change of $\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{v}/n$ is d/n times larger than the change of $\boldsymbol{v}$. One can then use simple calculations to show that the signal part $(\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}/n)_S$ corresponds to $(d/n)\boldsymbol{v}_S$. The non-signal part has the same factor but the ℓ_∞ norm there is small and hence bundled into the approximation error term.

5 Proof Sketch

In this section, we give the proof sketch of our main result Theorem 3 with several key proof ideas. We first combine the tools we discussed in Section 2 and the decomposition of $\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}/n$ and $\boldsymbol{v}$ defined in Section 4 to give the approximation of gradient. Then, we give the proof sketch of Stage 1 and Stage 2 in Section 5.1 and Section 5.2 respectively.

Approximation of gradient Given that $\boldsymbol{X}$ is a $(s+1,\delta)$-RIP matrix, the following lemma gives useful approximation that allows us to approximate the gradient in Lemma 6. The proof is a standard consequence of RIP property, which is deferred to Appendix A.2.

Lemma 5. *Given $n\times d$ matrix $\boldsymbol{X}/\sqrt{n}$ satisfying $(k+1,\delta)$-RIP, for any $\boldsymbol{\beta}\in\mathbb{R}^d$, let $\boldsymbol{\Delta} = \left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}-\boldsymbol{I}\right)\boldsymbol{\beta}$, then the following hold:*

- *If $\boldsymbol{\beta}$ is k-sparse, then $\|\boldsymbol{\Delta}\|_\infty \le \sqrt{k}\delta\|\boldsymbol{\beta}\|_2$.*
- *For any vector $\boldsymbol{\beta}$, we have $\|\boldsymbol{\Delta}\|_\infty \le \delta\|\boldsymbol{\beta}\|_1$.*

The following lemma gives the approximation of the gradient. For the gradient of $\boldsymbol{w}, \boldsymbol{u}$, it would become the same as the gradient on the population loss $\left\|\lambda \boldsymbol{w}^{\odot 2} - \lambda \boldsymbol{u}^{\odot 2} - \boldsymbol{\beta}^*\right\|_2^2/2$ if $(d/n)\boldsymbol{v}_S$ and $\boldsymbol{\Delta}_r$ are small. In particular, this suggests that the second-order term $\lambda \boldsymbol{w}^{\odot 2} - \lambda \boldsymbol{u}^{\odot 2}$ will learn the target when $\boldsymbol{v}$ remains small.

Lemma 6 (Gradient approximation)**.** *Under Assumption 1, we have the following gradients and their useful approximation:*

$$\begin{aligned}
\nabla_{\boldsymbol{w}} L &= \left(\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{r}\right) \odot (2\lambda \boldsymbol{w}) = 2\lambda \left(\frac{d}{n}\boldsymbol{v}_S + \lambda \boldsymbol{w}_{S_+}^{\odot 2} - \lambda \boldsymbol{u}_{S_-}^{\odot 2} - \boldsymbol{\beta}^* + \boldsymbol{\Delta}_r\right) \odot \boldsymbol{w},\\
\nabla_{\boldsymbol{u}} L &= -\left(\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{r}\right) \odot (2\lambda \boldsymbol{u}) = -2\lambda \left(\frac{d}{n}\boldsymbol{v}_S + \lambda \boldsymbol{w}_{S_+}^{\odot 2} - \lambda \boldsymbol{u}_{S_-}^{\odot 2} - \boldsymbol{\beta}^* + \boldsymbol{\Delta}_r\right) \odot \boldsymbol{u},\\
\nabla_{\boldsymbol{v}} L &= \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{r} = \frac{d}{n}\boldsymbol{v}_S + \lambda \boldsymbol{w}_{S_+}^{\odot 2} - \lambda \boldsymbol{u}_{S_-}^{\odot 2} - \boldsymbol{\beta}^* + \boldsymbol{\Delta}_r,
\end{aligned}$$

where

$$\begin{aligned}
\|\boldsymbol{\Delta}_r\|_\infty =& O\left((1 + |nb-1|)\, B_\xi\right) + \sqrt{s}\delta \frac{d}{n}\|\boldsymbol{v}_S\|_2 + s\delta \left\|\frac{d}{n}\boldsymbol{v}_S + \lambda \boldsymbol{w}_{S_+}^{\odot 2} - \lambda \boldsymbol{u}_{S_-}^{\odot 2} - \boldsymbol{\beta}^*\right\|_\infty\\
&+ O\left(\frac{d}{\sqrt{n}}\lambda\right)\left(\left\|\boldsymbol{w}_{e_+}\right\|_\infty^2 + \left\|\boldsymbol{u}_{e_-}\right\|_\infty^2\right) + \gamma,
\end{aligned}$$

b and $\|\boldsymbol{\Gamma}\|_\infty \le \gamma$ are defined in (3), and recall S_+, S_- are the set of positive and negative entries of $\boldsymbol{\beta}^*$ and $e_+ = [d] \setminus S_+$, $e_- = [d] \setminus S_-$ are the corresponding complement set.

Note that the factor d/n in front of $\boldsymbol{v}_S$ naturally arises when we using the decomposition of $\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{v}/n$ in (3). This suggests that the actual part to fit the signal $\boldsymbol{\beta}^*$ is $(d/n)\boldsymbol{v}_S + \lambda \boldsymbol{w}_S^{\odot 2} - \lambda \boldsymbol{u}_S^{\odot 2}$, instead of the naïve $\boldsymbol{v}_S + \lambda \boldsymbol{w}_S^{\odot 2} - \lambda \boldsymbol{u}_S^{\odot 2}$ from the form of learner model. On the other hand, since $\boldsymbol{v}_S$ remains small, it does not affect the final test error because they are all close to $\lambda \boldsymbol{w}_S^{\odot 2} - \lambda \boldsymbol{u}_S^{\odot 2}$.

The forms of gradients highlight the difference between the parametrization $\boldsymbol{v}$ and $\boldsymbol{w}^{\odot 2} - \boldsymbol{u}^{\odot 2}$. For each coordinate, w_i (same for u_i) moves according to $w_i \leftarrow (1+\eta\lambda_i)w_i$ for some growth rate λ_i, which would grow exponential fast when $\lambda_i > 0$. However, the gradient for $\boldsymbol{v}$ is not proportional to $\boldsymbol{v}$, so it only grows linearly with time. Such difference allows us to control the order of learning dynamics ($\boldsymbol{v}$ or $\boldsymbol{w}^{\odot 2} - \boldsymbol{u}^{\odot 2}$ grows up first). Thus, we could have the desired 2-stage learning dynamics by properly choosing the growth rate λ.

5.1 Stage 1: learning the signal

In Stage 1, our goal is to show that the linear term $\boldsymbol{v}$ will be characterized by the decompositions (3)(4), and the second-order term $\boldsymbol{w}^{\odot 2}, \boldsymbol{u}^{\odot 2}$ will recover the signal $\boldsymbol{\beta}^*$.

The following lemma gives the ending criteria for Stage 1. We can see only the signal entries $\boldsymbol{w}_{S_+}, \boldsymbol{u}_{S_-}$ grow large to recover $\boldsymbol{\beta}^*$ and others such as non-signal entries $\boldsymbol{w}_{e_+}, \boldsymbol{u}_{e_-}$ and linear term $\boldsymbol{v}$ are remain small. Also, the loss reduces to $O(\sigma\sqrt{n})$, which is essentially the norm of noise $\|\boldsymbol{\xi}\|_2$. The detailed proof is deferred to Appendix B.

Lemma 7 (Stage 1)**.** *Let C_1 be a large enough universal constant, denote*

$$T_1 := \inf\left\{t : \left\|\frac{d}{n}\boldsymbol{v}_S^{(T_1)} + \lambda \boldsymbol{w}_{S_+}^{(T_1)\odot 2} - \lambda \boldsymbol{u}_{S_-}^{(T_1)\odot 2} - \boldsymbol{\beta}^*\right\|_\infty = C_1(B_\xi + \sigma\sqrt{n/d})\right\}.$$

Then we know $T_1 = O(\log(1/\alpha)/\eta\lambda)$ and the following hold:

- $\left\|\boldsymbol{w}_{e_+}^{(T_1)}\right\|_\infty, \left\|\boldsymbol{u}_{e_-}^{(T_1)}\right\|_\infty = O(\alpha)$.
- $\left\|\boldsymbol{v}_S^{(T_1)}\right\|_2 = O(\sqrt{s}(n/d)\log^2(d)(B_\xi + \sigma\sqrt{n/d}))$ *and* $\left\|\boldsymbol{v}^{(T_1)}\right\|_2 = O(\sigma\sqrt{n/d})$.

- $\left\|\boldsymbol{r}^{(T_1)}\right\|_2 = O(\sigma\sqrt{n})$.

Recall B_ξ is the target infinity norm error for recovering the entries in $\boldsymbol{\beta}^*$, when $d \gg n$, $\frac{d}{n}\boldsymbol{v}_S+\lambda\boldsymbol{w}_{S_+}^{\odot 2}-\lambda\boldsymbol{u}_{S_-}^{\odot 2}$ achieves this error at the end of Stage 1. We focus on this term instead of $\boldsymbol{v}_S+\lambda\boldsymbol{w}_{S_+}^{\odot 2}-\lambda\boldsymbol{u}_{S_-}^{\odot 2}$ due to its connection with the gradient shown in Lemma 6. Given that $\boldsymbol{v}_S$ is small, these two terms are in fact roughly the same.

As we discussed, a key step in the analysis is to characterize each term in the decomposition of $\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}/n$ and $\boldsymbol{v}$, which would imply that $\boldsymbol{v}$ remains small in Stage 1. This is formalized in the following lemma.

Lemma 8 (Informal)**.** *Consider the decomposition of $\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}/n$ and $\boldsymbol{v}$ in* (3) (4)*, we have for $t \le O(\log(1/\alpha/\eta\lambda))$*

$$\begin{aligned} b_t &= (1-(1-\eta d/n)^t)/n \le 1/n,\\ a_t &= (1-(1-\eta d/n)^t)/d \le 1/d\\ \|\boldsymbol{\Gamma}_t\|_\infty &\le \gamma_t = O(\sigma\sqrt{n/d}+B_\xi),\\ \|\boldsymbol{\Delta}_v\|_\infty &\le \zeta_t = O(\sigma\sqrt{n}/d). \end{aligned}$$

Note that $\boldsymbol{v}$ will memorize the noise when $b_t = 1/n$ and $a_t = 1/d$ as $\boldsymbol{X}\boldsymbol{v}^{(t)} \approx \boldsymbol{X}(a_t\boldsymbol{X}^\top\boldsymbol{\xi}) \approx \boldsymbol{\xi}$. However, since $T_1 = \widetilde{O}(\eta\lambda) = o(n/\eta d)$, we know $a_t = o(1/d)$ in Stage 1. This shows that $\boldsymbol{v}$ is still small and does not yet interpolate the noise part.

Combine the above lemma with Lemma 6, we have the following gradient approximation

$$\begin{aligned} \nabla_{\boldsymbol{w}} L &= \left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{r}\right)\odot(2\lambda\boldsymbol{w}) = 2\lambda(\frac{d}{n}\boldsymbol{v}_S+\lambda\boldsymbol{w}_S^{\odot 2}-\lambda\boldsymbol{u}_S^{\odot 2}-\boldsymbol{\beta}^*+\boldsymbol{\Delta}_r)\odot\boldsymbol{w},\\ \nabla_{\boldsymbol{u}} L &= -\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{r}\right)\odot(2\lambda\boldsymbol{u}) = -2\lambda(\frac{d}{n}\boldsymbol{v}_S+\lambda\boldsymbol{w}_S^{\odot 2}-\lambda\boldsymbol{u}_S^{\odot 2}-\boldsymbol{\beta}^*+\boldsymbol{\Delta}_r)\odot\boldsymbol{u}, \end{aligned}$$

where

$$\|\boldsymbol{\Delta}_r\|_\infty = O(B_\xi+\sigma\sqrt{n/d}) + s\delta\left\|\frac{d}{n}\boldsymbol{v}_S+\lambda\boldsymbol{w}_{S_+}^{\odot 2}-\lambda\boldsymbol{u}_{S_-}^{\odot 2}-\boldsymbol{\beta}^*\right\|_\infty.$$

Intuitively, this suggests if a coordinate of the residual $\frac{d}{n}\boldsymbol{v}_S+\lambda\boldsymbol{w}_S^{\odot 2}-\lambda\boldsymbol{u}_S^{\odot 2}-\boldsymbol{\beta}^*$ has large absolute value, then one of $\boldsymbol{w}$ or $\boldsymbol{u}$ will grow exponentially depending on the sign of the residual.

Given such gradient approximation, our goal is to show that $\boldsymbol{v}_S$ and $\boldsymbol{\Delta}_r$ remain small so that $\boldsymbol{w}$ and $\boldsymbol{u}$ essentially follow the gradient on population loss $\left\|\lambda\boldsymbol{w}^{\odot 2}-\lambda\boldsymbol{u}^{\odot 2}-\boldsymbol{\beta}^*\right\|_2^2/2$ to recover the target $\boldsymbol{\beta}^*$.

In the simplest case of $s=1$, we can see that whenever the signal error $|(d/n)v_1+\lambda w_1^2-\lambda u_1^2-\beta_1^*| \ge O(B_\xi+\sigma\sqrt{n/d})$ is still large, it leads to a large gradient for either u_1 or w_1, which in turn decreases the error. Therefore, at the end the error will decrease to $O(B_\xi+\sigma\sqrt{n/d})$. In fact, due to the parameterization of $\boldsymbol{w}^{\odot 2},\boldsymbol{u}^{\odot 2}$, their growing rate would be exponential so they will grow up fast to recover the signal.

At the same time, we can control the growth of v_1 by choosing a large enough λ to ensure the length of Stage 1 T_1 is short. The non-signal entries $\boldsymbol{w}_{e_+},\boldsymbol{u}_{e_-}$ will also remain almost as small as their initialization, as their growth rate is much smaller compared with the signal entries.

For higher sparsity s, the analysis becomes significantly more complicated because of the signal error term $\left\|\frac{d}{n}\boldsymbol{v}_S+\lambda\boldsymbol{w}_{S_+}^{\odot 2}-\lambda\boldsymbol{u}_{S_-}^{\odot 2}-\boldsymbol{\beta}^*\right\|_\infty$ in $\|\boldsymbol{\Delta}_r\|_\infty$. Not all the entries of $\boldsymbol{\beta}^*$ are of the same size, which results in different growth rates in the entries of $\boldsymbol{w}$ and $\boldsymbol{u}$. The entries with larger β_i^* will be learned faster than the smaller ones, which could lead to the case where $\left\|\frac{d}{n}\boldsymbol{v}_S+\lambda\boldsymbol{w}_{S_+}^{\odot 2}-\lambda\boldsymbol{u}_{S_-}^{\odot 2}-\boldsymbol{\beta}^*\right\|_\infty$ is much larger than the error for a particular entry $k \in S$ of $(\frac{d}{n}\boldsymbol{v}_S+\lambda\boldsymbol{w}_{S_+}^{\odot 2}-\lambda\boldsymbol{u}_{S_-}^{\odot 2}-\boldsymbol{\beta}^*)_k$.

To deal with such issue, we show the following lemma that bound the time for reducing the signal error by half. Similar result was shown in Vaskevicius et al. (2019) where they do not have the linear term $\boldsymbol{v}$. The proof relies on the observation from the gradient approximation above that the signal error will monotone decrease before reaching $\|\boldsymbol{\Delta}_r\|_\infty$, and is made possible by the decomposition of $\boldsymbol{v}$.

Lemma 9 (Informal)**.** *Given any time t_0, assume $\left\|\frac{d}{n}\boldsymbol{v}_S^{(t_0)} + \lambda(\boldsymbol{w}_{S_+}^{(t_0)})^2 - \lambda(\boldsymbol{u}_{S_-}^{(t_0)})^2 - \boldsymbol{\beta}^*\right\|_\infty \geq \Omega(B_\xi + \sigma\sqrt{n/d})$. Let*

$$T' := \inf\left\{t - t_0 \geq 0 : \left\|\frac{d}{n}\boldsymbol{v}_S^{(t)} + \lambda(\boldsymbol{w}_{S_+}^{(t)})^2 - \lambda(\boldsymbol{u}_{S_-}^{(t)})^2 - \boldsymbol{\beta}^*\right\|_\infty \leq \left\|\frac{d}{n}\boldsymbol{v}_S^{(t_0)} + \lambda(\boldsymbol{w}_{S_+}^{(t_0)})^2 - \lambda(\boldsymbol{u}_{S_-}^{(t_0)})^2 - \boldsymbol{\beta}^*\right\|_\infty /2\right\}$$

be the time that signal error reduces by half. Then, we know $T' = O(1/\eta\lambda)$.

Repeatedly using the above lemma, we know it takes $T_1 = \widetilde{O}(1/\eta\lambda)$ time to reach the desired accuracy. Other claims follow directly. Detailed proofs are deferred to Appendix B.

5.2 Stage 2: memorizing the noise

Given that in Stage 1 we know $\lambda\boldsymbol{w}^{\odot 2} - \lambda\boldsymbol{u}^{\odot 2}$ has already recovered signal $\boldsymbol{\beta}^*$, in Stage 2 we show that the remaining noise will be memorized by the linear term $\boldsymbol{v}$ without increasing the test loss by too much. This allows us to recover the ground-truth $\boldsymbol{\beta}^*$ despite interpolating the data to ε training error, as formalized in the following lemma. The proof is deferred to Appendix C.

Lemma 10 (Stage 2)**.** *Let $T_2 := \inf\{t \geq 0 : L(\boldsymbol{w}^{(t)}, \boldsymbol{u}^{(t)}, \boldsymbol{v}^{(t)}) \leq \varepsilon\}$. Then, we have the length of Stage 2 is $T_2 - T_1 = O((n/\eta d)\log(n/\varepsilon))$ and the following hold for every $t \geq T_2$:*

- $\left\|\frac{d}{n}\boldsymbol{v}_S^{(t)} + \lambda\boldsymbol{w}_{S_+}^{(t)\odot 2} - \lambda\boldsymbol{u}_{S_-}^{(t)\odot 2} - \boldsymbol{\beta}^*\right\|_\infty = O(B_\xi + \sigma\sqrt{n/d})$
- $\left\|\boldsymbol{w}_{e_+}^{(t)}\right\|_\infty, \left\|\boldsymbol{u}_{e_-}^{(t)}\right\|_\infty = O(\alpha)$.
- $\left\|\boldsymbol{v}_S^{(t)}\right\|_2 = O(\sqrt{s}(n/d)\log^2(d)(B_\xi + \sigma\sqrt{n/d}))$ *and* $\left\|\boldsymbol{v}^{(t)}\right\|_2 = O(\sigma\sqrt{n/d})$.

Similar as in Stage 1, we still need to characterize each term in the decomposition of $\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}/n$ and $\boldsymbol{v}$.

Lemma 11 (Informal)**.** *Consider the decomposition of $\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}/n$ and $\boldsymbol{v}$ in* (3) (4)*, we have for $t \leq O((n/\eta d)\log(n/\varepsilon))$*

$$\begin{aligned}
b_t &= (1 - (1 - \eta d/n)^t)/n \leq 1/n, \\
a_t &= (1 - (1 - \eta d/n)^t)/d \leq 1/d \\
\|\boldsymbol{\Gamma}_t\|_\infty &\leq \gamma_t = O(\sigma\sqrt{n/d} + B_\xi), \\
\|\boldsymbol{\Delta}_v\|_\infty &\leq \zeta_t = O((B_\xi + \sigma\sqrt{n/d})n\log(n)/d).
\end{aligned}$$

Unlike in Stage 1, the signal has mostly been fitted in Stage 2. This makes the gradient smaller and the time it takes for Stage 2 ($T_2 - T_1 = O((n/\eta d)\log(n/\varepsilon))$) is much longer than Stage 1. Because of this longer time, we now have $b_t \approx 1/n$, $a_t \approx 1/d$ at the end of Stage 2. This implies that we essentially use linear term $\boldsymbol{v}$ to interpolate the noise as $\boldsymbol{X}\boldsymbol{v}^{(t)} \approx \boldsymbol{X}(a_t\boldsymbol{X}^\top\boldsymbol{\xi}) \approx \boldsymbol{\xi}$.

In the analysis of Stage 2, we have two major goals that are closely related: first, we want non-signal entries of $\boldsymbol{w}$, $\boldsymbol{u}$ to stay small; second, we want the residual $\|\boldsymbol{r}\|_2$ to decrease exponentially.

For $\boldsymbol{w}$, $\boldsymbol{u}$, combine the above lemma with Lemma 6, we know

$$\begin{aligned}
\nabla_{\boldsymbol{w}} L &= \left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{r}\right) \odot (2\lambda\boldsymbol{w}), \\
\nabla_{\boldsymbol{u}} L &= -\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{r}\right) \odot (2\lambda\boldsymbol{u}),
\end{aligned}$$

where

$$\left\|\frac{1}{n}\boldsymbol{X}^{\top}\boldsymbol{r}\right\|_{\infty} = \left\|\frac{d}{n}\boldsymbol{v}_S + \lambda\boldsymbol{w}_S^{\odot 2} - \lambda\boldsymbol{u}_S^{\odot 2} - \boldsymbol{\beta}^* + \boldsymbol{\Delta}_r\right\|_{\infty} = O(B_\xi + \sigma\sqrt{n/d}).$$

The infinity norm bound on $\frac{1}{n}\boldsymbol{X}^{\top}\boldsymbol{r}$ follows from a case analysis for signal and non-signal entries. For the signal entries, using the above gradient approximation similar as in Stage 1, we can show that the signal error $\left\|\frac{d}{n}\boldsymbol{v}_S + \lambda\boldsymbol{w}_S^{\odot 2} - \lambda\boldsymbol{u}_S^{\odot 2} - \boldsymbol{\beta}^* + \boldsymbol{\Delta}_r\right\|_{\infty}$ remains $O(B_\xi + \sigma\sqrt{n/d})$. For the non-signal entries $\boldsymbol{w}_{e_+}, \boldsymbol{u}_{e_-}$, we know its exponential growth rate is $O(\lambda(B_\xi + \sigma\sqrt{n/d}))$ from the gradient approximation.

The bound on $\left\|\frac{1}{n}\boldsymbol{X}^{\top}\boldsymbol{r}\right\|_{\infty}$ limits the movement of $\boldsymbol{u}$ and $\boldsymbol{w}$. As long as $O(\eta\lambda(B_\xi+\sigma\sqrt{n/d})(T_2-T_1)) < 1$, the non-signal part of $\boldsymbol{u}$ and $\boldsymbol{w}$ will remain small.

On the other hand, for the decrease rate of $\|\boldsymbol{r}\|_2$, the standard approach is to use ideas from Neural Tangent Kernel (NTK) (Jacot et al., 2018; Du et al., 2019; Allen-Zhu et al., 2019), and approximate the dynamics of $\boldsymbol{r}$ as $\boldsymbol{r}^{(t+1)} = (\boldsymbol{I} - \eta\boldsymbol{H}^{(t)})\boldsymbol{r}^{(t)}$ where $\boldsymbol{H}^{(t)}$ is the neural tangent kernel. The decreasing rate of $\|\boldsymbol{r}\|_2$ can then be bounded by lowerbounding the minimum eigenvalue of $\boldsymbol{H}^{(t)}$. However, bounding $\boldsymbol{H}^{(t)}$ naïvely by its distance to some initial $\boldsymbol{H}^{(t)}$ does not work in our case.

To fix this problem, we again rely on the dynamics of $\boldsymbol{v}$. Lemma 11 suggests that $\boldsymbol{v}^{(t)}$ gets close to $\boldsymbol{X}^{\top}\boldsymbol{\xi}$ with a rate of $\Omega(d/n)$ (this can also be viewed as the minimum eigenvalue of the NTK kernel restricted to $\boldsymbol{v}$). This convergence rate gives a bound on the length of $T_2 - T_1$, which allow us to choose an appropriate λ to keep $\boldsymbol{w}_{e_+}, \boldsymbol{u}_{e_-}$ small.

Once we have the bounds for the convergence rate and non-signal entries of $\boldsymbol{u}$,$\boldsymbol{w}$, other claims follow directly. Details are deferred to Appendix C.

Note that in the argument above, since the length of Stage 2 T_2-T_1 is proportional to $\log(1/\varepsilon)$, it cannot be used when ϵ is very close to 0 as λ is proportional to $1/(T_2 - T_1)$ and would become very small. In fact, we can get rid of the dependency on $\log(1/\varepsilon)$ with a more careful analysis. In the actual proof, we have two sub-stages for Stage 2, which uses different ways to bound the growth rate $\left\|\boldsymbol{X}^{\top}\boldsymbol{r}/n\right\|_{\infty}$. For Stage 2.1, we use the argument above until $\|\boldsymbol{r}\|_2 = O(\sigma)$. In Stage 2.2, given the training loss is already small enough, we use a NTK-type analysis to bound $\left\|\boldsymbol{X}^{\top}\boldsymbol{r}/n\right\|_{\infty} = (1-\Omega(\eta d/n))^{t-T_1}O(\sigma/\sqrt{n})$ as $\|\boldsymbol{r}\|_2$ decreases with rate $\Omega(d/n)$. See Appendix C for details.

6 Conclusion

In this paper, we give a new parametrization for the sparse linear regression problem, and showed that gradient descent for this new parametrization can learn an interpolator with near-optimal test loss. This highlights the importance of choosing the correct parametrization, especially the role of linear terms in fitting noise. Our proof is based on a new dynamic analysis that shows it is possible to first learn the features, and then fit the noise using an NTK-like process. We suspect similar training dynamics may apply to more complicated problems such as low-rank matrix factorization or neural networks.

Acknowledgement

This work is supported by NSF Award DMS-2031849, CCF-1845171 (CAREER), CCF-1934964 (Tripods) and a Sloan Research Fellowship.

References

Advani, M. S., Saxe, A. M., and Sompolinsky, H. (2020). High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446.

Allen-Zhu, Z., Li, Y., and Song, Z. (2019). A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252. PMLR.

Arora, S., Cohen, N., Hu, W., and Luo, Y. (2019). Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32.

Azulay, S., Moroshko, E., Nacson, M. S., Woodworth, B. E., Srebro, N., Globerson, A., and Soudry, D. (2021). On the implicit bias of initialization shape: Beyond infinitesimal mirror descent. In *International Conference on Machine Learning*, pages 468–477. PMLR.

Baraniuk, R., Davenport, M., DeVore, R., and Wakin, M. (2008). A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070.

Bartlett, P. L., Montanari, A., and Rakhlin, A. (2021). Deep learning: a statistical viewpoint. *Acta numerica*, 30:87–201.

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

Belkin, M., Hsu, D., and Xu, J. (2020). Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of statistics*, 37(4):1705–1732.

Chatterji, N. S. and Long, P. M. (2022). Foolish crowds support benign overfitting. *Journal of Machine Learning Research*, 23(125):1–12.

Chinot, G., Löffler, M., and van de Geer, S. (2022). On the robustness of minimum norm interpolators and regularized empirical risk minimizers. *The Annals of Statistics*, 50(4).

Dar, Y., Muthukumar, V., and Baraniuk, R. G. (2021). A farewell to the bias-variance tradeoff? an overview of the theory of overparameterized machine learning. *arXiv preprint arXiv:2109.02355*.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. (2019). Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*.

Frei, S., Vardi, G., Bartlett, P. L., Srebro, N., and Hu, W. (2022). Implicit bias in leaky relu networks trained on high-dimensional data. *arXiv preprint arXiv:2210.07082*.

Ge, R., Ren, Y., Wang, X., and Zhou, M. (2021). Understanding deflation process in over-parametrized tensor decomposition. *Advances in Neural Information Processing Systems*, 34:1299–1311.

Gidel, G., Bach, F., and Lacoste-Julien, S. (2019). Implicit regularization of discrete gradient dynamics in linear neural networks. *Advances in Neural Information Processing Systems*, 32.

Gissin, D., Shalev-Shwartz, S., and Daniely, A. (2020). The implicit bias of depth: How incremental learning drives generalization. In *International Conference on Learning Representations*.

Gunasekar, S., Lee, J., Soudry, D., and Srebro, N. (2018). Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841. PMLR.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31.

Ju, P., Lin, X., and Liu, J. (2020). Overfitting can be harmless for basis pursuit, but only to a degree. *Advances in Neural Information Processing Systems*, 33:7956–7967.

Koehler, F., Zhou, L., Sutherland, D. J., and Srebro, N. (2021). Uniform convergence of interpolators: Gaussian width, norm bounds and benign overfitting. *Advances in Neural Information Processing Systems*, 34:20657–20668.

Li, J., Nguyen, T., Hegde, C., and Wong, K. W. (2021a). Implicit sparse regularization: The impact of depth and early stopping. *Advances in Neural Information Processing Systems*, 34:28298–28309.

Li, Y. and Wei, Y. (2021). Minimum ℓ_1-norm interpolators: Precise asymptotics and multiple descent. *arXiv preprint arXiv:2110.09502*.

Li, Z., Luo, Y., and Lyu, K. (2021b). Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. In *International Conference on Learning Representations*.

Lounici, K., Pontil, M., Van De Geer, S., and Tsybakov, A. B. (2011). Oracle inequalities and optimal inference under group sparsity. *The annals of statistics*, 39(4):2164–2204.

Mitra, P. P. (2019). Understanding overfitting peaks in generalization error: Analytical risk curves for ℓ_2 and ℓ_1 penalized interpolation. *arXiv preprint arXiv:1906.03667*.

Muthukumar, V., Vodrahalli, K., Subramanian, V., and Sahai, A. (2020). Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1):67–83.

Negrea, J., Dziugaite, G. K., and Roy, D. (2020). In defense of uniform convergence: Generalization via derandomization with an application to interpolating predictors. In *International Conference on Machine Learning*, pages 7263–7272. PMLR.

Raskutti, G., Wainwright, M. J., and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE transactions on information theory*, 57(10):6976–6994.

Razin, N., Maman, A., and Cohen, N. (2021). Implicit regularization in tensor factorization. In *International Conference on Machine Learning*, pages 8913–8924. PMLR.

Razin, N., Maman, A., and Cohen, N. (2022). Implicit regularization in hierarchical tensor factorization and deep convolutional neural networks. *arXiv preprint arXiv:2201.11729*.

Timor, N., Vardi, G., and Shamir, O. (2022). Implicit regularization towards rank minimization in relu networks. *arXiv preprint arXiv:2201.12760*.

Tsigler, A. and Bartlett, P. L. (2020). Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*.

Vardi, G. (2022). On the implicit bias in deep-learning algorithms. *arXiv preprint arXiv:2208.12591*.

Vaskevicius, T., Kanade, V., and Rebeschini, P. (2019). Implicit regularization for optimal sparse recovery. *Advances in Neural Information Processing Systems*, 32.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

Wang, G., Donhauser, K., and Yang, F. (2022). Tight bounds for minimum ℓ_1-norm interpolation of noisy data. In *International Conference on Artificial Intelligence and Statistics*, pages 10572–10602. PMLR.

Woodworth, B., Gunasekar, S., Lee, J. D., Moroshko, E., Savarese, P., Golan, I., Soudry, D., and Srebro, N. (2020). Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673. PMLR.

Yun, C., Krishnan, S., and Mobahi, H. (2021). A unifying view on implicit bias in training linear neural networks. In *International Conference on Learning Representations*.

Zhou, L., Sutherland, D. J., and Srebro, N. (2020). On uniform convergence and low-norm interpolation learning. *Advances in Neural Information Processing Systems*, 33:6867–6877.

A Preliminary

In this section, we prepare some useful lemmas for the later analysis. In Section A.1, we show that Assumption 1 is true when data matrix $\boldsymbol{X}$ is a Gaussian random matrix and noise $\boldsymbol{\xi} \sim N(0, \sigma^2 \boldsymbol{I})$. In Section A.2, we give the proof of Lemma 5 and Lemma 6 for gradient approximation.

A.1 RIP and regularity conditions

In this subsection, we show that Assumption 1 can be satisfied when $\boldsymbol{X}$ is a Gaussian random matrix and $\boldsymbol{\xi}$ is a Gaussian random vector with variance σ^2.

We use standard proof to show the RIP property, and the rest of the properties follow from simple concentration. First, the following shows random Gaussian matrix is a $(s+1, \delta)$-RIP matrix with $\delta = \Theta(\sqrt{(s/n)\log(d/s)})$. To satisfy Assumption 1, with simple calculation we see that we only require $\widetilde{\Omega}(s^4) \le n \le \widetilde{O}(d/s^4)$.

Proposition A.1. *Let $\boldsymbol{X}$ be a $n \times d$ Gaussian random matrix. Then, there exists universal constant c_1, c_2 such that $\boldsymbol{X}/\sqrt{n}$ is (k, δ)-RIP for any $k \le c_1 n / \log(d/k)$ and $\delta \ge c_2 \sqrt{(k/n)\log(d/k)}$ with probability at least $1 - (k/d)^k \ge 1 - 1/d$.*

Proof. From the proof of Theorem 5.2 in (Baraniuk et al., 2008), we know the error probability is at most

$$e^{-c_0(\delta/2)n + k[\log(ed/k) + \log(12/\delta)] + \log(2)},$$

where $c_0(\varepsilon) = \varepsilon^2/4 - \varepsilon^3/6$. Note that it suffices to consider $\delta < 1$, which implies that $c_0(\delta/2) \ge \delta^2/48$ and $k \le n/c_2^2/\log(d/k)$. Then the exponent can be upper bounded by with $\delta \ge c_2\sqrt{(k/n)\log(d/k)}$

$$-n\delta^2/48 + (4 + \log(1/c_2))k\log(d/k) \le -(c_2^2/48)k\log(d/k) + (4 + \log(1/c_2))k\log(d/k) < -(c_2^2/50)k\log(d/k),$$

where the last inequality is true since we can choose c_2 to be large enough constant. □

The following lemma shows that the regularity conditions on $\boldsymbol{X}, \boldsymbol{\xi}$ in the second part of Assumption 1 are satisfied with high probability when $\boldsymbol{X}$ is a Gaussian matrix and $\boldsymbol{\xi}$ is sampled from $N(0, \sigma^2 \boldsymbol{I})$.

Lemma A.2 (Regularity conditions)**.** *Suppose $\boldsymbol{X}$ is a Gaussian matrix and $\boldsymbol{\xi} \sim N(0, \sigma^2 \boldsymbol{I})$. With probability at least $1 - de^{-\Omega(n)}$, We have*

$$\begin{aligned}
&\|\boldsymbol{\xi}\|_2 = O(\sigma\sqrt{n}),\\
&\left\|\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{\xi}\right\|_\infty \le B_\xi := O\left(\sigma\sqrt{\frac{\log d}{n}}\right),\\
&\left\|\boldsymbol{X}^\top \boldsymbol{\xi}\right\|_2 = O\left(\sigma\sqrt{dn}\right),\\
&\left\|\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{\beta}\right\|_\infty = O\left(\frac{\|\boldsymbol{\beta}\|_2}{\sqrt{n}}\right) \text{ for any vector } \boldsymbol{\beta},\\
&(1 - O(\sqrt{n/d}))d \le \lambda_{\min}(\boldsymbol{X}\boldsymbol{X}^\top) \le \lambda_{\max}(\boldsymbol{X}\boldsymbol{X}^\top) \le (1 + O(\sqrt{n/d}))d.
\end{aligned}$$

Proof. The first three and the last one are standard consequences of Gaussian vector/matrix concentration, see e.g., Lemma A.5 in Vaskevicius et al. (2019) for the proof of $\left\|\boldsymbol{X}^\top \boldsymbol{\xi}/n\right\|_\infty$ and Theorem 3.1.1 and Theorem 4.4.5 in Vershynin (2018) for the rest. For the third one, denote $\boldsymbol{X}[:, i]$ is the i-th column of $\boldsymbol{X}$. Then, $\left\|\boldsymbol{X}^\top \boldsymbol{\beta}/n\right\|_\infty \le \max_i |\boldsymbol{\beta}^\top \boldsymbol{X}[:, i]|/n \le \|\boldsymbol{\beta}\|_2 \max_i \|\boldsymbol{X}[:, i]\|_2 / n$. Then it follows from standard Gaussian concentration. □

Now we are ready to prove Lemma 2 that shows Assumption 1 holds under Gaussian input and Gaussian noise. It immediately follows from Proposition A.1 and Lemma A.2 above.

Lemma 2. *Suppose $\boldsymbol{X}$ is a Gaussian random matrix and $\boldsymbol{\xi} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$. Then if $\widetilde{\Omega}(s^4) \le n \le \widetilde{O}(d/s^4)$, we have Assumption 1 is satisfied with probability at least $1 - 1/d$.*

Proof. It suffices to combine Proposition A.1 and Lemma A.2. □

A.2 Gradient approximation

Lemma 5 and Lemma 6 give ways to approximate several important terms in the gradient. Here we give their proofs.

Lemma 5. *Given $n \times d$ matrix $\boldsymbol{X}/\sqrt{n}$ satisfying $(k+1, \delta)$-RIP, for any $\boldsymbol{\beta} \in \mathbb{R}^d$, let $\boldsymbol{\Delta} = \left(\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{I}\right)\boldsymbol{\beta}$, then the following hold:*

- *If $\boldsymbol{\beta}$ is k-sparse, then $\|\boldsymbol{\Delta}\|_\infty \le \sqrt{k}\delta \|\boldsymbol{\beta}\|_2$.*
- *For any vector $\boldsymbol{\beta}$, we have $\|\boldsymbol{\Delta}\|_\infty \le \delta \|\boldsymbol{\beta}\|_1$.*

Proof. For the first part, it is a standard consequence of the RIP condition, see e.g., Lemma A.3 in Vaskevicius et al. (2019). For the second part, notice that $\boldsymbol{\beta} = \sum_i \beta_i \boldsymbol{e}_i$ where $\{\boldsymbol{e}_i\}_{i=1}^d$ is the standard basis, it then follows from the first part. ☐

Lemma 6 (Gradient approximation)**.** *Under Assumption 1, we have the following gradients and their useful approximation:*

$$
\begin{aligned}
\nabla_{\boldsymbol{w}} L &= \left(\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{r}\right) \odot (2\lambda \boldsymbol{w}) = 2\lambda \left(\frac{d}{n}\boldsymbol{v}_S + \lambda \boldsymbol{w}_{S_+}^{\odot 2} - \lambda \boldsymbol{u}_{S_-}^{\odot 2} - \boldsymbol{\beta}^* + \boldsymbol{\Delta}_r\right) \odot \boldsymbol{w}, \\
\nabla_{\boldsymbol{u}} L &= -\left(\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{r}\right) \odot (2\lambda \boldsymbol{u}) = -2\lambda \left(\frac{d}{n}\boldsymbol{v}_S + \lambda \boldsymbol{w}_{S_+}^{\odot 2} - \lambda \boldsymbol{u}_{S_-}^{\odot 2} - \boldsymbol{\beta}^* + \boldsymbol{\Delta}_r\right) \odot \boldsymbol{u}, \\
\nabla_{\boldsymbol{v}} L &= \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{r} = \frac{d}{n}\boldsymbol{v}_S + \lambda \boldsymbol{w}_{S_+}^{\odot 2} - \lambda \boldsymbol{u}_{S_-}^{\odot 2} - \boldsymbol{\beta}^* + \boldsymbol{\Delta}_r,
\end{aligned}
$$

where

$$
\begin{aligned}
\|\boldsymbol{\Delta}_r\|_\infty =& O\left((1 + |nb - 1|)\, B_\xi\right) + \sqrt{s}\delta \frac{d}{n}\|\boldsymbol{v}_S\|_2 + s\delta \left\|\frac{d}{n}\boldsymbol{v}_S + \lambda \boldsymbol{w}_{S_+}^{\odot 2} - \lambda \boldsymbol{u}_{S_-}^{\odot 2} - \boldsymbol{\beta}^*\right\|_\infty \\
&+ O\left(\frac{d}{\sqrt{n}}\lambda\right)(\|\boldsymbol{w}_{e_+}\|_\infty^2 + \|\boldsymbol{u}_{e_-}\|_\infty^2) + \gamma,
\end{aligned}
$$

b and $\|\boldsymbol{\Gamma}\|_\infty \le \gamma$ are defined in (3), and recall S_+, S_- are the set of positive and negative entries of $\boldsymbol{\beta}^*$ and $e_+ = [d] \setminus S_+$, $e_- = [d] \setminus S_-$ are the corresponding complement set.

Proof. By the decomposition of $\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{v}/n$ in (3), we have

$$
\begin{aligned}
\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{r} =& \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{v} - \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{\xi} + \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X}(\lambda \boldsymbol{w}_{S_+}^{\odot 2} - \lambda \boldsymbol{u}_{S_-}^{\odot 2} - \boldsymbol{\beta}^*) + \frac{1}{n}\boldsymbol{X}^\top (\lambda \boldsymbol{X}\boldsymbol{w}_{e_+}^{\odot 2} - \lambda \boldsymbol{X}\boldsymbol{u}_{e_-}^{\odot 2}) \\
=& (b - \frac{1}{n})(\boldsymbol{X}^\top \boldsymbol{\xi})_e - \frac{1}{n}(\boldsymbol{X}^\top \boldsymbol{\xi})_S + \frac{d}{n}\boldsymbol{v}_S + \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X}(\lambda \boldsymbol{w}_{S_+}^{\odot 2} - \lambda \boldsymbol{u}_{S_-}^{\odot 2} - \boldsymbol{\beta}^*) \\
&+ \frac{1}{n}\boldsymbol{X}^\top (\lambda \boldsymbol{X}\boldsymbol{w}_{e_+}^{\odot 2} - \lambda \boldsymbol{X}\boldsymbol{u}_{e_-}^{\odot 2}) + \boldsymbol{\Gamma} \\
=& (b - \frac{1}{n})(\boldsymbol{X}^\top \boldsymbol{\xi})_e - \frac{1}{n}(\boldsymbol{X}^\top \boldsymbol{\xi})_S + \frac{d}{n}\boldsymbol{v}_S + \lambda \boldsymbol{w}_{S_+}^{\odot 2} - \lambda \boldsymbol{u}_{S_-}^{\odot 2} - \boldsymbol{\beta}^* \\
&+ (\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{I})(\frac{d}{n}\boldsymbol{v}_S + \lambda \boldsymbol{w}_{S_+}^{\odot 2} - \lambda \boldsymbol{u}_{S_-}^{\odot 2} - \boldsymbol{\beta}^*) - (\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{I})\frac{d}{n}\boldsymbol{v}_S \\
&+ \frac{1}{n}\boldsymbol{X}^\top (\lambda \boldsymbol{X}\boldsymbol{w}_{e_+}^{\odot 2} - \lambda \boldsymbol{X}\boldsymbol{u}_{e_-}^{\odot 2}) + \boldsymbol{\Gamma} \\
=& \frac{d}{n}\boldsymbol{v}_S + \lambda \boldsymbol{w}_{S_+}^{\odot 2} - \lambda \boldsymbol{u}_{S_-}^{\odot 2} - \boldsymbol{\beta}^* \\
&+ (b - \frac{1}{n})(\boldsymbol{X}^\top \boldsymbol{\xi})_e - \frac{1}{n}(\boldsymbol{X}^\top \boldsymbol{\xi})_S + (\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{I})(\frac{d}{n}\boldsymbol{v}_S + \lambda \boldsymbol{w}_{S_+}^{\odot 2} - \lambda \boldsymbol{u}_{S_-}^{\odot 2} - \boldsymbol{\beta}^*) - (\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{I})\frac{d}{n}\boldsymbol{v}_S \\
&+ \frac{1}{n}\boldsymbol{X}^\top (\lambda \boldsymbol{X}\boldsymbol{w}_{e_+}^{\odot 2} - \lambda \boldsymbol{X}\boldsymbol{u}_{e_-}^{\odot 2}) + \boldsymbol{\Gamma}.
\end{aligned}
$$

Denote the last two lines in in the last equation above as $\boldsymbol{\Delta}_r$. We know

$$\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{r} = \frac{d}{n}\boldsymbol{v}_S + \lambda \boldsymbol{w}_{S_+}^{\odot 2} - \lambda \boldsymbol{u}_{S_-}^{\odot 2} - \boldsymbol{\beta}^* + \boldsymbol{\Delta}_r.$$

To bound $\|\boldsymbol{\Delta}_r\|_\infty$, by Lemma 5 and Assumption 1, we know

$$\begin{aligned}
\left\|(b-\frac{1}{n})(\boldsymbol{X}^\top\boldsymbol{\xi})_e - \frac{1}{n}(\boldsymbol{X}^\top\boldsymbol{\xi})_S\right\|_\infty =& O\left((1+|nb-1|)\,B_\xi\right)\\
\left\|(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}-\boldsymbol{I})(\frac{d}{n}\boldsymbol{v}_S + \lambda \boldsymbol{w}_{S_+}^{\odot 2} - \lambda \boldsymbol{u}_{S_-}^{\odot 2} - \boldsymbol{\beta}^*)\right\|_\infty =& \sqrt{s}\delta\left\|\frac{d}{n}\boldsymbol{v}_S + \lambda \boldsymbol{w}_{S_+}^{\odot 2} - \lambda \boldsymbol{u}_{S_-}^{\odot 2} - \boldsymbol{\beta}^*\right\|_2\\
\leq& s\delta\left\|\frac{d}{n}\boldsymbol{v}_S + \lambda \boldsymbol{w}_{S_+}^{\odot 2} - \lambda \boldsymbol{u}_{S_-}^{\odot 2} - \boldsymbol{\beta}^*\right\|_\infty\\
\left\|(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}-\boldsymbol{I})\frac{d}{n}\boldsymbol{v}_S\right\|_\infty \leq& \sqrt{s}\delta\frac{d}{n}\left\|\boldsymbol{v}_S\right\|_2\\
\left\|\frac{1}{n}\boldsymbol{X}^\top(\lambda\boldsymbol{X}\boldsymbol{w}_{e_+}^{\odot 2} - \lambda\boldsymbol{X}\boldsymbol{u}_{e_-}^{\odot 2})\right\|_\infty =& O\left(\frac{\lambda}{\sqrt{n}}\left\|\boldsymbol{X}\boldsymbol{w}_{e_+}^{\odot 2} - \boldsymbol{X}\boldsymbol{u}_{e_-}^{\odot 2}\right\|_2\right)\\
=& O\left(\frac{d}{\sqrt{n}}\lambda\right)(\left\|\boldsymbol{w}_{e_+}\right\|_\infty^2 + \left\|\boldsymbol{u}_{e_-}\right\|_\infty^2),
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\|\boldsymbol{\Delta}_r\|_\infty =& O\left((1+|nb-1|)\,B_\xi\right) + s\delta\left\|\frac{d}{n}\boldsymbol{v}_S + \lambda \boldsymbol{w}_{S_+}^{\odot 2} - \lambda \boldsymbol{u}_{S_-}^{\odot 2} - \boldsymbol{\beta}^*\right\|_\infty + \sqrt{s}\delta\frac{d}{n}\left\|\boldsymbol{v}_S\right\|_2\\
&+ O\left(\frac{d}{\sqrt{n}}\lambda\right)(\left\|\boldsymbol{w}_{e_+}\right\|_\infty^2 + \left\|\boldsymbol{u}_{e_-}\right\|_\infty^2) + \gamma
\end{aligned}$$

□

B Proof for Stage 1

Recall that our goal in Stage 1 is to show (1) variables $\boldsymbol{w}_S$ and $\boldsymbol{u}_S$ grow large to recover $\boldsymbol{\beta}^*$ (specifically, $\boldsymbol{w}_S$ recovers the positive entries of $\boldsymbol{\beta}^*$ and $\boldsymbol{u}_S$ recovers the negative entries of $\boldsymbol{\beta}^*$); (2) the other variables $\boldsymbol{w}_e$, $\boldsymbol{u}_e$ and $\boldsymbol{v}$ remain small. This is summarized in the following main lemma:

Lemma 7 (Stage 1)**.** *Let C_1 be a large enough universal constant, denote*

$$T_1 := \inf\left\{t : \left\|\frac{d}{n}\boldsymbol{v}_S^{(T_1)} + \lambda \boldsymbol{w}_{S_+}^{(T_1)\odot 2} - \lambda \boldsymbol{u}_{S_-}^{(T_1)\odot 2} - \boldsymbol{\beta}^*\right\|_\infty = C_1(B_\xi + \sigma\sqrt{n/d})\right\}.$$

Then we know $T_1 = O(\log(1/\alpha)/\eta\lambda)$ and the following hold:

- $\left\|\boldsymbol{w}_{e_+}^{(T_1)}\right\|_\infty, \left\|\boldsymbol{u}_{e_-}^{(T_1)}\right\|_\infty = O(\alpha)$.
- $\left\|\boldsymbol{v}_S^{(T_1)}\right\|_2 = O(\sqrt{s}(n/d)\log^2(d)(B_\xi + \sigma\sqrt{n/d}))$ *and* $\left\|\boldsymbol{v}^{(T_1)}\right\|_2 = O(\sigma\sqrt{n/d})$.
- $\left\|\boldsymbol{r}^{(T_1)}\right\|_2 = O(\sigma\sqrt{n})$.

To prove this lemma, we need to maintain the following inductive hypothesis which assumes the approximate error comes from the non-signal entry is small and other regularity conditions. Later we will use these assumptions to bound different error terms and finish the induction.

Lemma B.1 (Inductive Hypothesis for Stage 1). *For $t \le \widetilde{T}_1 := C_{T_1}\log(1/\alpha)/\eta\lambda\beta_{\min}$ with large enough universal constant C_{T_1}, the following hold:*

- $\left\|\boldsymbol{w}_{e_+}^{(t)}\right\|_\infty, \left\|\boldsymbol{u}_{e_-}^{(t)}\right\|_\infty = O(\alpha)$.
- $\left\|\lambda\boldsymbol{w}_{S_+}^{(t)\odot 2} - \lambda\boldsymbol{u}_{S_-}^{(t)\odot 2} - \boldsymbol{\beta}^*\right\|_\infty = O(1)$, $\left\|\frac{d}{n}\boldsymbol{v}_S^{(t)} + \lambda\boldsymbol{w}_{S_+}^{(t)\odot 2} - \lambda\boldsymbol{u}_{S_-}^{(t)\odot 2} - \boldsymbol{\beta}^*\right\|_\infty = O(1)$.
- $\left\|\boldsymbol{r}^{(t)}\right\|_2 \le \left\|\boldsymbol{r}^{(0)}\right\|_2 = O(\sqrt{sn})$.

B.1 Dynamics of v

As we discussed earlier, even though in Stage 1 we hope to use the corresponding entries of $\boldsymbol{u}$, $\boldsymbol{w}$ to learn the signal, the same entries of $\boldsymbol{v}$ will also grow and it's important to understand the dynamics of $\boldsymbol{v}$.

The dynamics of $\boldsymbol{v}$ roughly follows the trajectory for optimizing $\|\boldsymbol{X}\boldsymbol{v} - \boldsymbol{\xi}\|_2^2/2n$. We formalize that in the following two lemmas. First, we give a decomposition of $\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{v}/n$ as follow. This term plays an important role when we estimate the gradient as shown in Lemma 6, therefore we here give a careful analysis.

Lemma B.2. *Recall the decomposition in* (3)

$$\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}^{(t)} = \frac{d}{n}\boldsymbol{v}_S^{(t)} + b_t(\boldsymbol{X}^\top\boldsymbol{\xi})_e + \boldsymbol{\Gamma}_t,$$
$$b_{t+1} = b_t - \frac{\eta d}{n}\left(b_t - \frac{1}{n}\right),$$

where $\left\|\boldsymbol{\Gamma}^{(t)}\right\|_\infty \le \gamma_t$ and recall the notation $\boldsymbol{\beta}_S = \sum_{i:\beta_i^*\ne 0}\beta_i\boldsymbol{e}_i$, $\boldsymbol{\beta}_e = \sum_{i:\beta_i^*=0}\beta_i\boldsymbol{e}_i$. Suppose Lemma B.1 holds. We have for $t \le \widetilde{T}_1$

$$b_t = (1-(1-\eta d/n)^t)/n \le 1/n,$$
$$\gamma_t \le O((\sqrt{sd/n} + ds\delta/n)\eta t) = O(\sigma\sqrt{n/d} + B_\xi).$$

We then give the decomposition of $\boldsymbol{v}$.

Lemma B.3. *Recall the decomposition in* (4)

$$\boldsymbol{v}^{(t)} = \boldsymbol{v}_S^{(t)} + a_t\boldsymbol{X}^\top\boldsymbol{\xi} + \boldsymbol{\Delta}_v^{(t)},$$
$$a_{t+1} = a_t - \eta(b_t - 1/n)$$

where $\left\|\boldsymbol{\Delta}_v^{(t)}\right\|_\infty \le \zeta_t$. and recall the notation $\boldsymbol{\beta}_S = \sum_{i:\beta_i^*\ne 0}\beta_i\boldsymbol{e}_i$, $\boldsymbol{\beta}_e = \sum_{i:\beta_i^*=0}\beta_i\boldsymbol{e}_i$. Suppose Lemma B.1 holds. We have for $t \le \widetilde{T}_1$

$$a_t = (1-(1-\eta d/n)^t)/d \le 1/d$$
$$\zeta_t = O((B_\xi + s\delta + \sigma\sqrt{n/d})\eta t) = O(\sigma\sqrt{n}/d).$$

Moreover, for every $t \le \widetilde{T}_1$, $\left\|\boldsymbol{v}^{(t)}\right\|_2 = O(\sigma\sqrt{n/d})$, $\left\|\boldsymbol{v}_S^{(t)}\right\|_2 = O(\sqrt{s}(n/d)\log^2(d)(B_\xi + \sigma\sqrt{n/d}))$.

B.2 Implications of Inductive Hypothesis Lemma B.1

Given the understanding of dynamics of $\boldsymbol{v}$ and $\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}$ in Appendix B.1, we have the following approximation of gradient, using Lemma 6.

Lemma B.4. *In the setting of Lemma B.2 and Lemma B.3, we have for* $t \leq \widetilde{T}_1$

$$\nabla_{\boldsymbol{w}} L = \left(\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{r}\right) \odot (2\lambda \boldsymbol{w}) = 2\lambda(\frac{d}{n}\boldsymbol{v}_S + \lambda \boldsymbol{w}_{S_+}^{\odot 2} - \lambda \boldsymbol{u}_{S_-}^{\odot 2} - \boldsymbol{\beta}^* + \boldsymbol{\Delta}_r) \odot \boldsymbol{w},$$

$$\nabla_{\boldsymbol{u}} L = -\left(\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{r}\right) \odot (2\lambda \boldsymbol{u}) = -2\lambda(\frac{d}{n}\boldsymbol{v}_S + \lambda \boldsymbol{w}_{S_+}^{\odot 2} - \lambda \boldsymbol{u}_{S_-}^{\odot 2} - \boldsymbol{\beta}^* + \boldsymbol{\Delta}_r) \odot \boldsymbol{u},$$

$$\nabla_v L = \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{r} = \frac{d}{n}\boldsymbol{v}_S + \lambda \boldsymbol{w}_{S_+}^{\odot 2} - \lambda \boldsymbol{u}_{S_-}^{\odot 2} - \boldsymbol{\beta}^* + \boldsymbol{\Delta}_r,$$

where

$$\left\|\boldsymbol{\Delta}_r^{(t)}\right\|_\infty = \underbrace{O\left(B_\xi + \sigma\sqrt{n/d}\right)}_{=:B_s} + s\delta \left\|\frac{d}{n}\boldsymbol{v}_S^{(t)} + \lambda \boldsymbol{w}_{S_+}^{(t)\odot 2} - \lambda \boldsymbol{u}_{S_-}^{(t)\odot 2} - \boldsymbol{\beta}^*\right\|_\infty,$$

Now we are ready to estimate the dynamics for the relevant entries of $\boldsymbol{u}$ and $\boldsymbol{w}$ using Lemma B.4. We first show in Lemma B.5 that $\boldsymbol{w}_{S_+}, \boldsymbol{u}_{S_-}$ will grow to $\Omega(\beta_{\min})$. Then in Lemma B.6 we show that it takes $O(1/\eta\lambda\beta_{\min})$ to decrease $\left\|\frac{d}{n}\boldsymbol{v}_S^{(t)} + \lambda(\boldsymbol{w}_{S_+}^{(t)})^2 - \lambda(\boldsymbol{u}_{S_-}^{(t)})^2 - \boldsymbol{\beta}^*\right\|_\infty$ by half. The proofs are deferred to Appendix B.4.

Lemma B.5. *Suppose Lemma B.1 hold. Then for every* $T_{11} \leq t \leq \widetilde{T}_1$ *with* $T_{11} = O(\log(1/\lambda\alpha^2)/\eta\lambda\beta_{\min})$, $\lambda(w_k^{(t)})^2 \geq \beta_{\min}/4$ *for* $k \in S_+$ *and* $\lambda(u_k^{(t)})^2 \geq \beta_{\min}/4$ *for* $k \in S_-$.

Lemma B.6. *Suppose Lemma B.1 and Lemma B.5 hold. Given any time* t_0*, assume at time* t_0 $B_0 := \left\|\frac{d}{n}\boldsymbol{v}_S^{(t_0)} + \lambda(\boldsymbol{w}_{S_+}^{(t_0)})^2 - \lambda(\boldsymbol{u}_{S_-}^{(t_0)})^2 - \boldsymbol{\beta}^*\right\|_\infty \geq 4B_s$ *where* B_s *is defined in Lemma B.4. Let*

$$T' := \inf\left\{t - t_0 \geq 0 : \left\|\frac{d}{n}\boldsymbol{v}_S^{(t)} + \lambda(\boldsymbol{w}_{S_+}^{(t)})^2 - \lambda(\boldsymbol{u}_{S_-}^{(t)})^2 - \boldsymbol{\beta}^*\right\|_\infty \leq \left\|\frac{d}{n}\boldsymbol{v}_S^{(t_0)} + \lambda(\boldsymbol{w}_{S_+}^{(t_0)})^2 - \lambda(\boldsymbol{u}_{S_-}^{(t_0)})^2 - \boldsymbol{\beta}^*\right\|_\infty /2\right\}$$

be the time that signal error reduces by half. Then, we know $T' = O(1/\eta\lambda\beta_{\min})$.

As a technical condition in proving the two lemmas above, we need to make sure that once we fit the signal using the corresponding entries in $\boldsymbol{u}, \boldsymbol{w}, \boldsymbol{v}$ up to error μ, the error will not become much worse later. We formalize this as the following stability lemma.

Lemma B.7 (Stability)**.** *Suppose Lemma B.4 and Lemma B.5 hold. Assume* $\left\|\frac{d}{n}\boldsymbol{v}^{(t_0)} + \lambda \boldsymbol{w}^{(t_0)\odot 2} - \lambda \boldsymbol{u}^{(t_0)\odot 2} - \boldsymbol{\beta}^*\right\|_\infty \leq \mu$ *at time* t_0*, then* $|\frac{d}{n}v_k^{(t)} + \lambda(w_k^{(t)})^2 - \lambda(u_k^{(t)})^2 - \beta_k^*| \leq \max\{\mu, 2(B_s + s\delta\mu)\}$ *for all* $t \geq t_0$ *and* $k \in S$*, where* B_s *is definied in Lemma B.4.*

Now we are ready to bound the time T_1 for Stage 1 using the above lemmas.

Lemma B.8. *Suppose Lemma B.1 hold. Recall*

$$T_1 := \inf\left\{t : \left\|\frac{d}{n}\boldsymbol{v}_S^{(t)} + \lambda(\boldsymbol{w}_{S_+}^{(t)})^2 - \lambda(\boldsymbol{u}_{S_-}^{(t)})^2 - \boldsymbol{\beta}^*\right\|_\infty \leq C_1(B_\xi + \sigma\sqrt{n/d})\right\},$$

where C_1 *is a large enough universal constant. Then, we know* $T_1 = O(\log(1/\alpha)/\eta\lambda\beta_{\min})$.

B.3 Proof of Inductive Hypothesis Lemma B.1 and Lemma 7

Finally, we are ready to prove in the induction hypothesis and finish the proof of Lemma 7.

Lemma B.1 (Inductive Hypothesis for Stage 1)**.** *For* $t \leq \widetilde{T}_1 := C_{T_1}\log(1/\alpha)/\eta\lambda\beta_{\min}$ *with large enough universal constant* C_{T_1}*, the following hold:*

- $\left\|\boldsymbol{w}_{e_+}^{(t)}\right\|_\infty, \left\|\boldsymbol{u}_{e_-}^{(t)}\right\|_\infty = O(\alpha)$.
- $\left\|\lambda\boldsymbol{w}_{S_+}^{(t)\odot 2} - \lambda\boldsymbol{u}_{S_-}^{(t)\odot 2} - \boldsymbol{\beta}^*\right\|_\infty = O(1)$, $\left\|\frac{d}{n}\boldsymbol{v}_S^{(t)} + \lambda\boldsymbol{w}_{S_+}^{(t)\odot 2} - \lambda\boldsymbol{u}_{S_-}^{(t)\odot 2} - \boldsymbol{\beta}^*\right\|_\infty = O(1)$.
- $\left\|\boldsymbol{r}^{(t)}\right\|_2 \le \left\|\boldsymbol{r}^{(0)}\right\|_2 = O(\sqrt{sn})$.

Proof. We claim $\left\|\boldsymbol{w}_{e_+}^{(t)}\right\|_\infty, \left\|\boldsymbol{u}_{e_-}^{(t)}\right\|_\infty = \alpha(1+O(\eta\lambda(B_\xi+\sigma\sqrt{n/d}+s\delta)))^t$ and $\left\|\frac{d}{n}\boldsymbol{v}_S^{(t)} + \lambda\boldsymbol{w}_{S_+}^{(t)\odot 2} - \lambda\boldsymbol{u}_{S_-}^{(t)\odot 2} - \boldsymbol{\beta}^*\right\|_\infty = O(1)$. If such claim is true, then we prove the first 2 points as $t \le \widetilde{T}_1$ and $(d/n)\left\|\boldsymbol{v}_S\right\|_\infty = O(1)$ by Lemma B.3.

We show the above claim by induction. We know all the conditions hold at $t = 0$. Suppose before time t it holds, then consider the time $t+1$.

For $\left\|\frac{d}{n}\boldsymbol{v}_S^{(t+1)} + \lambda\boldsymbol{w}_{S_+}^{(t+1)\odot 2} - \lambda\boldsymbol{u}_{S_-}^{(t+1)\odot 2} - \boldsymbol{\beta}^*\right\|_\infty$, if $\lambda(w_k^{(t+1)})^2 + \lambda(u_k^{(t+1)})^2 \le \beta_{\min}/4$, then it is easy to see it is bounded by $O(1)$. Otherwise, we can see it from the proof of Lemma B.6 and Lemma B.8.

Now consider $\left\|\boldsymbol{w}_{e_+}^{(t+1)}\right\|_\infty$ and $\left\|\boldsymbol{u}_{e_-}^{(t+1)}\right\|_\infty$. By Lemma B.4 we have for $k \notin S$

$$|w_k^{(t+1)}| \le (1 + 2\lambda\eta O(B_\xi + \sigma\sqrt{n/d} + s\delta))|w_k^{(t)}|,$$

which implies that $|w_k^{(t+1)}| \le (1 + O(\lambda\eta(B_\xi + \sigma\sqrt{n/d} + s\delta)))^{t+1}\alpha$ as $w_k^{(0)} = \alpha$. Similarly, we get the same bound for u_k with $k \notin S$.

It remains to consider w_k with $k \in S_-$ and u_k with $k \in S_+$. We will focus on w_k with $k \in S_-$, the other follows the same calculation. We have

$$w_k^{(t+1)}u_k^{(t+1)} = \left(1 - 2\eta\lambda\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{r}^{(t)}\right)_k\right)w_k^{(t)} \cdot \left(1 + 2\eta\lambda\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{r}^{(t)}\right)_k\right)u_k^{(t)} \le w_k^{(t)}u_k^{(t)} \le \alpha^2.$$

From the proof of Lemma B.8 we know $u_k^{(t)} \ge \alpha$. This implies that $|w_k^{(t)}| \le \alpha$.

We now prove the last part on $\left\|\boldsymbol{r}^{(t+1)}\right\|_2$. We have

$$\begin{aligned}
\boldsymbol{r}^{(t+1)} =& \boldsymbol{X}\boldsymbol{v}^{(t+1)} + \lambda\boldsymbol{X}\boldsymbol{w}^{(t+1)\odot 2} - \lambda\boldsymbol{X}\boldsymbol{u}^{(t+1)\odot 2} - \boldsymbol{\xi} \\
=& \boldsymbol{r}^{(t)} - \eta\boldsymbol{X}\cdot\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{r}^{(t)} + \lambda\boldsymbol{X}\left(-\eta\frac{4\lambda}{n}(\boldsymbol{X}^\top\boldsymbol{r})\odot\boldsymbol{w}^{\odot 2} + \eta^2\frac{4\lambda^2}{n^2}(\boldsymbol{X}^\top\boldsymbol{r})^{\odot 2}\odot\boldsymbol{w}^{\odot 2}\right) \\
&- \lambda\boldsymbol{X}\left(\eta\frac{4\lambda}{n}(\boldsymbol{X}^\top\boldsymbol{r})\odot\boldsymbol{u}^{\odot 2} + \eta^2\frac{4\lambda^2}{n^2}(\boldsymbol{X}^\top\boldsymbol{r})^{\odot 2}\odot\boldsymbol{u}^{\odot 2}\right).
\end{aligned}$$

This suggests that

$$\begin{aligned}
\left\|\boldsymbol{r}^{(t+1)}\right\|_2 \le& \left(1 - \frac{\eta}{n}\lambda_{\min}(\boldsymbol{X}\boldsymbol{X}^\top) - \frac{4\eta\lambda^2}{n}\lambda_{\min}(\boldsymbol{X}\mathrm{diag}(\boldsymbol{w}^{\odot 2} + \boldsymbol{u}^{\odot 2})\boldsymbol{X}^\top)\right)\left\|\boldsymbol{r}^{(t)}\right\|_2 + \lambda\sqrt{d}\cdot O\left(\eta^2\frac{\lambda^2}{n^2}d\left\|\boldsymbol{r}\right\|_2^2\right) \\
\le& \left(1 - \Omega\left(\frac{\eta d}{n}\right)\right)\left\|\boldsymbol{r}^{(t)}\right\|_2
\end{aligned}$$

where we use Lemma B.10 and Assumption 1. We finish the induction. □

Now we are ready to proof the main result Lemma 7 for Stage 1.

Lemma 7 (Stage 1)**.** *Let C_1 be a large enough universal constant, denote*

$$T_1 := \inf\left\{t : \left\|\frac{d}{n}\boldsymbol{v}_S^{(T_1)} + \lambda\boldsymbol{w}_{S_+}^{(T_1)\odot 2} - \lambda\boldsymbol{u}_{S_-}^{(T_1)\odot 2} - \boldsymbol{\beta}^*\right\|_\infty = C_1(B_\xi + \sigma\sqrt{n/d})\right\}.$$

Then we know $T_1 = O(\log(1/\alpha)/\eta\lambda)$ and the following hold:

- $\left\|\boldsymbol{w}_{e_+}^{(T_1)}\right\|_\infty, \left\|\boldsymbol{u}_{e_-}^{(T_1)}\right\|_\infty = O(\alpha)$.
- $\left\|\boldsymbol{v}_S^{(T_1)}\right\|_2 = O(\sqrt{s}(n/d)\log^2(d)(B_\xi + \sigma\sqrt{n/d}))$ *and* $\left\|\boldsymbol{v}^{(T_1)}\right\|_2 = O(\sigma\sqrt{n/d})$.
- $\left\|\boldsymbol{r}^{(T_1)}\right\|_2 = O(\sigma\sqrt{n})$.

Proof. Combine Lemma B.1, Lemma B.3, Lemma B.8 we prove the first 2 points and bound the time T_1. For the last point, by Lemma 5 and Assumption 1

$$\begin{aligned}\left\|\boldsymbol{r}^{(T_1)}\right\|_2 \le& \left\|\boldsymbol{X}\lambda\boldsymbol{w}_{e_+}^{(T_1)\odot 2}\right\|_2 + \left\|\boldsymbol{X}\lambda\boldsymbol{u}_{e_-}^{(T_1)\odot 2}\right\|_2 + \left\|\boldsymbol{X}(\boldsymbol{v}_S^{(T_1)} + \lambda\boldsymbol{w}_{S_+}^{(T_1)\odot 2} - \lambda\boldsymbol{u}_{S_-}^{(T_1)\odot 2} - \boldsymbol{\beta}^*)\right\|_2 + \left\|\boldsymbol{X}(\boldsymbol{v}^{(T_1)} - \boldsymbol{v}_S^{(T_1)}) - \boldsymbol{\xi}\right\|_2\\
\le& O(\lambda d\alpha^2) + O(\sqrt{ns}(B_\xi + \sigma\sqrt{n/d})) + (d/n - 1)\left\|\boldsymbol{X}\boldsymbol{v}_S^{(T_1)}\right\|_2 + \left\|(a_{T_1}\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{I})\boldsymbol{\xi} + \boldsymbol{\Delta}_v^{(T_1)}\right\|_2\\
\le& O(1)\left\|\boldsymbol{\xi}\right\|_2 + O(\lambda d\alpha^2) + O(\sqrt{ns}(B_\xi + \sigma\sqrt{n/d})) + \widetilde{O}(\sqrt{ns}(B_\xi + \sigma\sqrt{n/d})) + \sqrt{d}\zeta_{T_1}\\
=& O(\sigma\sqrt{n}),\end{aligned}$$

where we use $a_{T_1} \le 1/d$ and $\zeta_{T_1} = O(\sigma\sqrt{n}/d)$ from Lemma B.3. □

B.4 Omitted Proofs in Section B.1 and Section B.2

In this subsection, we give the proof of Lemma B.2, Lemma B.3, Lemma B.4, Lemma B.5, Lemma B.6, Lemma B.7 and Lemma B.8 in previous subsections.

Lemma B.2. *Recall the decomposition in* (3)

$$\begin{aligned}\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}^{(t)} &= \frac{d}{n}\boldsymbol{v}_S^{(t)} + b_t(\boldsymbol{X}^\top\boldsymbol{\xi})_e + \boldsymbol{\Gamma}_t,\\
b_{t+1} &= b_t - \frac{\eta d}{n}\left(b_t - \frac{1}{n}\right),\end{aligned}$$

where $\left\|\boldsymbol{\Gamma}^{(t)}\right\|_\infty \le \gamma_t$ *and recall the notation* $\boldsymbol{\beta}_S = \sum_{i:\beta_i^*\neq 0}\beta_i\boldsymbol{e}_i$, $\boldsymbol{\beta}_e = \sum_{i:\beta_i^*=0}\beta_i\boldsymbol{e}_i$. *Suppose Lemma B.1 holds. We have for* $t \le \widetilde{T}_1$

$$\begin{aligned}b_t &= (1 - (1 - \eta d/n)^t)/n \le 1/n,\\
\gamma_t &\le O((\sqrt{sd/n} + ds\delta/n)\eta t) = O(\sigma\sqrt{n/d} + B_\xi).\end{aligned}$$

Proof. We first write the update of b_t and $\boldsymbol{\Gamma}_t$ using the update of $\boldsymbol{v}$.

$$\begin{aligned}b_{t+1}(\boldsymbol{X}^\top\boldsymbol{\xi})_e + \boldsymbol{\Gamma}_{t+1} =& \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}^{(t+1)} - \frac{d}{n}\boldsymbol{v}_S^{(t+1)}\\
=& \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}^{(t)} - \frac{d}{n}\boldsymbol{v}_S^{(t)} - \eta\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{r}^{(t)} + \eta\frac{d}{n}\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{r}^{(t)}\right)_S\\
=& b_t(\boldsymbol{X}^\top\boldsymbol{\xi})_e + \boldsymbol{\Gamma}_t - \frac{\eta}{n^2}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{r}^{(t)} + \eta\frac{d}{n}\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{r}^{(t)}\right)_S\\
=& b_t(\boldsymbol{X}^\top\boldsymbol{\xi})_e + \boldsymbol{\Gamma}_t - \frac{\eta}{n^2}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top - d\boldsymbol{I})\boldsymbol{r}^{(t)} - \eta\frac{d}{n}\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{r}^{(t)}\right)_e.\end{aligned}$$

We bound the last two terms one by one. For $\frac{\eta}{n^2}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top - d\boldsymbol{I})\boldsymbol{r}^{(t)}$, we have by Assumption 1 and Lemma B.1

$$\left\|\frac{\eta}{n^2}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top - d\boldsymbol{I})\boldsymbol{r}^{(t)}\right\|_\infty \le \frac{\eta}{n}O(\frac{1}{\sqrt{n}}\cdot\sqrt{dn}\cdot\sqrt{sn}) = O(\eta\sqrt{sd/n}).$$

For $\eta\frac{d}{n}\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{r}^{(t)}\right)_e$, we have

$$\begin{aligned}\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{r}^{(t)}\right)_e &= \left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}^{(t)} + \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}(\lambda\boldsymbol{w}_{S_+}^{(t)\odot 2} - \lambda\boldsymbol{u}_{S_-}^{(t)\odot 2} - \boldsymbol{\beta}^*) - \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\xi} + \frac{1}{n}\boldsymbol{X}^\top(\lambda\boldsymbol{X}\boldsymbol{w}_{e_+}^{(t)\odot 2} - \lambda\boldsymbol{X}\boldsymbol{u}_{e_-}^{(t)\odot 2})\right)_e \\ &= \left(\frac{d}{n}\boldsymbol{v}_S^{(t)} + \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}(\lambda\boldsymbol{w}_{S_+}^{(t)\odot 2} - \lambda\boldsymbol{u}_{S_-}^{(t)\odot 2} - \boldsymbol{\beta}^*) + (b_t - \frac{1}{n})\boldsymbol{X}^\top\boldsymbol{\xi} + \boldsymbol{\Gamma}_t + \frac{1}{n}\boldsymbol{X}^\top(\lambda\boldsymbol{X}\boldsymbol{w}_{e_+}^{(t)\odot 2} - \lambda\boldsymbol{X}\boldsymbol{u}_{e_-}^{(t)\odot 2})\right)_e \\ &= (b_t - \frac{1}{n})(\boldsymbol{X}^\top\boldsymbol{\xi})_e + \left((\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X} - \boldsymbol{I})(\lambda\boldsymbol{w}_{S_+}^{(t)\odot 2} - \lambda\boldsymbol{u}_{S_-}^{(t)\odot 2} - \boldsymbol{\beta}^*) + \boldsymbol{\Gamma}_t + \frac{1}{n}\boldsymbol{X}^\top(\lambda\boldsymbol{X}\boldsymbol{w}_{e_+}^{(t)\odot 2} - \lambda\boldsymbol{X}\boldsymbol{u}_{e_-}^{(t)\odot 2})\right)_e.\end{aligned}$$

Therefore, by Lemma B.1 we know

$$\begin{aligned}b_{t+1} &= b_t - \frac{\eta d}{n}(b_t - \frac{1}{n}), \\ \gamma_{t+1} &\le \gamma_t + O(\eta\sqrt{sd/n}) + \eta\frac{d}{n}O(s\delta + (d/\sqrt{n})\lambda\alpha^2) = \gamma_t + \eta O(\sqrt{sd/n} + ds\delta/n).\end{aligned}$$

This implies

$$\begin{aligned}b_t &= (1 - (1 - \eta d/n)^t)/n \le 1/n, \\ \gamma_t &\le O((\sqrt{sd/n} + ds\delta/n)\eta t) = O(\sigma\sqrt{n/d} + B_\xi).\end{aligned}$$

□

Lemma B.3. *Recall the decomposition in* (4)

$$\begin{aligned}\boldsymbol{v}^{(t)} &= \boldsymbol{v}_S^{(t)} + a_t\boldsymbol{X}^\top\boldsymbol{\xi} + \boldsymbol{\Delta}_v^{(t)}, \\ a_{t+1} &= a_t - \eta(b_t - 1/n)\end{aligned}$$

where $\left\|\boldsymbol{\Delta}_v^{(t)}\right\|_\infty \le \zeta_t$. *and recall the notation* $\boldsymbol{\beta}_S = \sum_{i:\beta_i^*\neq 0}\beta_i\boldsymbol{e}_i$, $\boldsymbol{\beta}_e = \sum_{i:\beta_i^*=0}\beta_i\boldsymbol{e}_i$. *Suppose Lemma B.1 holds. We have for* $t \le \widetilde{T}_1$

$$\begin{aligned}a_t &= (1 - (1 - \eta d/n)^t)/d \le 1/d \\ \zeta_t &= O((B_\xi + s\delta + \sigma\sqrt{n/d})\eta t) = O(\sigma\sqrt{n}/d).\end{aligned}$$

Moreover, for every $t \le \widetilde{T}_1$, $\left\|\boldsymbol{v}^{(t)}\right\|_2 = O(\sigma\sqrt{n/d})$, $\left\|\boldsymbol{v}_S^{(t)}\right\|_2 = O(\sqrt{s}(n/d)\log^2(d)(B_\xi + \sigma\sqrt{n/d}))$.

Proof. We write the update of a_t and $\boldsymbol{\Delta}_v^{(t)}$ using the update of $\boldsymbol{v}$

$$\begin{aligned}a_{t+1}\boldsymbol{X}^\top\boldsymbol{\xi} + \boldsymbol{\Delta}_v^{(t+1)} =& \boldsymbol{v}^{(t+1)} - \boldsymbol{v}_S^{(t+1)} = \boldsymbol{v}^{(t)} - \boldsymbol{v}_S^{(t)} - \eta\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{r}^{(t)}\right)_e \\ =& a_t\boldsymbol{X}^\top\boldsymbol{\xi} + \boldsymbol{\Delta}_v^{(t)} - \eta\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{r}^{(t)}\right)_e.\end{aligned}$$

For $\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{r}^{(t)}\right)_e$, using the decomposition of $\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}/n$ in Lemma B.2, we have

$$\begin{aligned}\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{r}^{(t)}\right)_e &= \left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}^{(t)} + \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}(\lambda\boldsymbol{w}_{S_+}^{(t)\odot 2} - \lambda\boldsymbol{u}_{S_-}^{(t)\odot 2} - \boldsymbol{\beta}^*) - \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\xi} + \frac{1}{n}\boldsymbol{X}^\top(\lambda\boldsymbol{X}\boldsymbol{w}_{e_+}^{(t)\odot 2} - \lambda\boldsymbol{X}\boldsymbol{u}_{e_-}^{(t)\odot 2})\right)_e \\ &= \left(\frac{d}{n}\boldsymbol{v}_S^{(t)} + \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}(\lambda\boldsymbol{w}_{S_+}^{(t)\odot 2} - \lambda\boldsymbol{u}_{S_-}^{(t)\odot 2} - \boldsymbol{\beta}^*) + (b_t - \frac{1}{n})\boldsymbol{X}^\top\boldsymbol{\xi} + \boldsymbol{\Gamma}_t + \frac{1}{n}\boldsymbol{X}^\top(\lambda\boldsymbol{X}\boldsymbol{w}_{e_+}^{(t)\odot 2} - \lambda\boldsymbol{X}\boldsymbol{u}_{e_-}^{(t)\odot 2})\right)_e \\ &= (b_t - \frac{1}{n})(\boldsymbol{X}^\top\boldsymbol{\xi})_e + \left((\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X} - \boldsymbol{I})(\lambda\boldsymbol{w}_{S_+}^{(t)\odot 2} - \lambda\boldsymbol{u}_{S_-}^{(t)\odot 2} - \boldsymbol{\beta}^*) + \boldsymbol{\Gamma}_t + \frac{1}{n}\boldsymbol{X}^\top(\lambda\boldsymbol{X}\boldsymbol{w}_{e_+}^{(t)\odot 2} - \lambda\boldsymbol{X}\boldsymbol{u}_{e_-}^{(t)\odot 2})\right)_e.\end{aligned}$$

Therefore, we have the update of a_t and ζ_t by using Assumption 1, Lemma 5 and Lemma B.1

$$a_{t+1} = a_t - \eta(b_t - 1/n),$$
$$\zeta_{t+1} \le \zeta_t + \eta O(|nb_t - 1| B_\xi + s\delta + \sigma\sqrt{n/d} + B_\xi + (d/\sqrt{n})\lambda\alpha^2).$$

This implies

$$a_t = \eta t/n - \eta \sum_{\tau<t} b_\tau = (1-(1-\eta d/n)^t)/d \le 1/d$$
$$\zeta_t \le O((B_\xi + s\delta + \sigma\sqrt{n/d})\eta t) = O(\sigma\sqrt{n}/d).$$

Thus, we have $\left\|\boldsymbol{v}^{(t)} - \boldsymbol{v}_S^{(t)}\right\|_2 \le a_t O(\sigma\sqrt{dn}) + \zeta_t\sqrt{d} = O(\sigma\sqrt{n/d})$. We now bound $\|\boldsymbol{v}_S\|_2$. Since its gradient norm $\|\nabla_{\boldsymbol{v}_S} L\|_2 = \left\|(\boldsymbol{X}^\top \boldsymbol{r}/n)_S\right\|_2 \le O(\sqrt{s})$ by Lemma B.1 and Assumption 1, we can bound $\|\boldsymbol{v}_S\|_2$ as $\|\boldsymbol{v}_S\|_2 = \eta\sum_{\tau\le t}\left\|\nabla_{\boldsymbol{v}_S} L^{(\tau)}\right\|_2 \le O(\sqrt{s}\eta t) = O(\sqrt{s}(n/d)\log^2(d)(B_\xi + \sigma\sqrt{n/d}))$. This also implies $\|\boldsymbol{v}\|_2 = O(\sigma\sqrt{n/d})$. □

Lemma B.4. *In the setting of Lemma B.2 and Lemma B.3, we have for $t \le \widetilde{T}_1$*

$$\nabla_{\boldsymbol{w}} L = \left(\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{r}\right) \odot (2\lambda \boldsymbol{w}) = 2\lambda(\frac{d}{n}\boldsymbol{v}_S + \lambda \boldsymbol{w}_{S_+}^{\odot 2} - \lambda \boldsymbol{u}_{S_-}^{\odot 2} - \boldsymbol{\beta}^* + \boldsymbol{\Delta}_r) \odot \boldsymbol{w},$$
$$\nabla_{\boldsymbol{u}} L = -\left(\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{r}\right) \odot (2\lambda \boldsymbol{u}) = -2\lambda(\frac{d}{n}\boldsymbol{v}_S + \lambda \boldsymbol{w}_{S_+}^{\odot 2} - \lambda \boldsymbol{u}_{S_-}^{\odot 2} - \boldsymbol{\beta}^* + \boldsymbol{\Delta}_r) \odot \boldsymbol{u},$$
$$\nabla_v L = \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{r} = \frac{d}{n}\boldsymbol{v}_S + \lambda \boldsymbol{w}_{S_+}^{\odot 2} - \lambda \boldsymbol{u}_{S_-}^{\odot 2} - \boldsymbol{\beta}^* + \boldsymbol{\Delta}_r,$$

where

$$\left\|\boldsymbol{\Delta}_r^{(t)}\right\|_\infty = \underbrace{O\left(B_\xi + \sigma\sqrt{n/d}\right)}_{=:B_s} + s\delta \left\|\frac{d}{n}\boldsymbol{v}_S^{(t)} + \lambda \boldsymbol{w}_{S_+}^{(t)\odot 2} - \lambda \boldsymbol{u}_{S_-}^{(t)\odot 2} - \boldsymbol{\beta}^*\right\|_\infty,$$

Proof. By Lemma B.2 and Lemma B.3 and the choice of parameter, the result directly follows from Lemma 6. □

Lemma B.5. *Suppose Lemma B.1 hold. Then for every $T_{11} \le t \le \widetilde{T}_1$ with $T_{11} = O(\log(1/\lambda\alpha^2)/\eta\lambda\beta_{\min})$, $\lambda(w_k^{(t)})^2 \ge \beta_{\min}/4$ for $k \in S_+$ and $\lambda(u_k^{(t)})^2 \ge \beta_{\min}/4$ for $k \in S_-$.*

Proof. For $t \le \widetilde{T}_1$, by Lemma B.4, we have for $k \in S_+$ (note that $(\boldsymbol{u}_{S_-})_k = 0$ in this case. The case $k \in S_-$ is similar, we omit for simplicity)

$$w_k^{(t+1)} = \left(1 - 2\eta\lambda\left(\frac{d}{n}v_k^{(t)} + \lambda(w_k^{(t)})^2 - \beta_k^* \pm O(B_\xi + \sigma\sqrt{n/d} + s\delta)\right)\right) w_k^{(t)},$$
$$v_k^{(t+1)} = v_k^{(t)} - \eta\left(\frac{d}{n}v_k^{(t)} + \lambda(w_k^{(t)})^2 - \beta_k^* \pm O(B_\xi + \sigma\sqrt{n/d} + s\delta)\right).$$

Since $\left\|\boldsymbol{v}_S^{(t)}\right\|_\infty = O(\sqrt{s}(n/d)\log^2(d)(B_\xi + \sigma\sqrt{n/d}))$ by Lemma B.3, this implies that $(d/n)\left\|\boldsymbol{v}_S^{(t)}\right\|_\infty < \beta_{\min}/4$. Thus,

$$\lambda(w_k^{(t+1)})^2 = \left(1 - 2\eta\lambda\left(\frac{d}{n}v_k^{(t)} + \lambda(w_k^{(t)})^2 - \beta_k^* \pm O(B_\xi + \sigma\sqrt{n/d} + s\delta)\right)\right)^2 \lambda(w_k^{(t)})^2$$
$$\ge \left(1 - 2\eta\lambda\left(\lambda(w_k^{(t)})^2 - \beta_k^*/2\right)\right)^2 \lambda(w_k^{(t)})^2.$$

Therefore, by Lemma B.9 within time $O(\log(1/\lambda\alpha^2)/\eta\lambda\beta_{\min})$ we have $\lambda(w_k^{(t)})^2 \ge \beta_{\min}/4$ and will remain for $t \le \widetilde{T}_1$. □

Lemma B.6. *Suppose Lemma B.1 and Lemma B.5 hold. Given any time t_0, assume at time t_0 $B_0 := \left\|\frac{d}{n}\boldsymbol{v}_S^{(t_0)} + \lambda(\boldsymbol{w}_{S_+}^{(t_0)})^2 - \lambda(\boldsymbol{u}_{S_-}^{(t_0)})^2 - \boldsymbol{\beta}^*\right\|_\infty \geq 4B_s$ where B_s is defined in Lemma B.4. Let*

$$T' := \inf\left\{t - t_0 \geq 0 : \left\|\frac{d}{n}\boldsymbol{v}_S^{(t)} + \lambda(\boldsymbol{w}_{S_+}^{(t)})^2 - \lambda(\boldsymbol{u}_{S_-}^{(t)})^2 - \boldsymbol{\beta}^*\right\|_\infty \leq \left\|\frac{d}{n}\boldsymbol{v}_S^{(t_0)} + \lambda(\boldsymbol{w}_{S_+}^{(t_0)})^2 - \lambda(\boldsymbol{u}_{S_-}^{(t_0)})^2 - \boldsymbol{\beta}^*\right\|_\infty /2\right\}$$

be the time that signal error reduces by half. Then, we know $T' = O(1/\eta\lambda\beta_{\min})$.

Proof. For $t \leq t_0 + T'$, by Lemma B.4, we have for $k \in S_+$ (note that $(\boldsymbol{u}_{S_-})_k = 0$ in this case. The case $k \in S_-$ is similar, we omit for simplicity)

$$\begin{aligned}
w_k^{(t+1)} &= \left(1 - 2\eta\lambda\left(\frac{d}{n}v_k^{(t)} + \lambda(w_k^{(t)})^2 - \beta_k^* \pm \left(B_s + s\delta\left\|\frac{d}{n}\boldsymbol{v}_S^{(t)} + \lambda(\boldsymbol{w}_{S_+}^{(t)})^2 - \lambda(\boldsymbol{u}_{S_-}^{(t)})^2 - \boldsymbol{\beta}^*\right\|_\infty\right)\right)\right) w_k^{(t)},\\
v_k^{(t+1)} &= v_k^{(t)} - \eta\left(\frac{d}{n}v_k^{(t)} + \lambda(w_k^{(t)})^2 - \beta_k^* \pm \left(B_s + s\delta\left\|\frac{d}{n}\boldsymbol{v}_S^{(t)} + \lambda(\boldsymbol{w}_{S_+}^{(t)})^2 - \lambda(\boldsymbol{u}_{S_-}^{(t)})^2 - \boldsymbol{\beta}^*\right\|_\infty\right)\right).
\end{aligned}$$

We claim $\left\|\frac{d}{n}\boldsymbol{v}_S^{(t)} + \lambda(\boldsymbol{w}_{S_+}^{(t)})^2 - \lambda(\boldsymbol{u}_{S_-}^{(t)})^2 - \boldsymbol{\beta}^*\right\|_\infty \leq \left\|\frac{d}{n}\boldsymbol{v}_S^{(t_0)} + \lambda(\boldsymbol{w}_{S_+}^{(t_0)})^2 - \lambda(\boldsymbol{u}_{S_-}^{(t_0)})^2 - \boldsymbol{\beta}^*\right\|_\infty = B_0$ for $t_0 \leq t \leq t_0 + T'$. We show this by induction. At $t = t_0$ it holds. Suppose before t the claim holds. For time $t+1$,

$$\begin{aligned}
\frac{d}{n}v_k^{(t+1)} + \lambda(w_k^{(t+1)})^2 =& \frac{d}{n}v_k^{(t)} - \frac{d}{n}\eta\left(\frac{d}{n}v_k^{(t)} + \lambda(w_k^{(t)})^2 - \beta_k^* \pm B_0/3\right)\\
&+ \left(1 - 2\eta\lambda\left(\frac{d}{n}v_k^{(t)} + \lambda(w_k^{(t)})^2 - \beta_k^* \pm B_0/3\right)\right)^2 \lambda(w_k^{(t)})^2\\
\geq& \frac{d}{n}v_k^{(t)} + \lambda(w_k^{(t)})^2 - \eta\left(\frac{d}{n}v_k^{(t)} + \lambda(w_k^{(t)})^2 - \beta_k^* \pm B_0/3\right)\left(\frac{d}{n} + 4\lambda^2(w_k^{(t)})^2\right).
\end{aligned}$$

This implies for $t \leq t_0 + T'$

$$\begin{aligned}
\frac{d}{n}v_k^{(t+1)} + \lambda(w_k^{(t+1)})^2 - \beta_k^* &\geq \left(\frac{d}{n}v_k^{(t)} + \lambda(w_k^{(t)})^2 - \beta_k^*\right)\left(1 - \frac{\eta}{3}\left(\frac{d}{n} + 4\lambda^2(w_k^{(t)})^2\right)\right)\\
&\geq \left(\frac{d}{n}v_k^{(t)} + \lambda(w_k^{(t)})^2 - \beta_k^*\right)(1 - \Omega(\eta\lambda\beta_{\min})),
\end{aligned}$$

where in the last line we use Lemma B.5. Thus, if $\frac{d}{n}v_k^{(t)} + \lambda(w_k^{(t)})^2 - \beta_k^* < -B_0/2$, then it will increase so that $|\frac{d}{n}v_k^{(t)} + \lambda(w_k^{(t)})^2 - \beta_k^*| \leq B_0$. Similarly, one can show that if $\frac{d}{n}v_k^{(t)} + \lambda(w_k^{(t)})^2 - \beta_k^* > B_0/2$, then it will decrease so that $|\frac{d}{n}v_k^{(t)} + \lambda(w_k^{(t)})^2 - \beta_k^*| \leq B_0$. In this way, we finish the induction.

Moreover, from the above calculations, we can see that if $\frac{d}{n}v_k^{(t)} + \lambda(w_k^{(t)})^2 - \beta_k^* < -B_0/2$, then within time $O(1/\eta\lambda\beta_{\min})$, $|\frac{d}{n}v_k^{(t)} + \lambda(w_k^{(t)})^2 - \beta_k^*| \leq B_0/2$. Similarly, if $\frac{d}{n}v_k^{(t)} + \lambda(w_k^{(t)})^2 - \beta_k^* > B_0/2$, then within time $O(1/\eta\lambda\beta_{\min})$, $|\frac{d}{n}v_k^{(t)} + \lambda(w_k^{(t)})^2 - \beta_k^*| \leq B_0/2$. By Lemma B.7, we know once $|\frac{d}{n}v_k^{(t)} + \lambda(w_k^{(t)})^2 - \beta_k^*| \leq B_0/2$, it will remain bounded by $B_0/2$. Therefore, we know $T' = O(1/\eta\lambda\beta_{\min})$. ☐

Lemma B.7 (Stability)**.** *Suppose Lemma B.4 and Lemma B.5 hold. Assume $\left\|\frac{d}{n}\boldsymbol{v}^{(t_0)} + \lambda\boldsymbol{w}^{(t_0)\odot 2} - \lambda\boldsymbol{u}^{(t_0)\odot 2} - \boldsymbol{\beta}^*\right\|_\infty \leq \mu$ at time t_0, then $|\frac{d}{n}v_k^{(t)} + \lambda(w_k^{(t)})^2 - \lambda(u_k^{(t)})^2 - \beta_k^*| \leq \max\{\mu, 2(B_s + s\delta\mu)\}$ for all $t \geq t_0$ and $k \in S$, where B_s is definied in Lemma B.4.*

Proof. By Lemma B.4, we have for $k \in S_+$ (note that $(\boldsymbol{u}_{S_-})_k = 0$ in this case. The case $k \in S_-$ is similar, we omit for simplicity)

$$\begin{aligned}
w_k^{(t+1)} &= \left(1 - 2\eta\lambda\left(\frac{d}{n}v_k^{(t)} + \lambda(w_k^{(t)})^2 - \beta_k^* \pm (B_s + s\delta\mu)\right)\right) w_k^{(t)},\\
v_k^{(t+1)} &= v_k^{(t)} - \eta\left(\frac{d}{n}v_k^{(t)} + \lambda(w_k^{(t)})^2 - \beta_k^* \pm (B_s + s\delta\mu)\right).
\end{aligned}$$

Since $\lambda(w_k^{(t)})^2 = \beta_{\min}/4$ by Lemma B.5, we have

$$\begin{aligned}
\frac{d}{n}v_k^{(t+1)} + \lambda(w_k^{(t+1)})^2 =& \frac{d}{n}v_k^{(t)} - \frac{d}{n}\eta\left(\frac{d}{n}v_k^{(t)} + \lambda(w_k^{(t)})^2 - \beta_k^* \pm (B_s + s\delta\mu)\right) \\
&+ \left(1 - 2\eta\lambda\left(\frac{d}{n}v_k^{(t)} + \lambda(w_k^{(t)})^2 - \beta_k^* \pm (B_s + s\delta\mu)\right)\right)^2 \lambda(w_k^{(t)})^2 \\
\geq& \frac{d}{n}v_k^{(t)} + \lambda(w_k^{(t)})^2 - \eta\left(\frac{d}{n}v_k^{(t)} + \lambda(w_k^{(t)})^2 - \beta_k^* \pm \underbrace{(B_s + s\delta\mu)}_{=:\mathrm{err}}\right)\left(\frac{d}{n} + 4\lambda^2(w_k^{(t)})^2\right).
\end{aligned}$$

This implies for $t \geq t_0$, if $\frac{d}{n}v_k^{(t)} + \lambda(w_k^{(t)})^2 - \beta_k^* < -2\mathrm{err}$, we have

$$\begin{aligned}
\frac{d}{n}v_k^{(t+1)} + \lambda(w_k^{(t+1)})^2 - \beta_k^* &\geq \left(\frac{d}{n}v_k^{(t)} + \lambda(w_k^{(t)})^2 - \beta_k^*\right)\left(1 - \frac{\eta}{2}\left(\frac{d}{n} + 4\lambda^2(w_k^{(t)})^2\right)\right) \\
&\geq \left(\frac{d}{n}v_k^{(t)} + \lambda(w_k^{(t)})^2 - \beta_k^*\right)(1 - \Omega(\eta\lambda\beta_{\min}))
\end{aligned}$$

Thus, $\frac{d}{n}v_k^{(t)} + \lambda(w_k^{(t)})^2 - \beta_k^*$ will increase in this case. Therefore, we know $\frac{d}{n}v_k^{(t)} + \lambda(w_k^{(t)})^2 - \beta_k^* \geq -\max\{\mu, 2\mathrm{err}\} = -\max\{\mu, 2(B_s + s\delta\mu)\}$ for all $t \geq t_0$. Similarly, given η is small enough, we can also get a similar upper bound. Thus, we finish the proof. □

Lemma B.8. *Suppose Lemma B.1 hold. Recall*

$$T_1 := \inf\left\{t : \left\|\frac{d}{n}\boldsymbol{v}_S^{(t)} + \lambda(\boldsymbol{w}_{S_+}^{(t)})^2 - \lambda(\boldsymbol{u}_{S_-}^{(t)})^2 - \boldsymbol{\beta}^*\right\|_\infty \leq C_1(B_\xi + \sigma\sqrt{n/d})\right\},$$

where C_1 is a large enough universal constant. Then, we know $T_1 = O(\log(1/\alpha)/\eta\lambda\beta_{\min})$.

Proof. We can first use Lemma B.5 and then repeatedly using Lemma B.6 $\log(1/4B_s)$ times. We get within time $O(\log(1/\lambda\alpha^2(B_\xi + \sigma\sqrt{n/d}))/\eta\lambda\beta_{\min}) = O(\log(1/\alpha)/\eta\lambda\beta_{\min})$, $\left\|\frac{d}{n}\boldsymbol{v}_S^{(t)} + \lambda(\boldsymbol{w}_{S_+}^{(t)})^2 - \lambda(\boldsymbol{u}_{S_-}^{(t)})^2 - \boldsymbol{\beta}^*\right\|_\infty \leq 4B_s = C_1(B_\xi + \sigma\sqrt{n/d})$. □

B.5 Technical Lemmas

In this subsection, we collect several technical lemmas that are used in the proof.

Lemma B.9. *Suppose $z_{t+1} = (1 - \eta(z_t - \mu))^2 z_t$ with $\eta, \mu, z_0 > 0$ and $z_0 \leq \mu - \varepsilon$. Then if $\eta \leq \mu/2$, within time $T = O((1/\eta\mu)(\log(\mu/z_0) + \log(\mu/\varepsilon)))$ we have $|z_T - \mu| \leq \varepsilon$. Moreover, we have $|z_t - \mu| \leq \varepsilon$ for $t \geq T$.*

Proof. Denote $T_1 := \inf\{t : z_t \geq \mu/2\}$ and $T_2 := \inf\{t : |z_t - \mu| \leq \varepsilon\}$. We bound T_1 and $T_2 - T_1$ respectively in below.

For $t \leq T_1$, we have $z_{t+1} \geq (1 + \eta\mu/2)^2 z_t \geq (1 + \eta\mu/2)^{2t} z_0$. Therefore, $T_1 = O((1/\eta\mu)\log(\mu/z_0))$. For $T_1 \leq t \leq T_2$, we have $z_{t+1} \geq z_t - 2\eta(z_t - \mu)z_t \geq z_t - \eta(z_t - \mu)\mu$. This implies $z_{t+1} - \mu \geq (1 - \eta\mu)(z_t - \mu) \geq (1 - \eta\mu)^{t - T_1}(z_{T_1} - \mu)$. Therefore, $T_2 - T_2 = O((1/\eta\mu)\log(\mu/\varepsilon))$. Together we know $T = T_1 + T_2 = O((1/\eta\mu)(\log(\mu/z_0) + \log(\mu/\varepsilon)))$.

We then show once $|z_t - \mu| \leq \varepsilon$, it will stay close to μ. To see this, if $-\varepsilon \leq z_t - \mu < 0$, then from the above calculation we know $z_{t+1} - \mu \geq (1 - \eta\mu)(z_t - \mu) \geq -\varepsilon$. If $0 \leq z_t - \mu \leq \varepsilon$, then $z_{t+1} = (1 - \eta(z_t - \mu))^2 z_t \leq z_t \leq \mu + \varepsilon$. Therefore, we know $|z_t - \mu| \leq \varepsilon$ for $t \geq T_1$. □

Lemma B.10. *For $\alpha, \beta \in \mathbb{R}^d$, we have $\|\alpha \odot \beta\|_2 \leq \|\alpha\|_2 \|\beta\|_\infty$, $\left\|\alpha^{\odot k}\right\|_2 \leq \|\alpha\|_2^k$ for $k \geq 1$.*

Proof. We have

$$\|\alpha \odot \beta\|_2^2 = \sum_i \alpha_i^2 \beta_i^2 \le \|\alpha\|_2^2 \|\beta\|_\infty^2 ,$$

$$\left\|\alpha^{\odot k}\right\|_2^2 = \sum_i \alpha_i^{2k} = \|\alpha\|_{2k}^{2k} \le \|\alpha\|_2^{2k} .$$

□

C Proof for Stage 2

In Stage 2, we will show that the training loss goes to ε while the test loss $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2$ remains small. In particular, we will split into 2 sub-stages: in Stage 2.1, train loss decreases to $\|\boldsymbol{r}\|_2 = O(\sigma)$ (Lemma C.1), and in Stage 2.2 we use a NTK-type analysis (Lemma C.6). Note that it suffices to combine Lemma C.1 and Lemma C.6 to get Lemma 10.

Throughout Stage 2, we mostly rely on $\boldsymbol{v}_e$ to fit the noise in order to reduce the loss; at the same time, we show that the variables used in Stage 1 continue to fit the signal and all the other variables remain small. This can be done by an NTK-type analysis when the loss is very small. However, for the first part of Stage 2 we still need to track the dynamics of $\boldsymbol{v}$ and $\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{v}$ carefully.

C.1 Stage 2.1: train loss decreases to $\|r\|_2 = O(\sigma)$

Our goal in this stage is to show that the loss decreases to $O(\sigma^2)$ and that the non-signal entries remain small. We formalize this in the following main lemma.

Lemma C.1 (Stage 2.1)**.** *Let $T_{21} := \inf\{t : \left\|\boldsymbol{r}^{(t)}\right\|_2 \le C_{21}\sigma\}$ with large enough universal constant C_{21}. Then, we have $T_{21} - T_1 = O((n/\eta d)\log(n))$ and the following hold:*

- $\left\|\frac{d}{n}\boldsymbol{v}_S^{(T_{21})}\lambda \boldsymbol{w}_{S_+}^{(T_{21})\odot 2} - \lambda \boldsymbol{u}_{S_-}^{(T_{21})\odot 2} - \boldsymbol{\beta}^*\right\|_\infty = O(B_\xi + \sigma\sqrt{n/d})$
- $\left\|\boldsymbol{w}_{e_+}^{(T_{21})}\right\|_\infty, \left\|\boldsymbol{u}_{e_-}^{(T_{21})}\right\|_\infty = O(\alpha)$.
- $\left\|\boldsymbol{v}^{(T_{21})}\right\|_2 = O(\sigma\sqrt{n/d})$ *and* $\left\|\boldsymbol{v}_S^{(T_{21})}\right\|_2 = O(\sqrt{s}(n/d)\log^2(d)(B_\xi + \sigma\sqrt{n/d}))$.

To prove this, we will maintain the following inductive hypothesis, which shows the non-signal entries remain small. The overall strategy is to show that entries of $\boldsymbol{v}$ will allow us to fit the noise and hence reduce loss, and we do this by using a similar strategy to track the dynamics of $\boldsymbol{v}$ as in Stage 1.

Lemma C.2 (Inductive Hypothesis for Stage 2.1)**.** *For $T_1 \le t \le \widetilde{T}_{21} := T_1 + C_{T_{21}}(n\log(n)/\eta d)$ with a large enough universal constant $C_{T_{21}}$, we have the following hold:*

- $\left\|\frac{d}{n}\boldsymbol{v}_S^{(t)} + \lambda \boldsymbol{w}_{S_+}^{(t)\odot 2} - \lambda \boldsymbol{u}_{S_-}^{(t)\odot 2} - \boldsymbol{\beta}^*\right\|_\infty = O(B_\xi + \sigma\sqrt{n/d})$
- $\left\|\boldsymbol{w}_{e_+}^{(t)}\right\|_\infty, \left\|\boldsymbol{u}_{e_-}^{(t)}\right\|_\infty = O(\alpha)$.
- $\left\|\boldsymbol{v}_S^{(t)}\right\|_2 = O(\sqrt{s}(n/d)\log^2(d)(B_\xi + \sigma\sqrt{n/d}))$.
- $\left\|\boldsymbol{r}^{(t)}\right\|_2 = (1 - \Omega(\eta d/n))^{t-T_1} O(\sigma\sqrt{n})$.

In particular, the first point and third point imply that $\left\|\lambda \boldsymbol{w}_{S_+}^{(t)\odot 2} - \lambda \boldsymbol{u}_{S_-}^{(t)\odot 2} - \boldsymbol{\beta}^*\right\|_\infty = O(\sqrt{s}(B_\xi + \sigma\sqrt{n/d})\log^2(d))$. *The last point implies that* $T_{21} - T_1 = O((n/\eta d)\log(n))$. *Moreover, by the choice of parameters,* $O(d/\sqrt{n})\lambda\left\|\boldsymbol{w}_{e_+}^{(t)}\right\|_\infty^2 = O(B_\xi/\log d)$, $O(d/\sqrt{n})\lambda\left\|\boldsymbol{u}_{e_-}^{(t)}\right\|_\infty^2 = O(B_\xi/\log d)$.

C.1.1 Dynamics of v

As in Stage 1, we analyze the decomposition of $\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{v}/n$ and $\boldsymbol{v}$ separately. The proofs are very similar to Lemma B.2 and Lemma B.3 in Stage 1, but several terms will now have a tighter bound. We defer the proofs to Appendix C.1.4.

For the decomposition of $\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{v}/n$ we have

Lemma C.3. *Recall the decomposition in* (3)

$$\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{v}^{(t)} = \frac{d}{n}\boldsymbol{v}_S^{(t)} + b_t(\boldsymbol{X}^\top \boldsymbol{\xi})_e + \boldsymbol{\Gamma}_t,$$
$$b_{t+1} = b_t - \frac{\eta d}{n}(b_t - \frac{1}{n}),$$

where $\left\|\boldsymbol{\Gamma}^{(t)}\right\|_\infty \le \gamma_t$ *and recall the notation* $\boldsymbol{\beta}_S = \sum_{i:\beta_i^*\neq 0}\beta_i \boldsymbol{e}_i$, $\boldsymbol{\beta}_e = \sum_{i:\beta_i^*=0}\beta_i \boldsymbol{e}_i$. *Suppose Lemma C.2 holds. We have for* $T_1 \le t \le \widetilde{T}_{21}$

$$b_t = (1-(1-\eta d/n)^t)/n \le 1/n,$$
$$\gamma_t \le \gamma_{T_1} + O(\sigma\sqrt{d/n} + (dB_\xi/n\log d)\eta t) = O(\sigma\sqrt{n/d} + B_\xi).$$

For the decomposition of $\boldsymbol{v}$ we have

Lemma C.4. *Recall the decomposition in* (4)

$$\boldsymbol{v}^{(t)} = \boldsymbol{v}_S^{(t)} + a_t \boldsymbol{X}^\top \boldsymbol{\xi} + \boldsymbol{\Delta}_v^{(t)},$$
$$a_{t+1} = a_t - \eta(b_t - 1/n),$$

where $\left\|\boldsymbol{\Delta}_v^{(t)}\right\|_\infty \le \zeta_t$. *and recall the notation* $\boldsymbol{\beta}_S = \sum_{i:\beta_i^*\neq 0}\beta_i \boldsymbol{e}_i$, $\boldsymbol{\beta}_e = \sum_{i:\beta_i^*=0}\beta_i \boldsymbol{e}_i$. *Suppose Lemma C.2 holds. We have for* $T_1 \le t \le \widetilde{T}_{21}$

$$a_t = (1-(1-\eta d/n)^t)/d \le 1/d$$
$$\zeta_t = \zeta_{T_1} + O((B_\xi + \sigma\sqrt{n/d})\eta(t-T_1)) = O((B_\xi + \sigma\sqrt{n/d})n\log(n)/d).$$

In particular, we can show that $\left\|\boldsymbol{v}^{(t)}\right\|_2 = O(\sigma\sqrt{n/d})$.

C.1.2 Implications of Inductive Hypothesis Lemma C.2

Given the dynamics of $\boldsymbol{v}$, we now have the approximation of gradient by Lemma 6.

Lemma C.5. *In the setting of Lemma C.3 and Lemma C.4, we have for* $T_1 \le t \le \widetilde{T}_{21}$

$$\nabla_{\boldsymbol{w}} L = \left(\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{r}\right)\odot(2\lambda \boldsymbol{w}) = 2\lambda(\frac{d}{n}\boldsymbol{v}_S + \lambda \boldsymbol{w}_{S_+}^{\odot 2} - \lambda \boldsymbol{u}_{S_-}^{\odot 2} - \boldsymbol{\beta}^* + \boldsymbol{\Delta}_r)\odot \boldsymbol{w},$$
$$\nabla_{\boldsymbol{u}} L = -\left(\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{r}\right)\odot(2\lambda \boldsymbol{u}) = -2\lambda(\frac{d}{n}\boldsymbol{v}_S + \lambda \boldsymbol{w}_{S_+}^{\odot 2} - \lambda \boldsymbol{u}_{S_-}^{\odot 2} - \boldsymbol{\beta}^* + \boldsymbol{\Delta}_r)\odot \boldsymbol{u},$$
$$\nabla_v L = \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{r},$$

where

$$\left\|\boldsymbol{\Delta}_r^{(t)}\right\|_\infty = O\left(B_\xi + \sigma\sqrt{n/d}\right) + s\delta\left\|\frac{d}{n}\boldsymbol{v}_S^{(t)} + \lambda \boldsymbol{w}_{S_+}^{(t)\odot 2} - \lambda \boldsymbol{u}_{S_-}^{(t)\odot 2} - \boldsymbol{\beta}^*\right\|_\infty.$$

C.1.3 Proof of Inductive Hypothesis Lemma C.2 and Lemma C.1

Now we are ready to prove the induction hypothesis for Stage 2.1 and Lemma C.1.

Lemma C.2 (Inductive Hypothesis for Stage 2.1)**.** *For $T_1 \leq t \leq \widetilde{T}_{21} := T_1 + C_{T_{21}}(n\log(n)/\eta d)$ with a large enough universal constant $C_{T_{21}}$, we have the following hold:*

- $\left\|\frac{d}{n}\boldsymbol{v}_S^{(t)} + \lambda\boldsymbol{w}_{S_+}^{(t)\odot 2} - \lambda\boldsymbol{u}_{S_-}^{(t)\odot 2} - \boldsymbol{\beta}^*\right\|_\infty = O(B_\xi + \sigma\sqrt{n/d})$
- $\left\|\boldsymbol{w}_{e_+}^{(t)}\right\|_\infty, \left\|\boldsymbol{u}_{e_-}^{(t)}\right\|_\infty = O(\alpha)$.
- $\left\|\boldsymbol{v}_S^{(t)}\right\|_2 = O(\sqrt{s}(n/d)\log^2(d)(B_\xi + \sigma\sqrt{n/d}))$.
- $\left\|\boldsymbol{r}^{(t)}\right\|_2 = (1 - \Omega(\eta d/n))^{t-T_1} O(\sigma\sqrt{n})$.

In particular, the first point and third point imply that $\left\|\lambda\boldsymbol{w}_{S_+}^{(t)\odot 2} - \lambda\boldsymbol{u}_{S_-}^{(t)\odot 2} - \boldsymbol{\beta}^*\right\|_\infty = O(\sqrt{s}(B_\xi + \sigma\sqrt{n/d})\log^2(d))$. The last point implies that $T_{21} - T_1 = O((n/\eta d)\log(n))$. Moreover, by the choice of parameters, $O(d/\sqrt{n})\lambda\left\|\boldsymbol{w}_{e_+}^{(t)}\right\|_\infty^2 = O(B_\xi/\log d)$, $O(d/\sqrt{n})\lambda\left\|\boldsymbol{u}_{e_-}^{(t)}\right\|_\infty^2 = O(B_\xi/\log d)$.

Proof. We show these inductively on t. For $t = T_1$, we know it holds by Lemma 7. Suppose it holds before time t, then at time $t+1$ we will show it still hold.

For $\left\|\frac{d}{n}\boldsymbol{v}_S^{(t+1)} + \lambda\boldsymbol{w}_{S_+}^{(t+1)\odot 2} - \lambda\boldsymbol{u}_{S_-}^{(t+1)\odot 2} - \boldsymbol{\beta}^*\right\|_\infty$, let $k \in S_+$ (the case $k \in S_-$ can be handled similarly, we omit for simplicity). Since by the choice of parameter $(d/n)\left\|\boldsymbol{v}_S^{(t)}\right\|_\infty < \beta_{\min}/2$, we know $\lambda(w_k^{(t)})^2 = \beta_{\min}/4$. For $T_1 \leq t \leq \widetilde{T}_{21}$, by Lemma B.7 and Lemma C.5, we know $\left\|\frac{d}{n}\boldsymbol{v}^{(t+1)} + \lambda\boldsymbol{w}^{(t+1)\odot 2} - \lambda\boldsymbol{u}^{(t+1)\odot 2} - \boldsymbol{\beta}^*\right\|_\infty = O(B_\xi + \sigma\sqrt{n/d})$.

For $k \notin S$, consider w_k (u_k can be bounded similarly), we have the dynamics by Lemma C.5

$$w_k^{(t+1)} \leq \left(1 + 2\eta\lambda O(B_\xi + \sigma\sqrt{n/d})\right) w_k^{(t)}.$$

This means $|w_k^{(t)}| = (1 + O(\eta\lambda(B_\xi + \sigma\sqrt{n/d}))^{t-T_1} O(\alpha) = O(\alpha)$.

It remains to consider w_k with $k \in S_-$ and u_k with $k \in S_+$. We will focus on w_k with $k \in S_-$, the other follows the same calculation. Similar in the proof of Lemma B.1, we have

$$w_k^{(t+1)}u_k^{(t+1)} = \left(1 - 2\eta\lambda\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{r}^{(t)}\right)_k\right) w_k^{(t)} \cdot \left(1 + 2\eta\lambda\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{r}^{(t)}\right)_k\right) u_k^{(t)} \leq w_k^{(t)}u_k^{(t)} \leq \alpha^2.$$

We know $u_k^{(t)} \geq \alpha$. This implies that $|w_k^{(t)}| \leq \alpha$.

For $\|\boldsymbol{v}_S\|_2$, we have by Lemma C.5 and Lemma 7

$$\begin{aligned}\left\|\boldsymbol{v}_S^{(t+1)}\right\|_2 \leq \left\|\boldsymbol{v}_S^{(t)}\right\|_2 + \eta\left\|\frac{1}{n}(\boldsymbol{X}^\top\boldsymbol{r}^{(t)})_S\right\|_2 \leq \left\|\boldsymbol{v}_S^{(T_1)}\right\|_2 + O(\sqrt{s}(B_\xi + \sigma\sqrt{n/d})\eta(t - T_1))\\ =& O(\sqrt{s}(n/d)\log^2(d)(B_\xi + \sigma\sqrt{n/d}))\end{aligned}$$

For the bound on $\left\|\boldsymbol{r}^{(t+1)}\right\|_2$, using the same calculation as in the proof of Lemma B.1, we can show it is true. □

Given the above induction hypothesis, we are ready to prove the main result for Stage 2.1.

Lemma C.1 (Stage 2.1)**.** *Let $T_{21} := \inf\{t : \left\|\boldsymbol{r}^{(t)}\right\|_2 \leq C_{21}\sigma\}$ with large enough universal constant C_{21}. Then, we have $T_{21} - T_1 = O((n/\eta d)\log(n))$ and the following hold:*

- $\left\|\frac{d}{n}\boldsymbol{v}_S^{(T_{21})}\lambda\boldsymbol{w}_{S_+}^{(T_{21})\odot 2}-\lambda\boldsymbol{u}_{S_-}^{(T_{21})\odot 2}-\boldsymbol{\beta}^*\right\|_\infty = O(B_\xi+\sigma\sqrt{n/d})$
- $\left\|\boldsymbol{w}_{e_+}^{(T_{21})}\right\|_\infty, \left\|\boldsymbol{u}_{e_-}^{(T_{21})}\right\|_\infty = O(\alpha)$.
- $\left\|\boldsymbol{v}^{(T_{21})}\right\|_2 = O(\sigma\sqrt{n/d})$ *and* $\left\|\boldsymbol{v}_S^{(T_{21})}\right\|_2 = O(\sqrt{s}(n/d)\log^2(d)(B_\xi+\sigma\sqrt{n/d}))$.

Proof. The first two points and the bound on $T_{21}-T_1$ follow from Lemma C.2. The last point follow from Lemma C.4 and Lemma C.2. □

C.1.4 Omitted Proofs in Section C.1.1 and Section C.1.2

In this subsection, we give the proof of Lemma C.3, Lemma C.4 and Lemma C.5.

Lemma C.3. *Recall the decomposition in* (3)

$$\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}^{(t)} = \frac{d}{n}\boldsymbol{v}_S^{(t)} + b_t(\boldsymbol{X}^\top\boldsymbol{\xi})_e + \boldsymbol{\Gamma}_t,$$
$$b_{t+1} = b_t - \frac{\eta d}{n}(b_t - \frac{1}{n}),$$

where $\left\|\boldsymbol{\Gamma}^{(t)}\right\|_\infty \le \gamma_t$ *and recall the notation* $\boldsymbol{\beta}_S = \sum_{i:\beta_i^*\neq 0}\beta_i\boldsymbol{e}_i$, $\boldsymbol{\beta}_e = \sum_{i:\beta_i^*=0}\beta_i\boldsymbol{e}_i$. *Suppose Lemma C.2 holds. We have for* $T_1 \le t \le \widetilde{T}_{21}$

$$b_t = (1-(1-\eta d/n)^t)/n \le 1/n,$$
$$\gamma_t \le \gamma_{T_1} + O(\sigma\sqrt{d/n} + (dB_\xi/n\log d)\eta t) = O(\sigma\sqrt{n/d}+B_\xi).$$

Proof. The proof here is almost the same as in the proof of Lemma B.2 in Stage 1. The only difference is that we know have better bounds on the error terms. We first write the update of b_t and $\boldsymbol{\Gamma}_t$ using the update of $\boldsymbol{v}$.

$$\begin{aligned}
b_{t+1}(\boldsymbol{X}^\top\boldsymbol{\xi})_e + \boldsymbol{\Gamma}_{t+1} =& \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}^{(t+1)} - \frac{d}{n}\boldsymbol{v}_S^{(t+1)}\\
=& \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}^{(t)} - \frac{d}{n}\boldsymbol{v}_S^{(t)} - \eta\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{r}^{(t)} + \eta\frac{d}{n}\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{r}^{(t)}\right)_S\\
=& b_t(\boldsymbol{X}^\top\boldsymbol{\xi})_e + \boldsymbol{\Gamma}_t - \frac{\eta}{n^2}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{r}^{(t)} + \eta\frac{d}{n}\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{r}^{(t)}\right)_S\\
=& b_t(\boldsymbol{X}^\top\boldsymbol{\xi})_e + \boldsymbol{\Gamma}_t - \frac{\eta}{n^2}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top - d\boldsymbol{I})\boldsymbol{r}^{(t)} - \eta\frac{d}{n}\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{r}^{(t)}\right)_e.
\end{aligned}$$

We bound the last two terms one by one. For $\frac{\eta}{n^2}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top - d\boldsymbol{I})\boldsymbol{r}^{(t)}$, we have by Assumption 1 and Lemma C.2

$$\left\|\frac{\eta}{n^2}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top - d\boldsymbol{I})\boldsymbol{r}^{(t)}\right\|_\infty \le \frac{\eta}{n}O(\frac{1}{\sqrt{n}}\cdot\sqrt{dn})(1-\Omega(\eta d/n))^{t-T_1}O(\sigma\sqrt{n}) = O(\eta\sigma\sqrt{d/n})(1-\Omega(\eta d/n))^{t-T_1}.$$

For $\eta\frac{d}{n}\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{r}^{(t)}\right)_e$, we have

$$\begin{aligned}
\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{r}^{(t)}\right)_e =& \left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}^{(t)} + \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}(\lambda\boldsymbol{w}_{S_+}^{(t)\odot 2} - \lambda\boldsymbol{u}_{S_-}^{(t)\odot 2} - \boldsymbol{\beta}^*) - \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\xi} + \frac{1}{n}\boldsymbol{X}^\top(\lambda\boldsymbol{X}\boldsymbol{w}_{e_+}^{(t)\odot 2} - \lambda\boldsymbol{X}\boldsymbol{u}_{e_-}^{(t)\odot 2})\right)_e\\
=& \left(\frac{d}{n}\boldsymbol{v}_S^{(t)} + \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}(\lambda\boldsymbol{w}_{S_+}^{(t)\odot 2} - \lambda\boldsymbol{u}_{S_-}^{(t)\odot 2} - \boldsymbol{\beta}^*) + (b_t - \frac{1}{n})\boldsymbol{X}^\top\boldsymbol{\xi} + \boldsymbol{\Gamma}_t + \frac{1}{n}\boldsymbol{X}^\top(\lambda\boldsymbol{X}\boldsymbol{w}_{e_+}^{(t)\odot 2} - \lambda\boldsymbol{X}\boldsymbol{u}_{e_-}^{(t)\odot 2})\right)_e\\
=& (b_t - \frac{1}{n})(\boldsymbol{X}^\top\boldsymbol{\xi})_e + \left((\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X} - \boldsymbol{I})(\lambda\boldsymbol{w}_{S_+}^{(t)\odot 2} - \lambda\boldsymbol{u}_{S_-}^{(t)\odot 2} - \boldsymbol{\beta}^*) + \boldsymbol{\Gamma}_t + \frac{1}{n}\boldsymbol{X}^\top(\lambda\boldsymbol{X}\boldsymbol{w}_{e_+}^{(t)\odot 2} - \lambda\boldsymbol{X}\boldsymbol{u}_{e_-}^{(t)\odot 2})\right)_e.
\end{aligned}$$

Therefore, we know by Lemma C.2

$$
\begin{aligned}
b_{t+1} &= b_t - \frac{\eta d}{n}(b_t - \frac{1}{n}),\\
\gamma_{t+1} &\le \gamma_t + (1 - O(\eta d/n))^{t-T_1} O(\eta\sigma\sqrt{d/n}) + \eta\frac{d}{n} O(B_\xi/\log d + (d/\sqrt{n})\lambda\alpha^2)\\
&= \gamma_t + (1 - O(\eta d/n))^{t-T_1} O(\eta\sigma\sqrt{d/n}) + \eta O(dB_\xi/n\log d).
\end{aligned}
$$

By Lemma B.2, this implies

$$
\begin{aligned}
&b_t = (1-\eta d/n)^{t-T_1} b_{T_1} + (1-(1-\eta d/n)^{t-T_1})/n = (1-(1-\eta d/n)^t)/n \le 1/n,\\
&\gamma_t \le \gamma_{T_1} + O(\sigma\sqrt{n/d} + (dB_\xi/n\log d)\eta(t-T_1)) = O(\sigma\sqrt{n/d} + B_\xi).
\end{aligned}
$$

□

Lemma C.4. *Recall the decomposition in* (4)

$$
\begin{aligned}
\boldsymbol{v}^{(t)} &= \boldsymbol{v}_S^{(t)} + a_t \boldsymbol{X}^\top \boldsymbol{\xi} + \boldsymbol{\Delta}_v^{(t)},\\
a_{t+1} &= a_t - \eta(b_t - 1/n),
\end{aligned}
$$

where $\left\|\boldsymbol{\Delta}_v^{(t)}\right\|_\infty \le \zeta_t$. *and recall the notation* $\boldsymbol{\beta}_S = \sum_{i:\beta_i^*\neq 0}\beta_i \boldsymbol{e}_i$, $\boldsymbol{\beta}_e = \sum_{i:\beta_i^* = 0}\beta_i \boldsymbol{e}_i$. *Suppose Lemma C.2 holds. We have for* $T_1 \le t \le \widetilde{T}_{21}$

$$
\begin{aligned}
a_t &= (1-(1-\eta d/n)^t)/d \le 1/d\\
\zeta_t &= \zeta_{T_1} + O((B_\xi + \sigma\sqrt{n/d})\eta(t-T_1)) = O((B_\xi + \sigma\sqrt{n/d})n\log(n)/d).
\end{aligned}
$$

In particular, we can show that $\|\boldsymbol{v}^{(t)}\|_2 = O(\sigma\sqrt{n/d})$.

Proof. The proof here is almost the same as in the proof of Lemma B.3 in Stage 1. The only difference is that we know have better bounds on the error terms. We write the update of a_t and $\boldsymbol{\Delta}_v^{(t)}$ using the update of $\boldsymbol{v}$

$$
\begin{aligned}
a_{t+1}\boldsymbol{X}^\top\boldsymbol{\xi} + \boldsymbol{\Delta}_v^{(t+1)} &= \boldsymbol{v}^{(t+1)} - \boldsymbol{v}_S^{(t+1)} = \boldsymbol{v}^{(t)} - \boldsymbol{v}_S^{(t)} - \eta\left(\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{r}^{(t)}\right)_e\\
&= a_t\boldsymbol{X}^\top\boldsymbol{\xi} + \boldsymbol{\Delta}_v^{(t)} - \eta\left(\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{r}^{(t)}\right)_e.
\end{aligned}
$$

For $\left(\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{r}^{(t)}\right)_e$, using the decomposition of $\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}/n$ in Lemma C.3, we have

$$
\begin{aligned}
\left(\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{r}^{(t)}\right)_e &= \left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}^{(t)} + \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}(\lambda \boldsymbol{w}_{S_+}^{(t)\odot 2} - \lambda \boldsymbol{u}_{S_-}^{(t)\odot 2} - \boldsymbol{\beta}^*) - \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\xi} + \frac{1}{n}\boldsymbol{X}^\top(\lambda\boldsymbol{X}\boldsymbol{w}_{e_+}^{(t)\odot 2} - \lambda\boldsymbol{X}\boldsymbol{u}_{e_-}^{(t)\odot 2})\right)_e\\
&= \left(\frac{d}{n}\boldsymbol{v}_S^{(t)} + \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}(\lambda \boldsymbol{w}_{S_+}^{(t)\odot 2} - \lambda \boldsymbol{u}_{S_-}^{(t)\odot 2} - \boldsymbol{\beta}^*) + (b_t - \frac{1}{n})\boldsymbol{X}^\top\boldsymbol{\xi} + \boldsymbol{\Gamma}_t + \frac{1}{n}\boldsymbol{X}^\top(\lambda\boldsymbol{X}\boldsymbol{w}_{e_+}^{(t)\odot 2} - \lambda\boldsymbol{X}\boldsymbol{u}_{e_-}^{(t)\odot 2})\right)_e\\
&= (b_t - \frac{1}{n})(\boldsymbol{X}^\top\boldsymbol{\xi})_e + \left((\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X} - \boldsymbol{I})(\lambda \boldsymbol{w}_{S_+}^{(t)\odot 2} - \lambda \boldsymbol{u}_{S_-}^{(t)\odot 2} - \boldsymbol{\beta}^*) + \boldsymbol{\Gamma}_t + \frac{1}{n}\boldsymbol{X}^\top(\lambda\boldsymbol{X}\boldsymbol{w}_{e_+}^{(t)\odot 2} - \lambda\boldsymbol{X}\boldsymbol{u}_{e_-}^{(t)\odot 2})\right)_e.
\end{aligned}
$$

Therefore, we have the update of a_t and ζ_t by using Lemma 5, Assumption 1 and Lemma C.2

$$
\begin{aligned}
a_{t+1} &= a_t - \eta(b_t - 1/n),\\
\zeta_{t+1} &\le \zeta_t + \eta O(|nb_t - 1|B_\xi + B_\xi/\log d + \sigma\sqrt{n/d} + B_\xi + (d/\sqrt{n})\lambda\alpha^2).
\end{aligned}
$$

By Lemma B.3, this implies

$$a_t = \eta t/n - \eta \sum_{\tau<t} b_\tau = (1-(1-\eta d/n)^t)/d \le 1/d$$
$$\zeta_t \le \zeta_{T_1} + O((B_\xi + \sigma\sqrt{n/d})\eta(t-T_1)) = O((B_\xi + \sigma\sqrt{n/d}) n\log(n)/d).$$

We now bound $\|\boldsymbol{v}\|_2$. Since its gradient norm $\|\nabla_{\boldsymbol{v}} L\|_2 = \left\|\boldsymbol{X}^\top \boldsymbol{r}/n\right\|_2 \le (1-\Omega(\eta d/n))^{t-T_1} O(\sigma\sqrt{d/n})$ by Lemma C.2 and Assumption 1, we can bound $\left\|\boldsymbol{v}^{(t)}\right\|_2 \le \left\|\boldsymbol{v}^{(T_1)}\right\|_2 + \eta\sum_{T_1\le\tau\le t}\left\|\nabla_{\boldsymbol{v}} L^{(\tau)}\right\|_2 = \left\|\boldsymbol{v}^{(T_1)}\right\|_2 + O(\sigma\sqrt{n/d}) = O(\sigma\sqrt{n/d})$.

□

Lemma C.5. *In the setting of Lemma C.3 and Lemma C.4, we have for* $T_1 \le t \le \widetilde{T}_{21}$

$$\nabla_{\boldsymbol{w}} L = \left(\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{r}\right) \odot (2\lambda \boldsymbol{w}) = 2\lambda(\frac{d}{n}\boldsymbol{v}_S + \lambda \boldsymbol{w}_{S_+}^{\odot 2} - \lambda \boldsymbol{u}_{S_-}^{\odot 2} - \boldsymbol{\beta}^* + \boldsymbol{\Delta}_r) \odot \boldsymbol{w},$$
$$\nabla_{\boldsymbol{u}} L = -\left(\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{r}\right) \odot (2\lambda \boldsymbol{u}) = -2\lambda(\frac{d}{n}\boldsymbol{v}_S + \lambda \boldsymbol{w}_{S_+}^{\odot 2} - \lambda \boldsymbol{u}_{S_-}^{\odot 2} - \boldsymbol{\beta}^* + \boldsymbol{\Delta}_r) \odot \boldsymbol{u},$$
$$\nabla_v L = \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{r},$$

where

$$\left\|\boldsymbol{\Delta}_r^{(t)}\right\|_\infty = O\left(B_\xi + \sigma\sqrt{n/d}\right) + s\delta\left\|\frac{d}{n}\boldsymbol{v}_S^{(t)} + \lambda \boldsymbol{w}_{S_+}^{(t)\odot 2} - \lambda \boldsymbol{u}_{S_-}^{(t)\odot 2} - \boldsymbol{\beta}^*\right\|_\infty.$$

Proof. By Lemma C.3 and Lemma C.4 and the choice of parameter, the result directly follows from Lemma 6.

□

C.2 Stage 2.2

After Stage 2.1, the loss is already very small. This allows us to further tighten the bound of several terms and use an NTK-type analysis to show that the parameters do not move much while reduce the training loss to ε.

Lemma C.6. *Let* $T_{22} := \inf\{t : L(\boldsymbol{u}^{(t)}, \boldsymbol{w}^{(t)}, \boldsymbol{v}^{(t)}) = \left\|\boldsymbol{r}^{(t)}\right\|^2/n \le \varepsilon\}$. *Then* $T_{22} - T_{21} = O(n\log(\sigma/\varepsilon)/\eta d)$ *and the following hold:*

- $\left\|\frac{d}{n}\boldsymbol{v}_S^{(T_{22})} + \lambda \boldsymbol{w}_{S_+}^{(T_{22})\odot 2} - \lambda \boldsymbol{u}_{S_-}^{(T_{22})\odot 2} - \boldsymbol{\beta}^*\right\|_\infty = O(B_\xi + \sigma\sqrt{n/d})$
- $\left\|\boldsymbol{w}_{e_+}^{(T_{22})}\right\|_\infty, \left\|\boldsymbol{u}_{e_-}^{(T_{22})}\right\|_\infty = O(\alpha)$
- $\|\boldsymbol{v}^{(T_{22})}\|_2 = O(\sigma\sqrt{n/d})$, $\left\|\boldsymbol{v}_S^{(T_{22})}\right\|_2 = O(\sqrt{s}(n/d)\log^2(d)(B_\xi + \sigma\sqrt{n/d}))$
- $\|\boldsymbol{r}^{(t)}\|_2 = (1-\Omega(\eta d/n))^{t-T_{21}} O(\sigma)$

In particular, the above imply that $\left\|\boldsymbol{\beta}^{(T_{22})} - \boldsymbol{\beta}^*\right\|_2 = O(\sqrt{s}\log^2(d)(B_\xi + \sigma\sqrt{n/d}))$. *Moreover, for every* $t \ge T_{22}$*, the above still hold and train loss* $L^{(t)} \le \varepsilon$.

Proof. We show these by induction. At $t = T_{21}$, we know they hold by Lemma C.1. Suppose before time t they hold, then at time $t+1$ we know $\left\|\boldsymbol{X}^\top \boldsymbol{r}^{(\tau)}/n\right\|_\infty = (1-\Omega(\eta d/n))^{\tau - T_{21}} O(\sigma/\sqrt{n})$ for any $\tau \le t$ by Assumption 1.

For $\left\|\frac{d}{n}\boldsymbol{v}_S^{(t+1)} + \lambda \boldsymbol{w}_{S_+}^{(t+1)\odot 2} - \lambda \boldsymbol{u}_{S_-}^{(t+1)\odot 2} - \boldsymbol{\beta}^*\right\|_\infty$, consider the k-th entry with $k \in S_+$ ($k \in S_-$ can be bounded similarly). The proof is similar to the proof in Lemma C.2, we omit for simplicity.

We now consider $\left\|\boldsymbol{w}_{e_+}^{(t+1)}\right\|_\infty$ and $\left\|\boldsymbol{u}_{e_-}^{(t+1)}\right\|_\infty$. For $k \notin S$, consider w_k (u_k can be bounded similarly)

$$\begin{aligned}
|w_k^{(t+1)}| \leq & \left(1 + 2\lambda\eta \left\|\boldsymbol{X}^\top \boldsymbol{r}^{(t)}/n\right\|_\infty\right) w_k^{(t)} \\
\leq & \prod_{T_{21}\leq\tau\leq t} \left(1 + (1-\Omega(\eta d/n))^{\tau - T_{21}} O(\eta\lambda\sigma/\sqrt{n})\right) O(\alpha) \\
\leq & \left(1 + \sum_{T_{21}\leq\tau\leq t} (1-\Omega(\eta d/n))^{\tau - T_{21}} O(\eta\lambda\sigma/\sqrt{n})\right) O(\alpha) \\
\leq & O(\alpha + \lambda\sigma\sqrt{n}\alpha/d) = O(\alpha),
\end{aligned}$$

where in the second to last line we use the fact that $\prod_i(1+q_i) = e^{\sum_i \ln(1+q_i)} \leq e^{\sum_i q_i} \leq 1 + O(\sum_i q_i)$ for $\sum_i q_i = O(1)$.

It remains to consider w_k with $k \in S_-$ and u_k with $k \in S_+$. We will focus on w_k with $k \in S_-$, the other follows the same calculation. Similar in the proof of Lemma B.1, we have

$$w_k^{(t+1)} u_k^{(t+1)} = \left(1 - 2\eta\lambda \left(\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{r}^{(t)}\right)_k\right) w_k^{(t)} \cdot \left(1 + 2\eta\lambda \left(\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{r}^{(t)}\right)_k\right) u_k^{(t)} \leq w_k^{(t)} u_k^{(t)} \leq \alpha^2.$$

We know $u_k^{(t)} \geq \alpha$. This implies that $|w_k^{(t)}| \leq \alpha$.

For $\|\boldsymbol{r}\|_2$, we can bound it the same as in the proof of Lemma B.1.

For $\|\boldsymbol{v}_S\|$, we have by Lemma C.1

$$\begin{aligned}
\left\|\boldsymbol{v}_S^{(t+1)}\right\|_2 \leq \left\|\boldsymbol{v}_S^{(t)}\right\|_2 + \eta \left\|\frac{1}{n}(\boldsymbol{X}^\top \boldsymbol{r}^{(t)})_S\right\|_2 \leq & \left\|\boldsymbol{v}_S^{(T_1)}\right\|_2 + \sum_{T_{21}\leq\tau\leq t} (1-\Omega(\eta d/n))^{\tau - T_{21}} O(\eta\sigma/\sqrt{n}) \\
= & O(\sqrt{s}(n/d)\log^2(d)(B_\xi + \sigma\sqrt{n/d}))
\end{aligned}$$

For $\|\boldsymbol{v}\|_2$, we have

$$\left\|\boldsymbol{v}^{(t+1)} - \boldsymbol{v}^{(T_{21})}\right\|_2 \leq \eta \sum_{T_{21}\leq\tau\leq t} \left\|\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{r}^{(\tau)}\right\|_2 \leq \sum_{T_{21}\leq\tau\leq t} (1-\Omega(\eta d/n))^{\tau - T_{21}} O(\eta\sigma/\sqrt{n}) = O(\sigma\sqrt{n}/d).$$

Note that $\left\|\boldsymbol{v}^{(T_{21})}\right\|_2 = O(\sigma\sqrt{n/d})$, thus we have $\left\|\boldsymbol{v}^{(t+1)}\right\|_2 = O(\sigma\sqrt{n/d})$.

In this way, we finish the induction proof. It remains to bound $T_{22} - T_{21}$. Given $\left\|\boldsymbol{r}^{(t)}\right\|_2 = (1 - \Omega(\eta d/n))^t O(\sigma)$, we know $T_{22} - T_{21} = O(n\log(\sigma/\varepsilon)/\eta d)$. Moreover, we can see in the above proof that it will still hold after T_{22}, thanks to the geometric decreasing of $\|\boldsymbol{r}\|_2$. □

D Proof of main result Theorem 3

In this section, we give the proof of main result. Given that we have already characterized the training dynamics to the convergence in Stage 1 and Stage 2, it immediately follows from the results for Stage 1 (Lemma 7) and Stage 2 (Lemma 10).

Theorem 3 (Main result)**.** *Under Assumption 1, suppose there exists constant C such that $\sigma \leq C$. We train model* (1) *with initialization $\boldsymbol{v}^{(0)} = \mathbf{0}$, $\boldsymbol{w}^{(0)} = \boldsymbol{u}^{(0)} = \alpha\mathbf{1}$ and follow the gradient descent update* (2)*. If $\widetilde{\Omega}(s) \leq n \leq \widetilde{O}\left(\min\{d/s, d^{2/3}\}\right)$ and we choose $\lambda = \Theta\left(d/\sigma n(\sqrt{\log(d)/n} + \sqrt{n/d})\log(n)\right)$, $\alpha = 1/\mathrm{poly}(d)$, $\eta \leq O(\sqrt{n/sd}/\lambda^3)$, then for every $t \geq T = O(\log(n/\alpha\varepsilon)n/\eta d)$ with any given $\varepsilon > 0$ we have training loss $L(\boldsymbol{u}^{(t)}, \boldsymbol{w}^{(t)}, \boldsymbol{v}^{(t)}) \leq \varepsilon$ and test loss*

$$\left\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\right\|_2 = O\left(\sqrt{s}\log^2(d)\left(\sigma\sqrt{\frac{\log(d)}{n}} + \sigma\sqrt{\frac{n}{d}}\right)\right).$$

Proof. First note that Lemma 10 follows from the Lemma C.1 for Stage 2.1 and Lemma C.6 for Stage 2.2. Then, it suffices to combine Lemma 7 and Lemma 10 in Section 5, since

$$\begin{aligned}\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2 &\le \left\|\frac{d}{n}\boldsymbol{v}_S+\lambda\boldsymbol{w}_{S_+}^{\odot 2}-\lambda\boldsymbol{u}_{S_-}^{\odot 2}-\boldsymbol{\beta}^*\right\|_2+\left\|\frac{d}{n}\boldsymbol{v}_S\right\|_2+\|\boldsymbol{v}\|_2+\left\|\lambda\boldsymbol{w}_{e_+}^{\odot 2}-\lambda\boldsymbol{u}_{e_-}^{\odot 2}\right\|_2\\ &= O\left(\sqrt{s}\log(n)\log(n/\alpha)\left(\sigma\sqrt{\frac{\log(d)}{n}}+\sigma\sqrt{\frac{n}{d}}\right)\right)\end{aligned}$$

□

E Synthetic Experiments

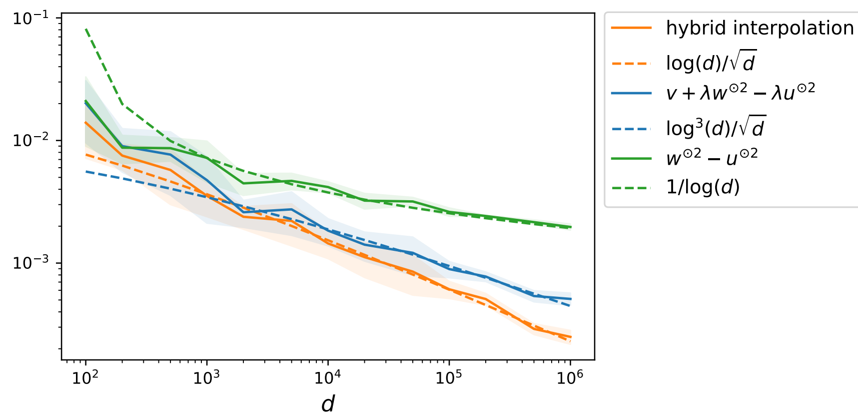

Figure 2: Test loss vs. dimension d when fixing the ratio $d/n=\sqrt{d}/3$ for 3 different interpolating method: hybrid interpolation with Lasso (Muthukumar et al., 2020), model $\boldsymbol{v}+\lambda\boldsymbol{w}^{\odot 2}-\lambda\boldsymbol{u}^{\odot 2}$ as we focused in the paper and model $\boldsymbol{w}^{\odot 2}-\boldsymbol{u}^{\odot 2}$ that only keeps the second order term. Solid lines represent the mean and shaded regions represent the standard deviation of test loss during 3 experiments. Dashed lines represent the corresponding order.

In this section, we run synthetic experiments to verify our theoretical results. We choose d from 100 to 10^6 and set $n=3\sqrt{d}$. The target $\boldsymbol{\beta}^*=(1/\sqrt{3},-1/\sqrt{3},1/\sqrt{3},0,\dots,0)^\top$, data $\boldsymbol{x}_i\sim N(\boldsymbol{0},\boldsymbol{I})$ sampled from Gaussian distribution and noise level $\sigma=0.1$. We compare 3 different interpolation method:

- hybrid interpolation (Muthukumar et al., 2020): As a 2-step procedure, we first use Lasso (implemented in `sklearn`) with ℓ_1 regularization coefficient on the order of $\Theta(\sigma\sqrt{\log(d)/n})$ (Theorems 7.13 and 7.20 in Wainwright (2019)). We choose the coefficient with the best test loss among the choice of $\{1/10,1/5,1/2,1,2,5,10\}*\sigma\sqrt{\log(d)/n}$. In the second step, we use the min-ℓ_2-norm interpolator to fit the residual.
- Model $\boldsymbol{v}+\lambda\boldsymbol{w}^{\odot 2}-\lambda\boldsymbol{u}^{\odot 2}$: As suggested in our main result, we initialize $\boldsymbol{v}=0$ and $\boldsymbol{w}=\boldsymbol{u}=\alpha\boldsymbol{1}$ with $\alpha=10^{-10}$. We set $\lambda=100d/\sigma n\log(n)(\sqrt{\log(d)/n}+\sqrt{n/d})$ and run gradient descent with stepsize $\eta=10^{-6}$ until training loss reaches 10^{-4}.
- Model $\boldsymbol{w}^{\odot 2}-\boldsymbol{u}^{\odot 2}$: We use small initialization that sets $\boldsymbol{w}=\boldsymbol{u}=\alpha\boldsymbol{1}$ with $\alpha=10^{-15}$. We run gradient descent with stepsize $\eta=10^{-6}$ until training loss reaches 10^{-4}.

Our results are shown in Figure 2. We can see that with fixed ratio $d/n=\sqrt{d}/3$, as d increases, the test loss of different method decreases with different rate. The hybrid interpolation gives the smallest test loss and our learner model $\boldsymbol{v} + \lambda \boldsymbol{w}^{\odot 2} - \lambda \boldsymbol{u}^{\odot 2}$ gives a similar performance. This agrees with what our theoretical result suggests. The model $\boldsymbol{w}^{\odot 2} - \boldsymbol{u}^{\odot 2}$ that only uses second-order term performs worse than others. This is expected as we know such parametrization with small initialization converges to min-ℓ_1-norm interpolator (Woodworth et al., 2020), and min-ℓ_1-norm interpolator gives large test loss $\Omega(\sigma^2/\log(d/n))$ in the sparse noisy regression setting (Chatterji and Long, 2022; Wang et al., 2022).