Recovery Guarantees for One-hidden-layer Neural Networks

Kai Zhong†
zhongkai@ices.utexas.edu
UT-Austin

Zhao Song‡
zhaos@utexas.edu
UT-Austin

Prateek Jain
prajain@microsoft.com
Microsoft Research, India

Peter L. Bartlett§
bartlett@cs.berkeley.edu
UC, Berkeley

Inderjit S. Dhillon¶
derjerit@cs.utexas.edu
UT-Austin

Abstract

In this paper, we consider regression problems with one-hidden-layer neural networks (1NNs). We distill some properties of activation functions that lead to local strong convexity in the neighborhood of the ground-truth parameters for the 1NN squared-loss objective and most popular nonlinear activation functions satisfy the distilled properties, including rectified linear units (ReLUs), leaky ReLUs, squared ReLUs and sigmoids. For activation functions that are also smooth, we show local linear convergence guarantees of gradient descent under a resampling rule. For homogeneous activations, we show tensor methods are able to initialize the parameters to fall into the local strong convexity region. As a result, tensor initialization followed by gradient descent is guaranteed to recover the ground truth with sample complexity $d \cdot \log(1/\epsilon) \cdot \text{poly}(k, \lambda)$ and computational complexity $n \cdot d \cdot \text{poly}(k, \lambda)$ for smooth homogeneous activations with high probability, where $d$ is the dimension of the input, $k$ ($k \leq d$) is the number of hidden nodes, $\lambda$ is a conditioning property of the ground-truth parameter matrix between the input layer and the hidden layer, $\epsilon$ is the targeted precision and $n$ is the number of samples. To the best of our knowledge, this is the first work that provides recovery guarantees for 1NNs with both sample complexity and computational complexity linear in the input dimension and logarithmic in the precision.

*Supported in part by NSF grants CCF-1320746, IIS-1546452 and CCF-1564000, and part of the work was done while interning in Microsoft research, India.

†Supported in part by UTCS TAship (CS361 Spring 17 Introduction to Computer Security).

‡Supported by Australian Research Council through an Australian Laureate Fellowship (FL110100281) and through the Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS), and NSF grants IIS-1619362.

¶Supported in part by NSF grants CCF-1320746, IIS-1546452 and CCF-1564000.

*A preliminary version of this paper appears in Proceedings of the Thirty-fourth International Conference on Machine Learning (ICML 2017).
## Contents

1 Introduction .................................................. 3

2 Related Work .................................................. 4
   2.1 Expressive Power ......................................... 4
   2.2 Achievability of Global Optima ....................... 4
   2.3 Generalization Bound / Recovery Guarantees ......... 5

3 Problem Formulation ........................................... 6

4 Positive Definiteness of Hessian ............................. 7

5 Tensor Methods for Initialization ........................... 9
   5.1 Preliminary .............................................. 9
   5.2 Algorithm ............................................ 10
   5.3 Theoretical Analysis .................................... 11

6 Global Convergence ........................................... 12

7 Numerical Experiments ........................................ 12

8 Conclusion ..................................................... 13

References ...................................................... 14

A Notation ....................................................... 19

B Preliminaries .................................................. 19
   B.1 Useful Facts ........................................... 19
   B.2 Matrix Bernstein ....................................... 22

C Properties of Activation Functions ......................... 25

D Local Positive Definiteness of Hessian ................. 26
   D.1 Main Results for Positive Definiteness of Hessian .... 26
      D.1.1 Bounding the Spectrum of the Hessian near the Ground Truth .... 26
      D.1.2 Local Linear Convergence of Gradient Descent ............ 27
   D.2 Positive Definiteness of Population Hessian at the Ground Truth .. 28
      D.2.1 Lower Bound on the Eigenvalues of Hessian for the Orthogonal Case ... 29
      D.2.2 Lower Bound on the Eigenvalues of Hessian for Non-orthogonal Case ... 33
      D.2.3 Upper Bound on the Eigenvalues of Hessian for Non-orthogonal Case ... 37
   D.3 Error Bound of Hessians near the Ground Truth for Smooth Activations ... 38
      D.3.1 Second-order Smoothness near the Ground Truth for Smooth Activations ... 38
      D.3.2 Empirical and Population Difference for Smooth Activations ........ 42
   D.4 Error Bound of Hessians near the Ground Truth for Non-smooth Activations ... 48
   D.5 Positive Definiteness for a Small Region .............. 51
E Tensor Methods
E.1 Tensor Initialization Algorithm .................................................. 55
E.2 Main Result for Tensor Methods .................................................. 57
E.3 Error Bound for the Subspace Spanned by the Weight Matrix .............. 59
   E.3.1 Error Bound for the Second-order Moment in Different Cases .......... 59
   E.3.2 Error Bound for the Second-order Moment ................................ 63
   E.3.3 Subspace Estimation Using Power Method .................................. 63
E.4 Error Bound for the Reduced Third-order Moment ............................... 66
   E.4.1 Error Bound for the Reduced Third-order Moment in Different Cases .... 66
   E.4.2 Final Error Bound for the Reduced Third-order Moment ................. 70
E.5 Error Bound for the Magnitude and Sign of the Weight Vectors .......... 71
   E.5.1 Robustness for Solving Linear Systems .................................... 71
   E.5.2 Error Bound for the First-order Moment ................................... 71
   E.5.3 Linear System for the First-order Moment .................................. 73
   E.5.4 Linear System for the Second-order Moment ............................... 74

F Acknowledgments ............................. 75
1 Introduction

Neural Networks (NNs) have achieved great practical success recently. Many theoretical contributions have been made very recently to understand the extraordinary performance of NNs. The remarkable results of NNs on complex tasks in computer vision and natural language processing inspired works on the expressive power of NNs [CSS16, CS16, RPK +16, DFS16, PLR +16, MPCB14, Tel16]. Indeed, several works found NNs are very powerful and the deeper the more powerful. However, due to the high non-convexity of NNs, knowing the expressivity of NNs doesn’t guarantee that the targeted functions will be learned. Therefore, several other works focused on the achievability of global optima. Many of them considered the over-parameterized setting, where the global optima or local minima close to the global optima will be achieved when the number of parameters is large enough, including [FB16, HV15, LSSS14, DPG +14, SS16, HM17]. This, however, leads to overfitting easily and can’t provide any generalization guarantees, which are actually the essential goal in most tasks.

A few works have considered generalization performance. For example, [XLS17] provide generalization bound under the Rademacher generalization analysis framework. Recently [ZBH +17] describe some experiments showing that NNs are complex enough that they actually memorize the training data but still generalize well. As they claim, this cannot be explained by applying generalization analysis techniques, like VC dimension and Rademacher complexity, to classification loss (although it does not rule out a margins analysis—see, for example, [Bar98]; their experiments involve the unbounded cross-entropy loss).

In this paper, we don’t develop a new generalization analysis. Instead we focus on parameter recovery setting, where we assume there are underlying ground-truth parameters and we provide recovery guarantees for the ground-truth parameters up to equivalent permutations. Since the parameters are exactly recovered, the generalization performance will also be guaranteed.

Several other techniques are also provided to recover the parameters or to guarantee generalization performance, such as tensor methods [JSA15] and kernel methods [AGMR17]. These methods require sample complexity $O(d^3)$ or computational complexity $\tilde{O}(n^2)$, which can be intractable in practice.

Recently [Sha16] show that neither specific assumptions on the niceness of the input distribution or niceness of the target function alone is sufficient to guarantee learnability using gradient-based methods. In this paper, we assume data points are sampled from Gaussian distribution and the parameters of hidden neurons are linearly independent.

Our main contributions are as follows,

1. We distill some properties for activation functions, which are satisfied by a wide range of activations, including ReLU, squared ReLU, sigmoid and tanh. With these properties we show positive definiteness (PD) of the Hessian in the neighborhood of the ground-truth parameters given enough samples (Theorem 4.2). Further, for activations that are also smooth, we show local linear convergence is guaranteed using gradient descent.

2. We propose a tensor method to initialize the parameters such that the initialized parameters fall into the local positive definiteness area. Our contribution is that we reduce the sample/computational complexity from cubic dependency on dimension to linear dependency (Theorem 5.6).

3. Combining the above two results, we provide a globally converging algorithm (Algorithm 2) for smooth homogeneous activations satisfying the distilled properties. The whole procedure requires sample/computational complexity linear in dimension and logarithmic in precision (Theorem 6.1).
2 Related Work

The recent empirical success of NNs has boosted their theoretical analyses [FZK+16, Bal16, BMBY16, SBL16, APVZ14, AGMR17, GKKT17]. In this paper, we classify them into three main directions.

2.1 Expressive Power

Expressive power is studied to understand the remarkable performance of neural networks on complex tasks. Although one-hidden-layer neural networks with sufficiently many hidden nodes can approximate any continuous function [Hor91], shallow networks can’t achieve the same performance in practice as deep networks. Theoretically, several recent works show the depth of NNs plays an essential role in the expressive power of neural networks [DFS16]. As shown in [CSS16, CS16, Tel16], functions that can be implemented by a deep network of polynomial size require exponential size in order to be implemented by a shallow network. [RPK+16, PLR+16, MPCB14, AGMR17] design some measures of expressivity that display an exponential dependence on the depth of the network. However, the increasing of the expressivity of NNs or its depth also increases the difficulty of the learning process to achieve a good enough model. In this paper, we focus on 1NNs and provide recovery guarantees using a finite number of samples.

2.2 Achievability of Global Optima

The global convergence is in general not guaranteed for NNs due to their non-convexity. It is widely believed that training deep models using gradient-based methods works so well because the error surface either has no local minima, or if they exist they need to be close in value to the global minima. [SCP16] present examples showing that for this to be true additional assumptions on the data, initialization schemes and/or the model classes have to be made. Indeed the achievability of global optima has been shown under many different types of assumptions.

In particular, [CHM+15] analyze the loss surface of a special random neural network through spin-glass theory and show that it has exponentially many local optima, whose loss is small and close to that of a global optimum. Later on, [Kaw16] eliminate some assumptions made by [CHM+15] but still require the independence of activations as [CHM+15], which is unrealistic. [SS16] study the geometric structure of the neural network objective function. They have shown that with high probability random initialization will fall into a basin with a small objective value when the network is over-parameterized. [LSSS14] consider polynomial networks where the activations are square functions, which are typically not used in practice. [HV15] show that when a local minimum has zero parameters related to a hidden node, a global optimum is achieved. [FB16] study the landscape of 1NN in terms of topology and geometry, and show that the level set becomes connected as the network is increasingly over-parameterized. [HM17] show that products of matrices don’t have spurious local minima and that deep residual networks can represent any function on a sample, as long as the number of parameters is larger than the sample size. [SC16] consider over-specified NNs, where the number of samples is smaller than the number of weights. [DPG+14] propose a new approach to second-order optimization that identifies and attacks the saddle point problem in high-dimensional non-convex optimization. They apply the approach to recurrent neural networks and show practical performance. [AGMR17] use results from tropical geometry to show global optimality of an algorithm, but it requires $(2n)^k \poly(n)$ computational complexity.

Almost all of these results require the number of parameters is larger than the number of points, which probably overfits the model and no generalization performance will be guaranteed. In this paper, we propose an efficient and provable algorithm for 1NNs that can achieve the underlying ground-truth parameters.
2.3 Generalization Bound / Recovery Guarantees

The achievability of global optima of the objective from the training data doesn’t guarantee the learned model to be able to generalize well on unseen testing data. In the literature, we find three main approaches to generalization guarantees.

1) Use generalization analysis frameworks, including VC dimension/Rademacher complexity, to bound the generalization error. A few works have studied the generalization performance for NNs. [XLS17] follow [SC16] but additionally provide generalization bounds using Rademacher complexity. They assume the obtained parameters are in a regularization set so that the generalization performance is guaranteed, but this assumption can’t be justified theoretically. [HRS16] apply stability analysis to the generalization analysis of SGD for convex and non-convex problems, arguing early stopping is important for generalization performance.

2) Assume an underlying model and try to recover this model. This direction is popular for many non-convex problems including matrix completion/sensing [JNS13, Har14, SL15, BLWZ17], mixed linear regression [ZJD16], subspace recovery [EV09] and other latent models [AGH+14].

Without making any assumptions, those non-convex problems are intractable [AGKM12, GV15, SWZ17a, GG11, SR11, HM13, AGM12, YCS14]. Recovery guarantees for NNs also need assumptions. Several different approaches under different assumptions are provided to have recovery guarantees on different NN settings.

Tensor methods [AGH+14, WTSA15, WA16, SWZ16] are a general tool for recovering models with latent factors by assuming the data distribution is known. Some existing recovery guarantees for NNs are provided by tensor methods [SA15, JSA15]. However, [SA15] only provide guarantees to recover the subspace spanned by the weight matrix and no sample complexity is given, while [JSA15] require $O(d^3/\epsilon^2)$ sample complexity. In this paper, we use tensor methods as an initialization step so that we don’t need very accurate estimation of the moments, which enables us to reduce the total sample complexity from $1/\epsilon^2$ to $\log(1/\epsilon)$.

[ABGM14] provide polynomial sample complexity and computational complexity bounds for learning deep representations in unsupervised setting, and they need to assume the weights are sparse and randomly distributed in $[-1, 1]$.

[Tia17] analyze 1NN by assuming Gaussian inputs in a supervised setting, in particular, regression and classification with a teacher. This paper also considers this setting. However, there are some key differences. a) [Tia17] require the second-layer parameters are all ones, while we can learn these parameters. b) In [Tia17], the ground-truth first-layer weight vectors are required to be orthogonal, while we only require linear independence. c) [Tia17] require a good initialization but doesn’t provide initialization methods, while we show the parameters can be efficiently initialized by tensor methods. d) In [Tia17], only the population case (infinite sample size) is considered, so there is no sample complexity analysis, while we show finite sample complexity.

Recovery guarantees for convolution neural network with Gaussian inputs are provided in [BG17], where they show a globally converging guarantee of gradient descent on a one-hidden-layer no-overlap convolution neural network. However, they consider population case, so no sample complexity is provided. Also their analysis depends on ReLU activations and the no-overlap case is very unlikely to be used in practice. In this paper, we consider a large range of activation functions, but for one-hidden-layer fully-connected NNs.

3) Improper Learning. In the improper learning setting for NNs, the learning algorithm is not restricted to output a NN, but only should output a prediction function whose error is not much larger than the error of the best NN among all the NNs considered. [ZLJ16, ZLW16] propose kernel methods to learn the prediction function which is guaranteed to have generalization performance close to that of the NN. However, the sample complexity and computational complexity are expo-
We consider the following regression problem. Given a set of \( n \) samples \( S = \{(x_1, y_1), (x_2, y_2), \cdots (x_n, y_n)\} \subseteq \mathbb{R}^d \times \mathbb{R}, \)
let \( D \) denote a underlying distribution over \( \mathbb{R}^d \times \mathbb{R} \) with parameters
\[
\{w_1^*, w_2^*, \cdots w_k^*\} \subseteq \mathbb{R}^d, \quad \{v_1^*, v_2^*, \cdots, v_k^*\} \subseteq \mathbb{R}
\]
such that each sample \( (x, y) \in S \) is sampled i.i.d. from this distribution, with
\[
D : \quad x \sim \mathcal{N}(0, I), \quad y = \sum_{i=1}^{k} v_i^* \cdot \phi(w_i^* \top x),
\]
where \( \phi(z) \) is the activation function, \( k \) is the number of nodes in the hidden layer. The main question we want to answer is: How many samples are sufficient to recover the underlying parameters?

It is well-known that, training one hidden layer neural network is NP-complete [BR88]. Thus, without making any assumptions, learning deep neural network is intractable. Throughout the paper, we assume \( x \) follows a standard normal distribution; the data is noiseless; the dimension of input data is at least the number of hidden nodes; and activation function \( \phi(z) \) satisfies some reasonable properties.

Actually our results can be easily extended to multivariate Gaussian distribution with positive definite covariance and zero mean since we can estimate the covariance first and then transform the input to a standard normal distribution but with some loss of accuracy. Although this paper focuses on the regression problem, we can transform classification problems to regression problems if a good teacher is provided as described in [Tia17]. Our analysis requires \( k \) to be no greater than \( d \), since the first-layer parameters will be linearly dependent otherwise.

For activation function \( \phi(z) \), we assume it is continuous and if it is non-smooth let its first derivative be left derivative. Furthermore, we assume it satisfies Property 3.1, 3.2, and 3.3. These properties are critical for the later analyses. We also observe that most activation functions actually satisfy these three properties.

**Property 3.1.** The first derivative \( \phi'(z) \) is nonnegative and homogeneously bounded, i.e., \( 0 \leq \phi'(z) \leq L_1|z|^p \) for some constants \( L_1 > 0 \) and \( p \geq 0 \).

**Property 3.2.** Let \( \alpha_q(\sigma) = \mathbb{E}_{z \sim \mathcal{N}(0, 1)}[\phi'(z \cdot \sigma)]^{q}, \forall q \in \{0, 1, 2\} \), and \( \beta_q(\sigma) = \mathbb{E}_{z \sim \mathcal{N}(0, 1)}[\phi''(z \cdot \sigma)]^{q}, \forall q \in \{0, 2\} \). Let \( \rho(\sigma) \) denote \( \min\{\beta_0(\sigma) - \alpha_0^2(\sigma) - \alpha_1^2(\sigma), \beta_0(\sigma) - \alpha_0^2(\sigma) - \alpha_2^2(\sigma), \alpha_0(\sigma) \cdot \alpha_2(\sigma) - \alpha_1^2(\sigma)\} \) The first derivative \( \phi'(z) \) satisfies that, for all \( \sigma > 0 \), we have \( \rho(\sigma) > 0 \).
**Property 3.3.** The second derivative $\phi''(z)$ is either (a) globally bounded $|\phi''(z)| \leq L_2$ for some constant $L_2$, i.e., $\phi(z)$ is $L_2$-smooth, or (b) $\phi''(z) = 0$ except for $c$ (a finite constant) points.

**Remark 3.4.** The first two properties are related to the first derivative $\phi'(z)$ and the last one is about the second derivative $\phi''(z)$. At high level, Property 3.1 requires $\phi$ to be non-decreasing with homogeneously bounded derivative; Property 3.2 requires $\phi$ to be highly non-linear; Property 3.3 requires $\phi$ to be either smooth or piece-wise linear.

**Theorem 3.5.** ReLU $\phi(z) = \max\{z, 0\}$, leaky ReLU $\phi(z) = \max\{z, 0.01z\}$, squared ReLU $\phi(z) = \max\{z, 0\}^2$ and any non-linear non-decreasing smooth functions with bounded symmetric $\phi'(z)$, like the sigmoid function $\phi(z) = 1/(1 + e^{-z})$, the tanh function and the erf function $\phi(z) = \int_0^z e^{-t^2} dt$, satisfy Property 3.1,3.2,3.3. The linear function, $\phi(z) = z$, doesn’t satisfy Property 3.2 and the quadratic function, $\phi(z) = z^2$, doesn’t satisfy Property 3.1 and 3.2.

The proof can be found in Appendix C.

## 4 Positive Definiteness of Hessian

In this section, we study the Hessian of empirical risk near the ground truth. We consider the case when $\nu^*$ is already known. Note that for homogeneous activations, we can assume $v_i^* \in \{-1, 1\}$ since $v_i^*(z) = \frac{1}{2p} \phi(|z|^{1/p})$, where $p$ is the degree of homogeneity. As $v_i^*$ only takes discrete values for homogeneous activations, in the next section, we show we can exactly recover $\nu^*$ using tensor methods with finite samples.

For a set of samples $S$, we define the Empirical Risk,

$$\hat{f}_S(W) = \frac{1}{2|S|} \sum_{(x,y) \in S} \left( \sum_{i=1}^k v_i^* \phi(w_i^\top x) - y \right)^2. \tag{2}$$

For a distribution $D$, we define the Expected Risk,

$$f_D(W) = \frac{1}{2} \mathbb{E}_{(x,y) \sim D} \left[ \left( \sum_{i=1}^k v_i^* \phi(w_i^\top x) - y \right)^2 \right]. \tag{3}$$

Let’s calculate the gradient and the Hessian of $\hat{f}_S(W)$ and $f_D(W)$. For each $j \in [k]$, the partial gradient of $f_D(W)$ with respect to $w_j$ can be represented as

$$\frac{\partial f_D(W)}{\partial w_j} = \mathbb{E}_{(x,y) \sim D} \left[ \left( \sum_{i=1}^k v_i^* \phi(w_i^\top x) - y \right) v_j^* \phi'(w_j^\top x)x \right].$$

For each $j, l \in [k]$ and $j \neq l$, the second partial derivative of $f_D(W)$ for the $(j,l)$-th off-diagonal block is,

$$\frac{\partial^2 f_D(W)}{\partial w_j \partial w_l} = \mathbb{E}_{(x,y) \sim D} \left[ v_j^* v_l^* \phi'(w_j^\top x) \phi'(w_l^\top x)xx^\top \right],$$

and for each $j \in [k]$, the second partial derivative of $f_D(W)$ for the $j$-th diagonal block is

$$\frac{\partial^2 f_D(W)}{\partial w_j^2} = \mathbb{E}_{(x,y) \sim D} \left[ \left( \sum_{i=1}^k v_i^* \phi(w_i^\top x) - y \right) v_j^* \phi''(w_j^\top x)xx^\top + (v_j^* \phi'(w_j^\top x))^2 xx^\top \right].$$
If \( \phi(z) \) is non-smooth, we use the Dirac function and its derivatives to represent \( \phi''(z) \). Replacing the expectation \( \mathbb{E}_{(x,y) \sim D} \) by the average over the samples \( |S|^{-1} \sum_{(x,y) \in S} \), we obtain the Hessian of the empirical risk.

Considering the case when \( W = W^* \in \mathbb{R}^{d \times k} \), for all \( j, l \in [k] \), we have,

\[
\frac{\partial^2 f_D(W^*)}{\partial w_j \partial w_l} = \mathbb{E}_{(x,y) \sim D} \left[ v_j^* v_l^* \phi'(w_j^T x) \phi'(w_l^T x) xx^T \right].
\]

If Property 3.3(b) is satisfied, \( \phi''(z) = 0 \) almost surely. So in this case the diagonal blocks of the empirical Hessian can be written as,

\[
\frac{\partial^2 f_S(W)}{\partial w_j^2} = \frac{1}{|S|} \sum_{(x,y) \in S} (v_j^* \phi'(w_j^T x))^2 xx^T.
\]

Now we show the Hessian of the objective near the global optimum is positive definite.

**Definition 4.1.** Given the ground truth matrix \( W^* \in \mathbb{R}^{d \times k} \), let \( \sigma_i(W^*) \) denote the \( i \)-th singular value of \( W^* \), often abbreviated as \( \sigma_i \). Let \( \kappa = \sigma_1/\sigma_k \), \( \lambda = (\prod_{i=1}^{k}\sigma_i)/\sigma_k^k \). Let \( v_{\max} \) denote \( \max_{i \in [k]} |v_i^*| \) and \( v_{\min} \) denote \( \min_{i \in [k]} |v_i^*| \). Let \( \nu = v_{\max}/v_{\min} \). Let \( \rho \) denote \( \rho(\sigma_k) \). Let \( \tau = (3\sigma_1/2)^{k/\nu} / \min_{i \in [1/2,\tau]} \rho(\sigma_i) \).

**Theorem 4.2** (Informal version of Theorem D.1). For any \( W \in \mathbb{R}^{d \times k} \) with \( \|W - W^*\| \leq \text{poly}(1/k, 1/\nu, 1/\rho(\sigma_k^2/p)) \cdot ||W^*|| \), let \( S \) denote a set of i.i.d. samples from distribution \( D \) (defined in (1)) and let the activation function satisfy Property 3.1,3.2,3.3. Then for any \( t \geq 1 \), if \( |S| \geq d \cdot \text{poly}(\log d, t, k, \nu, \tau, \lambda, \sigma_1^{2p}/\rho) \), we have with probability at least \( 1 - d^{-\Omega(t)} \),

\[
\Omega(v_{\min}^2 \rho(\sigma_k)/(\kappa^2 \lambda)) I \preceq \nabla^2 f_S(W) \preceq O(kv_{\max}^2 \sigma_1^{2p})I.
\]

**Remark 3.3.** As we can see from Theorem 4.2, \( \rho(\sigma_k) \) from Property 3.2 plays an important role for positive definite (PD) property. Interestingly, many popular activations, like ReLU, sigmoid and tanh, have \( \rho(\sigma_k) > 0 \), while some simple functions like linear \( (\phi(z) = z) \) and square \( (\phi(z) = z^2) \) functions have \( \rho(\sigma_k) = 0 \) and their Hessians are rank-deficient. Another important numbers are \( \kappa \) and \( \lambda \), two different condition numbers of the weight matrix, which directly influences the positive definiteness. If \( W^* \) is rank deficient, \( \lambda \to \infty, \kappa \to \infty \) and we don’t have PD property. In the best case when \( W^* \) is orthogonal, \( \lambda = \kappa = 1 \). In the worse case, \( \lambda \) can be exponential in \( k \). Also \( W \) should be close enough to \( W^* \). In the next section, we provide tensor methods to initialize \( w_i^* \) and \( v_i^* \) such that they satisfy the conditions in Theorem 4.2.

For the PD property to hold, we need the samples to be independent of the current parameters. Therefore, we need to do resampling at each iteration to guarantee the convergence in iterative algorithms like gradient descent. The following theorem provides the linear convergence guarantee of gradient descent for smooth activations.

**Theorem 4.4** (Linear convergence of gradient descent, informal version of Theorem D.2). Let \( W \) be the current iterate satisfying \( \|W - W^*\| \leq \text{poly}(1/\nu, 1/k, 1/\lambda, \rho(\sigma_k^2)) \cdot ||W^*|| \). Let \( S \) denote a set of i.i.d. samples from distribution \( D \) (defined in (1)) with \( |S| \geq d \cdot \text{poly}(\log d, t, k, \nu, \tau, \lambda, \sigma_1^{2p}/\rho) \) and let the activation function satisfy Property 3.1,3.2 and 3.3. Define \( \eta_0 := \Theta(v_{\min}^2 \rho(\sigma_k)/(\kappa^2 \lambda)) \) and \( M_0 := \Theta(kv_{\max}^2 \sigma_1^{2p}) \). If we perform gradient descent step size \( 1/M_0 \) on \( f_S(W) \) and obtain the next iterate,

\[
\bar{W} = W - \frac{1}{M_0} \nabla f_S(W),
\]
then with probability at least \(1 - d^{-\Omega(i)}\),
\[
\|\hat{W} - W^*\|_F^2 \leq (1 - \frac{m_0}{M_0})\|W - W^*\|_F^2.
\]

We provide the proofs in the Appendix D.1

5 Tensor Methods for Initialization

In this section, we show that Tensor methods can recover the parameters \(W^*\) to some precision and exactly recover \(v^*\) for homogeneous activations.

It is known that most tensor problems are NP-hard [Hås90, HL13] or even hard to approximate [SWZ17b]. However, by making some assumptions, tensor decomposition method becomes efficient [AGH+14, WTSA15, WA16, SWZ16]. Here we utilize the noiseless assumption and Gaussian inputs assumption to show a provable and efficient tensor methods.

5.1 Preliminary

Let’s define a special outer product \(\otimes\) for simplification of the notation. If \(v \in \mathbb{R}^d\) is a vector and \(I\) is the identity matrix, then \(v \otimes I = \sum_{j=1}^{d} [v \otimes e_j] e_j\). If \(M\) is a symmetric rank-\(r\) matrix factorized as \(M = \sum_{i=1}^{r} s_i v_i v_i^\top\) and \(I\) is the identity matrix, then
\[
M \otimes I = \sum_{i=1}^{r} s_i \sum_{j=1}^{d} \sum_{l=1}^{6} A_{i,j,l},
\]
where \(A_{1,i,j} = v_i \otimes v_j \otimes e_i \otimes e_j\), \(A_{2,i,j} = v_i \otimes e_j \otimes v_j \otimes e_i\), \(A_{3,i,j} = e_j \otimes v_i \otimes v_j \otimes e_i\), \(A_{4,i,j} = e_i \otimes e_j \otimes v_i \otimes v_j\), \(A_{5,i,j} = e_j \otimes v_i \otimes e_j \otimes v_i\) and \(A_{6,i,j} = e_j \otimes e_j \otimes v_i \otimes v_i\).

Denote \(\overline{w} = w/\|w\|\). Now let’s calculate some moments.

**Definition 5.1.** We define \(M_1, M_2, M_3, M_4\) and \(m_{1,i}, m_{2,i}, m_{3,i}, m_{4,i}\) as follows:
\[
M_1 = E_{(x,y) \sim D}[y \cdot x],
\]
\[
M_2 = E_{(x,y) \sim D}[y \cdot (x \otimes x - I)],
\]
\[
M_3 = E_{(x,y) \sim D}[y \cdot (x \otimes^3 - x \otimes I)],
\]
\[
M_4 = E_{(x,y) \sim D}[y \cdot (x \otimes^4 - (x \otimes x) \otimes I + I \otimes I)].
\]
\[
\gamma_j(\sigma) = E_{z \sim N(0,1)}[\phi(\sigma \cdot z)^j], \quad \forall j = 0, 1, 2, 3, 4.
\]
\[
m_{1,i} = \gamma_1(\|w_i^\ast\|),
\]
\[
m_{2,i} = \gamma_2(\|w_i^\ast\|) - \gamma_0(\|w_i^\ast\|).
\]
\[
m_{3,i} = \gamma_3(\|w_i^\ast\|) - 3\gamma_1(\|w_i^\ast\|).
\]
\[
m_{4,i} = \gamma_4(\|w_i^\ast\|) + 3\gamma_0(\|w_i^\ast\|) - 6\gamma_2(\|w_i^\ast\|).
\]

According to Definition 5.1, we have the following results,

**Claim 5.2.** For each \(j \in [4]\), \(M_j = \sum_{i=1}^{k} v_i^\ast m_{j,i} \overline{w}_i^\otimes j\).

Note that some \(m_{j,i}\)’s will be zero for specific activations. For example, for activations with symmetric first derivatives, i.e., \(\phi'(z) = \phi'(-z)\), like sigmoid and erf, we have \(\phi(z) + \phi(-z)\) being a constant and \(M_2 = 0\) since \(\gamma_0(\sigma) = \gamma_2(\sigma)\). Another example is ReLU. ReLU functions have vanishing \(M_3\), i.e., \(M_3 = 0\), as \(\gamma_3(\sigma) = 3\gamma_1(\sigma)\). To make tensor methods work, we make the following assumption.
Assumption 5.3. Assume the activation function $\phi(z)$ satisfies the following conditions:
1. If $M_j \neq 0$, then $m_{j,i} \neq 0$ for all $i \in [k]$.
2. At least one of $M_3$ and $M_4$ is non-zero.
3. If $M_1 = M_2 = 0$, then $\phi(z)$ is an even function, i.e., $\phi(z) = \phi(-z)$.
4. If $M_2 = M_4 = 0$, then $\phi(z)$ is an odd function, i.e., $\phi(z) = -\phi(-z)$.

If $\phi(z)$ is an odd function then $\phi(z) = -\phi(-z)$ and $v\phi(w^\top x) = -v\phi(-w^\top x)$. Hence we can always assume $v > 0$. If $\phi(z)$ is an even function, then $v\phi(w^\top x) = v\phi(-w^\top x)$. So if $w$ recovers $w^*$ then $-w$ also recovers $w^*$. Note that ReLU, leaky ReLU and squared ReLU satisfy Assumption 5.3. We further define the following non-zero moments.

Definition 5.4. Let $\alpha \in \mathbb{R}^d$ denote a randomly picked vector. We define $P_2$ and $P_3$ as follows: $P_2 = M_{j_2}(I, I, \alpha, \ldots, \alpha)$, where $j_2 = \min\{j \geq 2 | M_j \neq 0\}$ and $P_3 = M_{j_3}(I, I, I, \alpha, \ldots, \alpha)$, where $j_3 = \min\{j \geq 3 | M_j \neq 0\}$.

According to Definition 5.1 and 5.4, we have,

Claim 5.5. $P_2 = \sum_{i=1}^{k} v_i^* m_{j_2,i}(\alpha^\top m_i^*)^{j_2-2} m_i^* \otimes 2$ and $P_3 = \sum_{i=1}^{k} v_i^* m_{j_3,i}(\alpha^\top m_i^*)^{j_3-3} m_i^* \otimes 3$.

In other words for the above definition, $P_2$ is equal to the first non-zero matrix in the ordered sequence $\{M_2, M_3(I, I, \alpha), M_4(I, I, \alpha, \alpha)\}$. $P_3$ is equal to the first non-zero tensor in the ordered sequence $\{M_3, M_4(I, I, I, \alpha)\}$. Since $\alpha$ is randomly picked up, $w_i^* \alpha \neq 0$ and we view this number as a constant throughout this paper. So by construction and Assumption 5.3, both $P_2$ and $P_3$ are rank-$k$. Also, let $\hat{P}_2 \in \mathbb{R}^{d \times d}$ and $\hat{P}_3 \in \mathbb{R}^{d \times d \times d}$ denote the corresponding empirical moments of $P_2 \in \mathbb{R}^{d \times d}$ and $P_3 \in \mathbb{R}^{d \times d \times d}$ respectively.

5.2 Algorithm

Now we briefly introduce how to use a set of samples with size linear in dimension to recover the ground truth parameters to some precision. As shown in the previous section, we have a rank-$k$ 3rd-order moment $P_3$ that has tensor decomposition formed by $\{w_1^*, w_2^*, \ldots, w_k^*\}$. Therefore, we can use the non-orthogonal decomposition method [KCL15] to decompose the corresponding estimated tensor $\hat{P}_3$ and obtain an approximation of the parameters. The precision of the obtained parameters depends on the estimation error of $P_3$, which requires $\Omega(d^3/\epsilon^2)$ samples to achieve $\epsilon$ error. Also, the time complexity for tensor decomposition on a $d \times d \times d$ tensor is $\Omega(d^6)$.

In this paper, we reduce the cubic dependency of sample/computational complexity in dimension [JSA15] to linear dependency. Our idea follows the techniques used in [ZJD16], where they first used a 2nd-order moment $P_2$ to approximate the subspace spanned by $\{w_1^*, \ldots, w_k^*\}$, denoted as $V$, then use $V$ to reduce a higher-dimensional third-order tensor $P_3 \in \mathbb{R}^{d \times d \times d}$ to a lower-dimensional tensor $R_3 := P_3(V, V, V) \in \mathbb{R}^{k \times k \times k}$. Since the tensor decomposition and the tensor estimation are conducted on a lower-dimensional $\mathbb{R}^{k \times k \times k}$ space, the sample complexity and computational complexity are reduced.

The detailed algorithm is shown in Algorithm 1. First, we randomly partition the dataset into three subsets each with size $\tilde{O}(d)$. Then apply the power method on $\hat{P}_2$, which is the estimation of $P_2$ from $S_2$, to estimate $V$. After that, the non-orthogonal tensor decomposition (KCL)[KCL15] on $R_3$ outputs $\hat{u}_i$ which estimates $s_i V^\top w_i^*$ for $i \in [k]$ with unknown sign $s_i \in \{-1, 1\}$. Hence $w_i^*$ can be estimated by $s_i V \hat{u}_i$. Finally we estimate the magnitude of $w_i^*$ and the signs $s_i, v_i^*$ in the RecMagSign function for homogeneous activations. We discuss the details of each procedure and provide PowerMethod and RecMagSign algorithms in Appendix E.
Algorithm 1 Initialization via Tensor Method

1: procedure INITIALIZATION($S$) \[\triangleright\text{Theorem 5.6}\]
2: $S_2, S_3, S_4 \leftarrow \text{PARTITION}(S, 3)$
3: $\hat{P}_2 \leftarrow E_{S_2}[P_2]$
4: $V \leftarrow \text{POWERMETHOD}(\hat{P}_2, k)$
5: $\hat{R}_3 \leftarrow E_{S_3}[P_3(V, V, V)]$
6: $\{\hat{u}_i\}_{i \in [k]} \leftarrow \text{KCL}(\hat{R}_3)$
7: $\{w_i^{(0)}, v_i^{(0)}\}_{i \in [k]} \leftarrow \text{RECMAGSIGN}(V, \{\hat{u}_i\}_{i \in [k]}, S_4)$
8: return $\{w_i^{(0)}, v_i^{(0)}\}_{i \in [k]}$
9: end procedure

Algorithm 2 Globally Converging Algorithm

1: procedure LEARNING1NN($S, d, k, \epsilon$) \[\triangleright\text{Theorem 6.1}\]
2: $T \leftarrow \log(1/\epsilon) \cdot \text{poly}(k, \nu, \lambda, \sigma_1^2 / \rho)$.
3: $\eta \leftarrow 1/(k v_{\text{max}}^2 \sigma_1^2)$.
4: $S_0, S_1, \cdots, S_q \leftarrow \text{PARTITION}(S, q + 1)$.
5: $W^{(0)}, v^{(0)} \leftarrow \text{INITIALIZATION}(S_0)$.
6: Set $v_i^{*} \leftarrow v_i^{(0)}$ in Eq. (2) for all $f_{S_q}(W)$, $q \in [T]$.
7: for $q = 0, 1, 2, \cdots, T - 1$ do
8:     $W^{(q+1)} = W^{(q)} - \eta \nabla f_{S_{q+1}}(W^{(q)})$
9: end for
10: return $\{w_i^{(T)}, v_i^{(0)}\}_{i \in [k]}$
11: end procedure

(a) Sample complexity for recovery \hspace{1cm} (b) Tensor initialization error \hspace{1cm} (c) Objective v.s. iterations

Figure 1: Numerical Experiments

5.3 Theoretical Analysis

We formally present our theorem for Algorithm 1, and provide the proof in the Appendix E.2.

**Theorem 5.6.** Let the activation function be homogeneous satisfying Assumption 5.3. For any $0 < \epsilon < 1$ and $t \geq 1$, if $|S| \geq \epsilon^{-2} \cdot d \cdot \text{poly}(t, k, \nu, \log d)$, then there exists an algorithm (Algorithm 1) that takes $|S| k \cdot O(d)$ time and outputs a matrix $W^{(0)} \in \mathbb{R}^{d \times k}$ and a vector $v^{(0)} \in \mathbb{R}^k$ such that, with probability at least $1 - d^{-\Omega(t)}$,

$$
\|W^{(0)} - W^*\|_F \leq \epsilon \cdot \text{poly}(k, \nu)\|W^*\|_F, \quad \text{and} \quad v_i^{(0)} = v_i^*.
$$
6 Global Convergence

Combining the positive definiteness of the Hessian near the global optimal in Section 4 and the tensor initialization methods in Section 5, we come up with the overall globally converging Algorithm 2 and its guarantee Theorem 6.1.

**Theorem 6.1** (Global convergence guarantees). Let $S$ denote a set of i.i.d. samples from distribution $D$ (defined in (1)) and let the activation function be homogeneous satisfying Property 3.1, 3.2, 3.3(a) and Assumption 5.3. Then for any $t \geq 1$ and any $\epsilon > 0$, if $|S| \geq d \log(1/\epsilon) \cdot \text{poly}(\log d, t, k, \lambda)$, $T \geq \log(1/\epsilon) \cdot \text{poly}(k, \nu, \lambda, \sigma^2_1 / \rho)$ and $0 < \eta \leq 1/(k \sigma^2_{\max} \sigma^2_1)$, then there is an Algorithm (procedure LEARNING1NN in Algorithm 2) taking $|S| \cdot d \cdot \text{poly}(\log d, k, \lambda)$ time and outputting a matrix $W^{(T)} \in \mathbb{R}^{d \times k}$ and a vector $v^{(0)} \in \mathbb{R}^k$ satisfying

$$\|W^{(T)} - W^*\|_F \leq \epsilon \|W^*\|_F,$$

and $v^{(0)}_i = v^*_i$.

with probability at least $1 - d^{-\Omega(t)}$.

This follows by combining Theorem 4.4 and Theorem 5.6.

7 Numerical Experiments

In this section we use synthetic data to verify our theoretical results. We generate data points $\{x_i, y_i\}_{i=1}^{n}$ from Distribution $D$ (defined in Eq. (1)). We set $W^* = U \Sigma V^\top$, where $U \in \mathbb{R}^{d \times k}$ and $V \in \mathbb{R}^{k \times k}$ are orthogonal matrices generated from QR decomposition of Gaussian matrices, $\Sigma$ is a diagonal matrix whose diagonal elements are $1, 1 + \frac{\kappa - 1}{k - 1}, 1 + \frac{2(\kappa - 1)}{k - 1}, \ldots, \kappa$. In this experiment, we set $\kappa = 2$ and $k = 5$. We set $v^*_i$ to be randomly picked from $\{-1, 1\}$ with equal chance. We use squared ReLU $\phi(z) = \max\{z, 0\}^2$, which is a smooth homogeneous function. For non-orthogonal tensor methods, we directly use the code provided by [KCL15] with the number of random projections fixed as $L = 100$. We pick the stepsize $\eta = 0.02$ for gradient descent. In the experiments, we don’t do the resampling since the algorithm still works well without resampling.

First we show the number of samples required to recover the parameters for different dimensions. We fix $k = 5$, change $d$ for $d = 10, 20, \cdots, 100$ and $n$ for $n = 1000, 2000, \cdots, 10000$. For each pair of $d$ and $n$, we run 10 trials. We say a trial successfully recovers the parameters if there exists a permutation $\pi : [k] \rightarrow [k]$, such that the returned parameters $W$ and $v$ satisfy

$$\max_{j \in [k]} \{\|w^*_j - w^*_{\pi(j)}\| / \|w^*_j\|\} \leq 0.01 \text{ and } v^*_{\pi(j)} = v^*_j.$$

We record the recovery rates and represent them as grey scale in Fig. 1(a). As we can see from Fig. 1(a), the least number of samples required to have 100% recovery rate is about proportional to the dimension.

Next we test the tensor initialization. We show the error between the output of the tensor method and the ground truth parameters against the number of samples under different dimensions in Fig 1(b). The pure dark blocks indicate, in at least one of the 10 trials, $\sum_{i=1}^{k} v^{(0)}_i \neq \sum_{i=1}^{k} v^*_i$, which means $v^{(0)}_i$ is not correctly initialized. Let $\Pi(k)$ denote the set of all possible permutations $\pi : [k] \rightarrow [k]$. The grey scale represents the averaged error,

$$\min_{\pi \in \Pi(k)} \max_{j \in [k]} \{\|w^*_j - w^{(0)}_{\pi(j)}\| / \|w^*_j\|\},$$
over 10 trials. As we can see, with a fixed dimension, the more samples we have the better initialization we obtain. We can also see that to achieve the same initialization error, the sample complexity required is about proportional to the dimension.

We also compare different initialization methods for gradient descent in Fig. 1(c). We fix $d = 10, k = 5, n = 10000$ and compare three different initialization approaches, (I) Let both $v$ and $W$ be initialized from tensor methods, and then do gradient descent for $W$ while $v$ is fixed; (II) Let both $v$ and $W$ be initialized from random Gaussian, and then do gradient descent for both $W$ and $v$; (III) Let $v = v^*$ and $W$ be initialized from random Gaussian, and then do gradient descent for $W$ while $v$ is fixed. As we can see from Fig 1(c), Approach (I) is the fastest and Approach (II) doesn’t converge even if more iterations are allowed. Both Approach (I) and (III) have linear convergence rate when the objective value is small enough, which verifies our local linear convergence claim.

8 Conclusion

As shown in Theorem 6.1, the tensor initialization followed by gradient descent will provide a globally converging algorithm with linear time/sample complexity in dimension, logarithmic in precision and polynomial in other factors for smooth homogeneous activation functions. Our distilled properties for activation functions include a wide range of non-linear functions and hopefully provide an intuition to understand the role of non-linear activations played in optimization. Deeper neural networks and convergence for SGD will be considered in the future.
References


A  Notation

For any positive integer $n$, we use $[n]$ to denote the set $\{1, 2, \cdots, n\}$. For random variable $X$, let $\mathbb{E}[X]$ denote the expectation of $X$ (if this quantity exists). For any vector $x \in \mathbb{R}^n$, we use $\|x\|$ to denote its $\ell_2$ norm.

We provide several definitions related to matrix $A$. Let $\det(A)$ denote the determinant of a square matrix $A$. Let $A^T$ denote the transpose of $A$. Let $A^\dagger$ denote the Moore-Penrose pseudoinverse of $A$. Let $A^{-1}$ denote the inverse of a full rank square matrix. Let $\|A\|_F$ denote the Frobenius norm of matrix $A$. Let $\|A\|$ denote the spectral norm of matrix $A$. Let $\sigma_i(A)$ to denote the $i$-th largest singular value of $A$. We often use capital letter to denote the stack of corresponding small letter vectors, e.g., $W = [w_1 \ w_2 \ \cdots \ w_k]$. For two same-size matrices, $A, B \in \mathbb{R}^{d_1 \times d_2}$, we use $A \circ B \in \mathbb{R}^{d_1 \times d_2}$ to denote element-wise multiplication of these two matrices.

We use $\otimes$ to denote outer product and $\cdot$ to denote dot product. Given two column vectors $u, v \in \mathbb{R}^n$, then $u \otimes v \in \mathbb{R}^{n \times n}$ and $(u \otimes v)_{i,j} = u_i \cdot v_j$, and $u^\top v = \sum_{i=1}^n u_i v_i \in \mathbb{R}$. Given three column vectors $u, v, w \in \mathbb{R}^n$, then $u \otimes v \otimes w \in \mathbb{R}^{n \times n \times n}$ and $(u \otimes v \otimes w)_{i,j,k} = u_i \cdot v_j \cdot w_k$. We use $\otimes^r \in \mathbb{R}^{n^r}$ to denote the vector $u$ outer product with itself $r - 1$ times.

Tensor $T \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ is symmetric if and only if for any $i, j, k$, $T_{i,j,k} = T_{i,k,j} = T_{j,i,k} = T_{j,k,i} = T_{k,i,j} = T_{k,j,i}$. Given a third order tensor $T \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ and three matrices $A \in \mathbb{R}^{n_1 \times d_1}, B \in \mathbb{R}^{n_2 \times d_2}, C \in \mathbb{R}^{n_3 \times d_3}$, we use $T(A, B, C)$ to denote a $d_1 \times d_2 \times d_3$ tensor where the $(i,j,k)$-th entry is,

$$
\sum_{i'=1}^{n_1} \sum_{j'=1}^{n_2} \sum_{k'=1}^{n_3} T_{i',j',k'} A_{i',i} B_{j',j} C_{k',k}.
$$

We use $\|T\|$ to denote the operator norm of the tensor $T$, i.e.,

$$
\|T\| = \max_{\|a\|=1} |T(a, a, a)|.
$$

For tensor $T \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, we use matrix $T^{(1)} \in \mathbb{R}^{n_1 \times n_2 n_3}$ to denote the flattening of tensor $T$ along the first dimension, i.e., $T^{(1)}_{i,(j-1)n_3+k} = T_{i,j,k}, \forall i \in [n_1], j \in [n_2], k \in [n_3]$. Similarly for matrices $T^{(2)} \in \mathbb{R}^{n_2 \times n_3 n_1}$ and $T^{(3)} \in \mathbb{R}^{n_3 \times n_1 n_2}$.

We use $1_f$ to denote the indicator function, which is 1 if $f$ holds and 0 otherwise. Let $I_d \in \mathbb{R}^{d \times d}$ denote the identity matrix. We use $\phi(z)$ to denote an activation function. We define $(z)_+ := \max\{0, z\}$. We use $\mathcal{D}$ to denote a Gaussian distribution $\mathcal{N}(0, I_d)$ or to denote a joint distribution of $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$, where the marginal distribution of $X$ is $\mathcal{N}(0, I_d)$.

For any function $f$, we define $\tilde{O}(f)$ to be $f \cdot \log^{O(1)}(f)$. In addition to $O(\cdot)$ notation, for two functions $f, g$, we use the shorthand $f \lesssim g$ (resp. $\gtrsim$) to indicate that $f \leq C g$ (resp. $\geq$) for an absolute constant $C$. We use $f \approx g$ to mean $cf \leq g \leq Cf$ for constants $c, C$.

B  Preliminaries

In this section, we introduce some lemmata and corollaries that will be used in the proofs.

B.1  Useful Facts

We provide some facts that will be used in the later proofs.
\textbf{Fact B.1.} \textit{Let }$z$\textit{ denote a }$d$\textit{-dimensional vector, then for any }$C \geq 1$\textit{ and }$n \geq 1$, we have}

$$\Pr_{x \sim \mathcal{N}(0, I_d)}[\|x\|^2 \leq 5Cn \log n] \geq 1 - 1/(nd^C).$$

\textit{Proof.} This follows by Proposition 1.1 in \cite{HKZ12}. \hfill \Box

\textbf{Fact B.2.} \textit{For any }$C \geq 1$\textit{ and }$n \geq 1$, we have}

$$\Pr_{x \sim \mathcal{N}(0, I_d)}[\|x\|^2 \leq 5Cd \log n] \geq 1 - 1/(nd^C).$$

\textit{Proof.} This follows by Proposition 1.1 in \cite{HKZ12}. \hfill \Box

\textbf{Fact B.3.} \textit{Given a full column-rank matrix }$W = [w_1, w_2, \ldots, w_k] \in \mathbb{R}^{d \times k}$, \textit{let }$\overline{W} = \left[ \frac{w_1}{\|w_1\|}, \frac{w_2}{\|w_2\|}, \ldots, \frac{w_k}{\|w_k\|} \right]$. \textit{Then, we have: (I) for any }$i \in [k]$, $\sigma_k(W) \leq \|w_i\| \leq \sigma_1(W)$; (II) $1/\kappa(W) \leq \sigma_k(\overline{W}) \leq \sigma_1(W) \leq \sqrt{k}$.

\textit{Proof.} Part (I). We have,

$$\sigma_k(W) \leq \|\text{diag}(W)\| = \left\| \sum_{i=1}^{k} s_i \frac{w_i}{\|w_i\|} \right\|$$

$$\geq \min_{\|s\| = 1} \sigma_k(W) \left( \sum_{i=1}^{k} \left( \frac{s_i}{\|w_i\|} \right)^2 \right)^{\frac{1}{2}} \quad \text{by definition of } \overline{W}$$

$$\geq \min_{\|s\| = 1} \sigma_k(W) \left( \sum_{i=1}^{k} \frac{s_i}{\max_{j \in [k]} \|w_j\|} \right)^{\frac{1}{2}} \quad \text{by } \max_{j \in [k]} \|w_j\| \geq \|w_i\|$$

$$= \sigma_k(W) / \max_{j \in [k]} \|w_j\| \quad \text{by } \|s\| = 1$$

$$\geq \sigma_k(W) / \sigma_1(W) \quad \text{by } \max_{j \in [k]} \|w_j\| \leq \sigma_1(W)$$

$$= 1/\kappa(W).$$

It remains to upper bound $\sigma_1(\overline{W})$,

$$\sigma_1(\overline{W}) \leq \left( \sum_{i=1}^{k} \sigma_i^2(\overline{W}) \right)^{\frac{1}{2}} = \|\overline{W}\|_F \leq \sqrt{k}.$$

\textbf{Fact B.4.} \textit{Let }$U \in \mathbb{R}^{d \times k}$\textit{ and }$V \in \mathbb{R}^{d \times k}$\textit{ ($k \leq d$) denote two orthogonal matrices. Then }$\|UU^\top - VV^\top\| = \|(I - UU^\top)V\| = \|(I - VV^\top)U\| = \sqrt{1 - \sigma_k^2(U^\top V)}$.\hfill \Box
Proof. Let $U_\perp \in \mathbb{R}^{d \times (d-k)}$ and $V_\perp \in \mathbb{R}^{d \times (d-k)}$ be the orthogonal complementary matrices of $U, V \in \mathbb{R}^{d \times k}$ respectively.

\[ \|UU^T - VV^T\| = \|(I - VV^T)UU^T - VV^T(I - UU^T)\| = \|V_\perp V_\perp^T UU^T - VV^T U_\perp U_\perp^T\| \]
\[ = \left\| [V_\perp V] \begin{bmatrix} V_\perp^T U & 0 \\ 0 & -V^T U_\perp \end{bmatrix} \begin{bmatrix} U_\perp^T \end{bmatrix} \right\| = \max(\|V_\perp^T U\|, \|V^T U_\perp\|). \]

We show how to simplify $\|V_\perp^T U\|$,
\[ \|V_\perp^T U\| = \|(I - VV^T)U\| = \sqrt{\|U^T(I - VV^T)U\|} = \max_{\|a\|=1} \sqrt{1 - \|V^T Ua\|^2} = \sqrt{1 - \sigma_k^2(V^T U)}. \]

Similarly we can simplify $\|U_\perp^T V\|$,
\[ \|U_\perp^T V\| = \sqrt{1 - \sigma_k^2(U^T V)} = \sqrt{1 - \sigma_k^2(V^T U)}. \]

\hfill \Box

**Fact B.5.** Let $C \in \mathbb{R}^{d_1 \times d_2}, B \in \mathbb{R}^{d_2 \times d_3}$ be two matrices. Then $\|CB\| \leq \|C\|\|B\|_F$ and $\|CB\| \geq \sigma_{\min}(C)\|B\|_F$.

Proof. For each $i \in [d_3]$, let $b_i$ denote the $i$-th column of $B$. We can upper bound $\|CB\|$,
\[ \|CB\|_F = \left( \sum_{i=1}^{d_2} \|Cbi\|^2 \right)^{1/2} \leq \left( \sum_{i=1}^{d_2} \|C\|^2 \|b_i\|^2 \right)^{1/2} = \|C\|\|B\|_F. \]

We show how to lower bound $\|CB\|$,
\[ \|CB\| = \left( \sum_{i=1}^{d_2} \|Cbi\|^2 \right)^{1/2} \geq \left( \sum_{i=1}^{d_2} \sigma_k^2(C) \|b_i\|^2 \right)^{1/2} = \sigma_{\min}(C)\|B\|_F. \]

\hfill \Box

**Fact B.6.** Let $a, b, c \geq 0$ denote three constants, let $u, v, w \in \mathbb{R}^d$ denote three vectors, let $\mathcal{D}_d$ denote Gaussian distribution $\mathcal{N}(0, I_d)$ then
\[ \mathbb{E}_{x \sim \mathcal{D}_d} \left[ |u^T x|^a |v^T x|^b |w^T x|^c \right] \leq \left( \mathbb{E}_{x \sim \mathcal{D}_d} [|u^T x|^{2a}] \right)^{1/2} \cdot \left( \mathbb{E}_{x \sim \mathcal{D}_d} [|u^T x|^{2b}] \right)^{1/4} \cdot \left( \mathbb{E}_{x \sim \mathcal{D}_d} [|u^T x|^{2c}] \right)^{1/4} \]
\[ \leq \|u\|^a \|v\|^b \|w\|^c, \]

where the first step follows by Hölder’s inequality, i.e., $\mathbb{E}[|XYZ|] \leq (\mathbb{E}[|X|^2])^{1/2} \cdot (\mathbb{E}[|Y|^4])^{1/4} \cdot (\mathbb{E}[|Z|^4])^{1/4}$, the third step follows by calculating the expectation and $a, b, c$ are constants.

Since all the three components $|u^T x|, |v^T x|, |w^T x|$ are positive and related to a common random vector $x$, we can show a lower bound,
\[ \mathbb{E}_{x \sim \mathcal{D}_d} \left[ |u^T x|^a |v^T x|^b |w^T x|^c \right] \geq \|u\|^a \|v\|^b \|w\|^c. \]

\hfill \Box
B.2 Matrix Bernstein

In many proofs we need to bound the difference between some population matrices/tensors and their empirical versions. Typically, the classic matrix Bernstein inequality requires the norm of the random matrix be bounded almost surely (e.g., Theorem 6.1 in [Tro12]) or the random matrix satisfies subexponential property (Theorem 6.2 in [Tro12]). However, in our cases, most of the random matrices don’t satisfy these conditions. So we derive the following lemmata that can deal with random matrices that are not bounded almost surely or follow subexponential distribution, but are bounded with high probability.

**Lemma B.7** (Matrix Bernstein for unbounded case (A modified version of bounded case, Theorem 6.1 in [Tro12])). Let $B$ denote a distribution over $\mathbb{R}^{d_1 \times d_2}$. Let $d = d_1 + d_2$. Let $B_1, B_2, \cdots B_n$ be i.i.d. random matrices sampled from $B$. Let $\hat{B} = \mathbb{E}_{B \sim B}[B]$ and $\hat{B} = \frac{1}{n} \sum_{i=1}^{n} B_i$. For parameters $m \geq 0, \gamma \in (0,1), \nu > 0, L > 0$, if the distribution $B$ satisfies the following four properties,

(I) $\Pr_{B \sim B}[\|B\| \leq m] \geq 1 - \gamma$;

(II) $\left\| \mathbb{E}_{B \sim B}[B] \right\| > 0$;

(III) $\max \left( \left\| \mathbb{E}_{B \sim B}[BB^\top] \right\|, \left\| \mathbb{E}_{B \sim B}[B^\top B] \right\| \right) \leq \nu$;

(IV) $\max_{\|a\| = \|b\| = 1} \left( \mathbb{E}_{B \sim B}\left[ (a^\top Bb)^2 \right] \right)^{1/2} \leq L$.

Then we have for any $0 < \epsilon < 1$ and $t \geq 1$, if

$$n \geq (18t \log d) \cdot \left( \nu + \|B\|^2 + m \|B\|/\epsilon^2 \|B\|^2 \right) / (2L)^2$$

with probability at least $1 - 1/d^2t - n\gamma$,

$$\|\hat{B} - B\| \leq \epsilon \|B\|.$$

**Proof.** Define the event

$$\xi_i = \{\|B_i\| \leq m\}, \forall i \in [n].$$

Define $M_i = 1_{\|B_i\| \leq m}B_i$. Let $\hat{M} = \mathbb{E}_{B \sim B}[1_{\|B\| \leq m}B] \text{ and } \hat{M} = \frac{1}{n} \sum_{i=1}^{n} M_i$. By triangle inequality, we have

$$\|\hat{B} - B\| \leq \|\hat{B} - \hat{M}\| + \|\hat{M} - M\| + \|M - B\|. \quad (4)$$

In the next few paragraphs, we will upper bound the above three terms.

**The first term in Eq. (4).** Denote $\xi^c$ as the complementary set of $\xi$, thus $\Pr[\xi^c] \leq \gamma$. By a union bound over $i \in [n]$, with probability $1 - n\gamma$, $\|B_i\| \leq m$ for all $i \in [n]$. Thus $\hat{M} = \hat{B}$.

**The second term in Eq. (4).** For a matrix $B$ sampled from $B$, we use $\xi$ to denote the event
that \( \xi = \{ \|B\| \leq m \} \). Then, we can upper bound \( \|\overline{M} - \overline{B}\| \) in the following way,

\[
\|\overline{M} - \overline{B}\| = \left\| \mathbb{E}_{B \sim \mathcal{B}} [1_{\|B\| \leq m} \cdot B] - \mathbb{E}_{B \sim \mathcal{B}} [B] \right\| \\
= \left\| \mathbb{E}_{B \sim \mathcal{B}} [B \cdot 1_{\xi^c}] \right\| \\
= \max_{\|a\| = \|b\| = 1} \mathbb{E}_{B \sim \mathcal{B}} [a^\top B b 1_{\xi^c}] \\
\leq \max_{\|a\| = \|b\| = 1} \mathbb{E}_{B \sim \mathcal{B}} [(a^\top B b)^2]^{1/2} \cdot \mathbb{E}_{B \sim \mathcal{B}} [1_{\xi^c}]^{1/2} \quad \text{by Hölder’s inequality} \\
\leq L \mathbb{E}_{B \sim \mathcal{B}} [1_{\xi^c}]^{1/2} \quad \text{by Property (IV)} \\
\leq L\gamma^{1/2}, \quad \text{by \( \Pr[\xi^c] \leq \gamma \)} \\
\leq \frac{1}{2} \epsilon \|\overline{B}\|, \quad \text{by \( \gamma \leq (\epsilon \|\overline{B}\|/(2L))^2 \)}
\]

which implies

\[
\|\overline{M} - \overline{B}\| \leq \frac{\epsilon}{2} \|\overline{B}\|.
\]

Since \( \epsilon < 1 \), we also have \( \|\overline{M} - \overline{B}\| \leq \frac{1}{2} \|\overline{B}\| \) and \( \frac{1}{2} \|\overline{B}\| \geq \|\overline{M}\| \geq \frac{1}{2} \|\overline{B}\| \).

**The third term in Eq. (4).** We can bound \( \|\overline{M} - \overline{M}\| \) by Matrix Bernstein’s inequality [Tro12]. We define \( Z_i = M_i - \overline{M} \). Thus we have \( \mathbb{E}_{B_i \sim \mathcal{B}} [Z_i] = 0 \), \( \|Z_i\| \leq 2m \), and

\[
\left\| \mathbb{E}_{B_i \sim \mathcal{B}} [Z_i Z_i^\top] \right\| = \left\| \mathbb{E}_{B_i \sim \mathcal{B}} [M_i M_i^\top] - \overline{M} \cdot \overline{M}^\top \right\| \leq \nu + \|\overline{M}\|^2 \leq \nu + 3\|\overline{B}\|^2.
\]

Similarly, we have \( \left\| \mathbb{E}_{B_i \sim \mathcal{B}} [Z_i^\top Z_i] \right\| \leq \nu + 3\|\overline{B}\|^2 \). Using matrix Bernstein’s inequality, for any \( \epsilon > 0 \),

\[
\Pr_{B_1, \ldots, B_n \sim \mathcal{B}} \left[ \frac{1}{n} \left\| \sum_{i=1}^{n} Z_i \right\| \geq \epsilon \|\overline{B}\| \right] \leq d \exp \left( -\frac{\epsilon^2 \|\overline{B}\|^2 n/2}{\nu + 3\|\overline{B}\|^2 + 2m\|\overline{B}\|\epsilon/3} \right).
\]

By choosing

\[
n \geq (3t \log d) \cdot \frac{\nu + 3\|\overline{B}\|^2 + 2m\|\overline{B}\|\epsilon/3}{\epsilon^2 \|\overline{B}\|^2/2},
\]

for \( t \geq 1 \), we have with probability at least \( 1 - 1/d^{2t} \),

\[
\left\| \frac{1}{n} \sum_{i=1}^{n} M_i - \overline{M} \right\| \leq \frac{\epsilon}{2} \|\overline{B}\|
\]

Putting it all together, we have for \( 0 < \epsilon < 1 \), if

\[
n \geq (18t \log d) \cdot (\nu + \|\overline{B}\|^2 + m\|\overline{B}\|\epsilon)/(\epsilon^2 \|\overline{B}\|^2) \quad \text{and} \quad \gamma \leq (\epsilon \|\overline{B}\|/(2L))^2
\]

with probability at least \( 1 - 1/d^{2t} - n\gamma \),

\[
\left\| \frac{1}{n} \sum_{i=1}^{n} B_i - \mathbb{E}_{B \sim \mathcal{B}} [B] \right\| \leq \epsilon \left\| \mathbb{E}_{B \sim \mathcal{B}} [B] \right\|.
\]

\[\square\]
Corollary B.8 (Error bound for symmetric rank-one random matrices). Let $x_1, x_2, \ldots, x_n$ denote $n$ i.i.d. samples drawn from Gaussian distribution $N(0, I_d)$. Let $h(x) : \mathbb{R}^d \to \mathbb{R}$ be a function satisfying the following properties (I), (II) and (III).

(I) $\Pr_{x \sim N(0, I_d)} \left[ |h(x)| \leq m \right] \geq 1 - \gamma$

(II) $\left\| \mathbb{E}_{x \sim N(0, I_d)} [h(x)xx^\top] \right\| > 0$;

(III) $\left( \mathbb{E}_{x \sim N(0, I_d)} [h^4(x)] \right)^{1/4} \leq L$.

Define function $B(x) = h(x)xx^\top \in \mathbb{R}^{d \times d}$, $\forall i \in [n]$. Let $\overline{B} = \mathbb{E}_{x \sim N(0, I_d)} [h(x)xx^\top]$. For any $0 < \epsilon < 1$ and $t \geq 1$, if

$$n \geq (t \log d) \cdot (L^2d + \|\overline{B}\|^2 + (mtd \log n)\|\overline{B}\|/(\epsilon^2\|\overline{B}\|^2)),$$

and $\gamma + 1/(nd^2t) \leq (\epsilon\|\overline{B}\|/L)^2$ then

$$\Pr_{x_1, \ldots, x_n \sim N(0, I_d)} \left[ \left\| \overline{B} - \frac{1}{n} \sum_{i=1}^n B(x_i) \right\| \leq \epsilon\|\overline{B}\| \right] \geq 1 - 2/(d^2t) - n\gamma.$$ 

Proof. We show that the four Properties in Lemma B.7 are satisfied. Define function $B(x) = h(x)xx^\top$.

(I) $\|B(x)\| = h(x)||x||^2$.

By using Fact B.2, we have

$$\Pr_{x \sim N(0, I_d)} [||x||^2 \leq 10td \log n] \geq 1 - 1/(nd^2t)$$

Therefore,

$$\Pr_{x \sim N(0, I_d)} [\|B(x)\| \leq m \cdot 10td \log(n)] \geq 1 - \gamma - 1/(nd^2t).$$

(II) $\left\| \mathbb{E}_{B \sim B^T} [B] \right\| = \left\| \mathbb{E}_{x \sim N(0, I_d)} [h(x)xx^\top] \right\| > 0$.

(III) $\max \left( \left\| \mathbb{E}_{B \sim B^T} [BB^\top] \right\|, \left\| \mathbb{E}_{B \sim B^T} [B^\top B] \right\| \right)$

$$= \max_{\|a\|=1} \mathbb{E}_{x \sim N(0, I_d)} [(h(x))^2 \|x\|^2 (a^\top x)^2]$$

$$\leq \left( \mathbb{E}_{x \sim N(0, I_d)} [(h(x))^4] \right)^{1/2} \cdot \left( \mathbb{E}_{x \sim N(0, I_d)} [\|x\|^8] \right)^{1/4} \cdot \max_{\|a\|=1} \left( \mathbb{E}_{x \sim N(0, I_d)} [(a^\top x)^8] \right)^{1/4}$$

$$\leq L^2d.$$
Applying Lemma B.7, we obtain, for any $0 < \epsilon < 1$ and $t \geq 1$, if

$$n \gtrsim (t \log d) \cdot (L^2 d + \|B\|^2 + (n t \log n)\|B\|\epsilon)/(\epsilon^2\|B\|^2),$$

then

$$\Pr_{x_1, \ldots, x_n \sim \mathcal{N}(0, I_d)} \left[ \left\| B - \frac{1}{n} \sum_{i=1}^{n} B(x_i) \right\| \leq \epsilon\|B\| \right] \geq 1 - 2/(d^{2t}) - n\gamma.$$  

\[\square\]

## C Properties of Activation Functions

**Theorem 3.5.** ReLU $\phi(z) = \max\{z, 0\}$, leaky ReLU $\phi(z) = \max\{z, 0.01z\}$, squared ReLU $\phi(z) = \max\{z, 0\}^2$ and any non-linear non-decreasing smooth functions with bounded symmetric $\phi'(z)$, like the sigmoid function $\phi(z) = 1/(1 + e^{-z})$, the tanh function and the erf function $\phi(z) = \int_0^z e^{-t^2}dt$, satisfy Property 3.1, 3.2, 3.3. The linear function, $\phi(z) = z$, doesn't satisfy Property 3.2 and the quadratic function, $\phi(z) = z^2$, doesn't satisfy Property 3.1 and 3.2.

**Proof.** We can easily verify that ReLU, leaky ReLU and squared ReLU satisfy Property 3.2 by calculating $\rho(\sigma)$ in Property 3.2, which is shown in Table 1. Property 3.1 for ReLU, leaky ReLU and squared ReLU can be verified since they are non-decreasing with bounded first derivative. ReLU and leaky ReLU are piece-wise linear, so they satisfy Property 3.3(b). Squared ReLU is smooth so it satisfies Property 3.3(a).

<table>
<thead>
<tr>
<th>Activations</th>
<th>ReLU</th>
<th>Leaky ReLU</th>
<th>squared ReLU</th>
<th>erf</th>
<th>sigmoid ($\sigma = 0.1$)</th>
<th>sigmoid ($\sigma = 1$)</th>
<th>sigmoid ($\sigma = 10$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_0(\sigma)$</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1.01}{2}$</td>
<td>$\sigma\sqrt{\frac{2}{\pi}}$</td>
<td>$\frac{1}{(2\sigma^2+1)^{1/2}}$</td>
<td>0.99</td>
<td>0.605706</td>
<td>0.079</td>
</tr>
<tr>
<td>$a_1(\sigma)$</td>
<td>$\frac{1}{\sqrt{2\pi}}$</td>
<td>$\frac{0.99}{\sqrt{2\pi}}$</td>
<td>$\sigma$</td>
<td>$\sigma$</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$a_2(\sigma)$</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1.01}{2}$</td>
<td>$2\sigma\sqrt{\frac{2}{\pi}}$</td>
<td>$\frac{1}{(4\sigma^2+1)^{1/2}}$</td>
<td>0.97</td>
<td>0.24</td>
<td>0.00065</td>
</tr>
<tr>
<td>$\beta_0(\sigma)$</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1.0001}{2}$</td>
<td>$2\sigma$</td>
<td>$\frac{1}{(4\sigma^2+1)^{1/2}}$</td>
<td>0.98</td>
<td>0.46</td>
<td>0.053</td>
</tr>
<tr>
<td>$\beta_2(\sigma)$</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1.0001}{2}$</td>
<td>$6\sigma$</td>
<td>$\frac{1}{(4\sigma^2+1)^{1/2}}$</td>
<td>0.94</td>
<td>0.11</td>
<td>0.00017</td>
</tr>
<tr>
<td>$\rho(\sigma)$</td>
<td>0.091</td>
<td>0.089</td>
<td>0.27$\sigma$</td>
<td>$\rho_{\text{erf}}(\sigma)$</td>
<td>$1.8E-4$</td>
<td>$4.9E-2$</td>
<td>$5.1E-5$</td>
</tr>
</tbody>
</table>

Table 1: $\rho(\sigma)$ values for different activation functions. Note that we can calculate the exact values for ReLU, Leaky ReLU, squared ReLU and erf. We can’t find a closed-form value for sigmoid or tanh, but we calculate the numerical values of $\rho(\sigma)$ for $\sigma = 0.1, 1, 10$. $\rho_{\text{erf}}(\sigma) = \min\{(4\sigma^2+1)^{-1/2} - (2\sigma^2+1)^{-1}, (4\sigma^2+1)^{-3/2} - (2\sigma^2+1)^{-3}, (2\sigma^2+1)^{-2}\}$.
Smooth non-decreasing activations with bounded first derivatives automatically satisfy Property 3.1 and 3.3. For Property 3.2, since their first derivatives are symmetric, we have $E[\phi'(\sigma \cdot z)] = 0$. Then by Hölder’s inequality and $\phi'(z) \geq 0$, we have

$$E_{z \sim D_1} [\phi^2(\sigma \cdot z)] \geq \left( E_{z \sim D_1} [\phi'(\sigma \cdot z)] \right)^2,$$

$$E_{z \sim D_1} [\phi^2(\sigma \cdot z)^2] \cdot E_{z \sim D_1} [z^2] \geq \left( E_{z \sim D_1} [\phi'(\sigma \cdot z)] \right)^2,$$

$$E_{z \sim D_1} [\phi'(\sigma \cdot z)] \cdot E_{z \sim D_1} [\phi'(\sigma \cdot z)] = E_{z \sim D_1} \left( |\sqrt{\phi'(\sigma \cdot z)}|^2 \right) \geq \left( E_{z \sim D_1} [\phi'(\sigma \cdot z)] \right)^2.$$

The equality in the first inequality happens when $\phi'(\sigma \cdot z)$ is a constant a.e. The equality in the second inequality happens when $|\phi'(\sigma \cdot z)|$ is a constant a.e., which is invalidated by the non-linearity and smoothness condition. The equality in the third inequality holds only when $\phi'(z) = 0$ a.e., which leads to a constant function under non-decreasing condition. Therefore, $\rho(\sigma) > 0$ for any smooth non-decreasing non-linear activations with bounded symmetric first derivatives. The statements about linear activations and quadratic activation follow direct calculations.

---

**D Local Positive Definiteness of Hessian**

**D.1 Main Results for Positive Definiteness of Hessian**

**D.1.1 Bounding the Spectrum of the Hessian near the Ground Truth**

**Theorem D.1** (Bounding the spectrum of the Hessian near the ground truth). For any $W \in \mathbb{R}^{d \times k}$ with $\|W - W^*\| \lesssim v_{\min}^4 k^5 \lambda_2^2 v_{\max}^4 \rho(\sigma_k)/(k^2 \kappa^{5} \lambda^2 \rho(\sigma_k)) \cdot \|W^*\|$, let $S$ denote a set of i.i.d. samples from distribution $D$ (defined in (1)) and let the activation function satisfy Property 3.1,3.3. Then for any $t \geq 1$, if $|S| \geq d \cdot \text{poly}(d, t) \cdot k^2 v_{\max}^4 \rho(\sigma_k)/(v_{\min}^4 k^5 \lambda_2^2)$, we have with probability at least $1 - d^{-\Omega(t)}$,

$$\Omega(v_{\min}^4 k^5 \lambda_2^2) I \preceq \nabla^2 f_S(W) \preceq O(k v_{\max}^2 \rho(\sigma_k)).$$

**Proof.** The main idea of the proof follows the following inequalities,

$$\nabla^2 f_D(W) - \nabla^2 \tilde{f}_S(W) - \nabla^2 f_D(W^*) \preceq \nabla^2 \tilde{f}_S(W) - \nabla^2 f_D(W^*) \preceq \nabla^2 f_D(W) - \nabla^2 f_D(W^*)$$

The proof sketch is first to bound the range of the eigenvalues of $\nabla^2 f_D(W^*)$ (Lemma D.3) and then bound the spectral norm of the remaining error, $\|\nabla^2 \tilde{f}_S(W) - \nabla^2 f_D(W^*)\|$. $\|\nabla^2 \tilde{f}_S(W) - \nabla^2 f_D(W^*)\|$ can be further decomposed into two parts, $\|\nabla^2 \tilde{f}_S(W) - H\|$ and $\|H - \nabla^2 f_D(W^*)\|$, where $H$ is $\nabla^2 f_D(W)$ if $\phi$ is smooth, otherwise $H$ is a specially designed matrix. We can upper bound them when $W$ is close enough to $W^*$ and there are enough samples. In particular, if the activation satisfies Property 3.3(a), see Lemma D.10 for bounding $\|H - \nabla^2 f_D(W^*)\|$ and Lemma D.11 for bounding $\|H - \nabla^2 f_S(W)\|$. If the activation satisfies Property 3.3(b), see Lemma D.15.

Finally, we can complete the proof by setting $\delta = O(v_{\min}^4 k^5 \lambda_2^2) / (k v_{\max}^2)$ in Lemma D.11 and Lemma D.15, setting $\|W - W^*\| \lesssim v_{\min}^4 \rho(\sigma_k) / (k^2 \kappa^{5} \lambda^2 \rho(\sigma_k))$ in Lemma D.10 and setting $\|W - W^*\| \lesssim v_{\min}^4 k^5 \lambda_2^2$ in Lemma D.15.
D.1.2 Local Linear Convergence of Gradient Descent

Although Theorem D.1 gives upper and lower bounds for the spectrum of the Hessian w.h.p., it only holds when the current set of parameters $W$ are independent of samples. When we use iterative methods, like gradient descent, to optimize this objective, the next iterate calculated from the current set of samples will depend on this set of samples. Therefore, we need to do resampling at each iteration. Here we show that for activations that satisfies Properties 3.1, 3.2 and 3.3(a), linear convergence of gradient descent is guaranteed. To the best of our knowledge, there is no linear convergence guarantees for general non-smooth objective. So the following proposition also applies to smooth objectives only, which excludes ReLU.

**Theorem D.2** (Linear convergence of gradient descent, formal version of Theorem 4.4). Let $W^c \in \mathbb{R}^{d \times k}$ be the current iterate satisfying

$$\|W^c - W^*\| \leq v_{\min}^4 \rho^2(\sigma_k)/(k^2 \kappa^5 \lambda^2 v_{\max}^4 \sigma_1^4 \|W^*\|).$$

Let $S$ denote a set of i.i.d. samples from distribution $D$ (defined in (1)) Let the activation function satisfy Property 3.1, 3.2 and 3.3(a). Define

$$m_0 = \Theta(v_{\min}^2 \rho(\sigma_k)/(\kappa \lambda))$$

and $M_0 = \Theta(k v_{\max}^2 \sigma_1^2 \rho)$.  

For any $t \geq 1$, if we choose

$$|S| \geq d \cdot \text{poly}(\log d, t) \cdot k^2 v_{\max}^4 \kappa \lambda^2 \sigma_1^4 \rho^2/(v_{\min}^4 \rho^2(\sigma_k))$$

and perform gradient descent with step size $1/M_0$ on $\hat{f}_S(W^c)$ and obtain the next iterate,

$$\tilde{W} = W^c - \frac{1}{M_0} \nabla \hat{f}_S(W^c),$$

then with probability at least $1 - d^{-\Omega(t)}$,

$$\|\tilde{W} - W^*\|_F^2 \leq (1 - \frac{m_0}{M_0})\|W^c - W^*\|_F^2.$$

**Proof.** To prove Theorem D.2, we need to show the positive definite properties on the entire line between the current iterate and the optimum by constructing a set of anchor points, which are independent of the samples. Then we apply traditional analysis for the linear convergence of gradient descent.

In particular, given a current iterate $W^c$, we set $d^{(p+1)/2}$ anchor points $\{W^a\}_{a=1,2,\ldots,d^{(p+1)/2}}$ equally along the line $\xi W^* + (1 - \xi)W^c$ for $\xi \in [0,1]$.

According to Theorem D.1, by setting $t \leftarrow t + (p + 1)/2$, we have with probability at least $1 - d^{-(t+(p+1)/2)}$ for each anchor point $\{W^a\}$,

$$m_0 I \preceq \nabla^2 \hat{f}_S(W^a) \preceq M_0 I.$$  

Then given an anchor point $W^a$, according to Lemma D.16, we have with probability $1 - 2d^{-(t+(p+1)/2)}$, for any points $W$ between $(W^a-1 + W^a)/2$ and $(W^a + W^a+1)/2$,

$$m_0 I \preceq \nabla^2 \hat{f}_S(W) \preceq M_0 I. \quad (6)$$

Finally by applying union bound over these $d^{(p+1)/2}$ small intervals, we have with probability at least $1 - d^{-t}$ for any points $W$ on the line between $W^c$ and $W^*$,

$$m_0 I \preceq \nabla^2 \hat{f}_S(W) \preceq M_0 I.$$
Now we can apply traditional analysis for linear convergence of gradient descent. Let \( \eta \) denote the stepsize.

\[
\|\tilde{W} - W^*\|_F^2 \\
= \|W^c - \eta \nabla \hat{f}_S(W^c) - W^*\|_F^2 \\
= \|W^c - W^*\|_F^2 - 2\eta \langle \nabla \hat{f}_S(W^c), (W^c - W^*) \rangle + \eta^2 \|\nabla \hat{f}_S(W^c)\|_F^2
\]

We can rewrite \( \hat{f}_S(W^c) \),

\[
\nabla \hat{f}_S(W^c) = \left( \int_0^1 \nabla^2 \hat{f}_S(W^* + \gamma(W^c - W^*))d\gamma \right) \text{vec}(W^c - W^*).
\]

We define function \( \hat{H}_S : \mathbb{R}^{d \times k} \to \mathbb{R}^{dk \times dk} \) such that

\[
\hat{H}_S(W^c - W^*) = \left( \int_0^1 \nabla^2 \hat{f}_S(W^* + \gamma(W^c - W^*))d\gamma \right).
\]

According to Eq. (6),

\[
m_0 I \preceq \hat{H} \preceq M_0 I.
\]

We can upper bound \( \|\nabla \hat{f}_S(W^c)\|_F^2 \),

\[
\|\nabla \hat{f}_S(W^c)\|_F^2 = \langle \hat{H}_S(W^c - W^*), \hat{H}_S(W^c - W^*) \rangle \leq M_0 \langle W^c - W^*, \hat{H}_S(W^c - W^*) \rangle.
\]

Therefore,

\[
\|\tilde{W} - W^*\|_F^2 \\
\leq \|W^c - W^*\|_F^2 - (\eta^2 M_0 + 2\eta) \langle W^c - W^*, \hat{H}(W^c - W^*) \rangle \\
\leq \|W^c - W^*\|_F^2 - (\eta^2 M_0 + 2\eta)m_0 \|W^c - W^*\|_F^2 \\
= \|W^c - W^*\|_F^2 - \frac{m_0}{M_0} \|W^c - W^*\|_F^2 \\
\leq (1 - \frac{m_0}{M_0}) \|W^c - W^*\|_F^2
\]

where the third equality holds by setting \( \eta = \frac{1}{M_0} \).

\[\square\]

**D.2 Positive Definiteness of Population Hessian at the Ground Truth**

The goal of this Section is to prove Lemma D.3.

**Lemma D.3** (Positive definiteness of population Hessian at the ground truth). If \( \phi(z) \) satisfies Property 3.1, 3.2 and 3.3, we have the following property for the second derivative of function \( f_D(W) \) at \( W^* \),

\[
\Omega(v^2_{\min} \rho(\sigma_k)/(\kappa^2 \lambda)) I \preceq \nabla^2 f_D(W^*) \preceq O(kv^2_{\max} \sigma_1^2 p) I.
\]

**Proof.** The proof directly follows Lemma D.6 (Section D.2.2) and Lemma D.7(Section D.2.3).  \[\square\]
D.2.1 Lower Bound on the Eigenvalues of Hessian for the Orthogonal Case

**Lemma D.4.** Let $\mathcal{D}_1$ denote Gaussian distribution $\mathcal{N}(0, 1)$. Let $\alpha_0 = E_{z \sim \mathcal{D}_1}[\phi'(z)]$, $\alpha_1 = E_{z \sim \mathcal{D}_1}[\phi''(z)z]$, $\alpha_2 = E_{z \sim \mathcal{D}_1}[\phi''(z)z^2]$, $\beta_0 = E_{z \sim \mathcal{D}_1}[\phi''(z)]$, $\beta_2 = E_{z \sim \mathcal{D}_1}[\phi''(z)z^2]$. Let $\rho$ denote $\min\{\beta_0 - \alpha_0^2 - \alpha_1^2, (\beta_2 - \alpha_2^2) - \alpha_1^2\}$. Let $P = [p_1 \ p_2 \ \cdots \ p_k] \in \mathbb{R}^{k \times k}$. Then we have,

$$\mathbb{E}_{u \sim \mathcal{D}_k} \left[ \left( \sum_{i=1}^{k} p_i^T u \cdot \phi'(u_i) \right)^2 \right] \geq \rho \|P\|_F^2 \tag{8}$$

*Proof.* The main idea is to explicitly calculate the LHS of Eq (8), then reformulate the equation and find a lower bound represented by $\alpha_0, \alpha_1, \alpha_2, \beta_0, \beta_2$.

Further, we can rewrite the diagonal term in the following way,

$$A = \sum_{i=1}^{k} \mathbb{E}_{u \sim \mathcal{D}_k} \left[ p_i^T (\phi'(u_i)^2 \cdot uu^T) p_i \right]$$

$$= \sum_{i=1}^{k} \mathbb{E}_{u \sim \mathcal{D}_k} \left[ p_i^T \phi'(u_i)^2 \cdot \left( u_i^2 e_i e_i^T + \sum_{j \neq i} u_i u_j (e_i e_j^T + e_j e_i^T) + \sum_{j \neq i} \sum_{l \neq i} u_j u_l e_l e_l^T \right) \right] p_i$$

$$= \sum_{i=1}^{k} \left[ p_i^T \phi'(u_i)^2 \cdot \left( u_i^2 e_i e_i^T + \sum_{j \neq i} u_i u_j e_j e_j^T \right) \right] p_i$$

$$= \sum_{i=1}^{k} \left[ p_i^T \phi'(u_i)^2 \cdot \left( \sum_{j \neq i} u_i u_j e_j e_j^T \right) \right] p_i$$

$$= \sum_{i=1}^{k} \left[ p_i^T (\beta_2 e_i e_i^T + \sum_{j \neq i} \beta_0 e_j e_j^T) \right] p_i$$

$$= \sum_{i=1}^{k} \left[ p_i^T ((\beta_2 - \beta_0) e_i e_i^T + \beta_0 I_k) p_i \right]$$

$$= (\beta_2 - \beta_0) \sum_{i=1}^{k} p_i^T e_i e_i^T p_i + \beta_0 \sum_{i=1}^{k} p_i^T p_i$$

$$= (\beta_2 - \beta_0) \|\text{diag}(P)\|^2 + \beta_0 \|P\|_F^2,$$
where the second step follows by rewriting $uu^\top = \sum_{i=1}^{k} \sum_{j=1}^{k} u_i u_j e_i e_j^\top$, the third step follows by

$$E_{u \sim D_k} [\phi'(u_i)^2 u_i u_j] = 0, \forall j \neq i$$

and the last step follows by

$$\sum_{i=1}^{k} p_i^2 p_i = \sum_{i=1}^{k} \|p_i\|^2 = \|P\|_F^2.$$

We can rewrite the off-diagonal term in the following way,

$$B = \sum_{i \neq l} E_{u \sim D_k} [p_i^\top (\phi'(u_i) \phi'(u_i) \cdot uu^\top) p_l]$$

$$= \sum_{i \neq l} E_{u \sim D_k} \left[ p_i^\top \left( \phi'(u_i) \phi'(u_i) \cdot \left( u_i^2 e_i e_i^\top + u_i^2 e_i e_l^\top + u_i u_l (e_i e_l^\top + e_l e_i^\top) + \sum_{j \neq l} u_i u_j e_i e_j^\top \right) \right) p_l \right]$$

$$+ \sum_{i \neq l} E_{u \sim D_k} \left[ p_i^\top \left( \phi'(u_i) \phi'(u_i) \cdot \left( u_i^2 e_i e_i^\top + u_i^2 e_i e_l^\top + u_i u_l (e_i e_l^\top + e_l e_i^\top) + \sum_{j \neq l} u_j u_l e_j e_l^\top \right) \right) p_l \right]$$

$$= \sum_{i \neq l} E_{u \sim D_k} \left[ p_i^\top \left( \phi'(u_i) \phi'(u_i) \cdot \left( u_i^2 e_i e_i^\top + u_i^2 e_i e_l^\top + u_i u_l (e_i e_l^\top + e_l e_i^\top) + \sum_{j \neq l} u_i u_j e_i e_j^\top \right) \right) p_l \right]$$

$$+ \sum_{i \neq l} E_{u \sim D_k} \left[ \phi'(u_i) \phi'(u_i) u_i u_l (e_i e_l^\top + e_l e_i^\top) + \sum_{j \neq l} E_{u \sim D_k} [\phi'(u_i) \phi'(u_i) u_j^2 e_j e_j^\top] p_l \right]$$

$$= \sum_{i \neq l} E_{u \sim D_k} \left[ p_i^\top \left( \alpha_0 \alpha_2 (e_i e_i^\top + e_l e_l^\top) + \alpha_1^2 (e_i e_l^\top + e_l e_i^\top) + \sum_{j \neq i,l} \alpha_0^2 e_j e_j^\top \right) p_l \right]$$

$$= \sum_{i \neq l} E_{u \sim D_k} \left[ p_i^\top \left( (\alpha_0 \alpha_2 - \alpha_0^2) (e_i e_i^\top + e_l e_l^\top) + \alpha_1^2 (e_i e_l^\top + e_l e_i^\top) + \alpha_0^2 I_k \right) p_l \right]$$

$$= (\alpha_0 \alpha_2 - \alpha_0^2) \sum_{i \neq l} p_i^\top (e_i e_i^\top + e_l e_l^\top) p_l + \alpha_1^2 \sum_{i \neq l} p_i^\top (e_i e_l^\top + e_l e_i^\top) p_l + \alpha_0^2 \sum_{i \neq l} p_i^\top p_l,$$

where the third step follows by

$$E_{u \sim D_k} [\phi'(u_i) \phi'(u_i) u_i u_j] = 0$$

and

$$E_{u \sim D_k} [\phi'(u_i) \phi'(u_i) u_j' u_j] = 0$$

for $j' \neq j$. 
For the term $B_1$, we have
\[
B_1 = (\alpha_0 \alpha_2 - \alpha_0^2) \sum_{i \neq l} p_i^\top (e_i e_i^\top + e_l e_l^\top) p_l \\
= 2(\alpha_0 \alpha_2 - \alpha_0^2) \sum_{i \neq l} p_i^\top e_i e_i^\top p_l \\
= 2(\alpha_0 \alpha_2 - \alpha_0^2) \sum_{i=1}^{k} p_i^\top e_i e_i^\top \left( \sum_{l=1}^{k} p_l - p_i \right) \\
= 2(\alpha_0 \alpha_2 - \alpha_0^2) \left( \sum_{i=1}^{k} p_i^\top e_i e_i^\top \sum_{l=1}^{k} p_l - \sum_{i=1}^{k} p_i^\top e_i e_i^\top p_i \right) \\
= 2(\alpha_0 \alpha_2 - \alpha_0^2) (\text{diag}(P)^\top \cdot P \cdot 1 - \|\text{diag}(P)\|^2)
\]

For the term $B_2$, we have
\[
B_2 = \alpha_2^2 \sum_{i \neq l} p_i^\top (e_i e_i^\top + e_l e_l^\top) p_l \\
= \alpha_2^2 \left( \sum_{i \neq l} p_i^\top e_i e_i^\top p_l + \sum_{i \neq l} p_i^\top e_l e_l^\top p_l \right) \\
= \alpha_2^2 \left( \sum_{i=1}^{k} \sum_{l=1}^{k} p_i^\top e_i e_i^\top p_l - \sum_{j=1}^{k} p_j^\top e_j e_j^\top p_j + \sum_{i=1}^{k} \sum_{l=1}^{k} p_i^\top e_l e_l^\top p_l - \sum_{j=1}^{k} p_j^\top e_j e_j^\top p_j \right) \\
= \alpha_2^2 (\|\text{diag}(P)^\top \cdot 1\|^2 - \|\text{diag}(P)\|^2 + \langle P, P^\top \rangle - \|\text{diag}(P)\|^2)
\]

For the term $B_3$, we have
\[
B_3 = \alpha_0^2 \sum_{i \neq l} p_i^\top p_l \\
= \alpha_0^2 \left( \sum_{i=1}^{k} p_i^\top \sum_{l=1}^{k} p_l - \sum_{i=1}^{k} p_i^\top p_i \right) \\
= \alpha_0^2 \left( \|\sum_{i=1}^{k} p_i\|_2^2 - \sum_{i=1}^{k} \|p_i\|^2 \right) \\
= \alpha_0^2 (\|P \cdot 1\|^2 - \|P\|_F^2)
\]

Let $\text{diag}(P)$ denote a length $k$ column vector where the $i$-th entry is the $\langle i, i \rangle$-th entry of...
$P \in \mathbb{R}^{k \times k}$. Furthermore, we can show $A + B$ is,

$$A + B = A + B_1 + B_2 + B_3$$

$$= (\beta_2 - \beta_0)\|\text{diag}(P)\|^2 + \beta_0\|P\|^2_2 + 2(\alpha_0 \alpha_2 - \alpha_0^2)(\text{diag}(P)\top \cdot P \cdot 1 - \|\text{diag}(P)\|^2)$$

$$+ \alpha_1^2((\text{diag}(P)\top \cdot 1)^2 - \|\text{diag}(P)\|^2 + (P, P\top) - \|\text{diag}(P)\|^2) + \alpha_0^2\|P \cdot 1\|^2 - \|P\|^2_F$$

$$= \underbrace{\|\alpha_0 P \cdot 1 + (\alpha_2 - \alpha_0)\text{diag}(P)\|^2 + \alpha_1^2(\text{diag}(P)\top \cdot 1)^2}_{C_1} + \underbrace{\alpha_0^2\|P + P\top - 2\text{diag}(\text{diag}(P))\|^2}_{C_2}$$

$$+ \underbrace{(\beta_0 - \alpha_0^2 - \alpha_1^2)\|P - \text{diag}(\text{diag}(P))\|^2_F + (\beta_2 - \alpha_1^2 - \alpha_2^2)\|\text{diag}(P)\|^2}_{C_3}$$

$$\geq (\beta_0 - \alpha_0^2 - \alpha_1^2)\|P - \text{diag}(\text{diag}(P))\|^2_F + (\beta_2 - \alpha_1^2 - \alpha_2^2)\|\text{diag}(P)\|^2$$

$$\geq \min\{(\beta_0 - \alpha_0^2 - \alpha_1^2), (\beta_2 - \alpha_1^2 - \alpha_2^2)\} \cdot (\|P - \text{diag}(\text{diag}(P))\|^2_F + \|\text{diag}(\text{diag}(P))\|^2)$$

$$= \min\{(\beta_0 - \alpha_0^2 - \alpha_1^2), (\beta_2 - \alpha_1^2 - \alpha_2^2)\} \cdot (\|P - \text{diag}(\text{diag}(P))\|^2_F + \|\text{diag}(\text{diag}(P))\|^2)$$

$$\geq \min\{(\beta_0 - \alpha_0^2 - \alpha_1^2), (\beta_2 - \alpha_1^2 - \alpha_2^2)\} \cdot \|P\|^2_F$$

$$= \rho\|P\|^2_F,$$

where the first step follows by $B = B_1 + B_2 + B_3$, and the second step follows by the definition of $A, B_1, B_2, B_3$ the third step follows by $A + B_1 + B_2 + B_3 = C_1 + C_2 + C_3 + C_4 + C_5$, the fourth step follows by $C_1, C_2, C_3 \geq 0$, the fifth step follows $a \geq \min(a, b)$, the sixth step follows by $\|\text{diag}(P)\|^2 = \|\text{diag}(\text{diag}(P))\|^2$, the seventh step follows by triangle inequality, and the last step follows the definition of $\rho$. \qed

**Claim D.5.** $A + B_1 + B_2 + B_3 = C_1 + C_2 + C_3 + C_4 + C_5$.

**Proof.** The key properties we need are, for two vectors $a, b$, $\|a + b\|^2 = \|a\|^2 + 2\langle a, b \rangle + \|b\|^2$; for two
matrices $A, B, \|A + B\|_F^2 = \|A\|_F^2 + 2\langle A, B \rangle + \|B\|_F^2$. Then, we have

$$
C_1 + C_2 + C_3 + C_4 + C_5
= \frac{1}{2} \left( \|\alpha_0 P \cdot 1\|_F^2 + 2(\alpha_0 \alpha_2 - \alpha_0^2) \langle P \cdot 1, \text{diag}(P) \rangle + (\alpha_2 - \alpha_0)^2 \|\text{diag}(P)\|_F^2 + \alpha_1^2 \langle \text{diag}(P)^\top \cdot 1 \rangle^2 \right) \\
+ \frac{\alpha_1^2}{2} \left( \|P\|_F^2 + 4 \|\text{diag}(\text{diag}(P))\|_F^2 + 2 \langle P, P^\top \rangle - 4(P, \text{diag}(\text{diag}(P))) - 4(P^\top, \text{diag}(\text{diag}(P))) \right) \\
+ (\beta_0 - \alpha_0^2 - \alpha_1^2)(\|P\|_F^2 - 2\langle \text{diag}(\text{diag}(P))\rangle^2 + \|\text{diag}(P)\|_F^2) + (\beta_2 - \alpha_1^2 - \alpha_1^2) \|\text{diag}(P)\|_F^2 \\
= \alpha_0^2 \|P \cdot 1\|_F^2 + 2(\alpha_0 \alpha_2 - \alpha_0^2) \langle P \cdot 1, \text{diag}(P) \rangle + (\alpha_2 - \alpha_0)^2 \|\text{diag}(P)\|_F^2 + \alpha_1^2 \langle \text{diag}(P)^\top \cdot 1 \rangle^2 \\
+ \frac{\alpha_1^2}{2} \left( \|P\|_F^2 + 2\langle P, P^\top \rangle - 8\|\text{diag}(P)\|_F^2 \right) \\
+ (\beta_0 - \alpha_0^2 - \alpha_1^2)(\|P\|_F^2 - 2\langle \text{diag}(\text{diag}(P))\rangle^2 + \|\text{diag}(P)\|_F^2) + (\beta_2 - \alpha_1^2 - \alpha_1^2) \|\text{diag}(P)\|_F^2 \\
= \frac{\beta_2 - \beta_0 - 2(\alpha_2 \alpha_0 - \alpha_0^2 + \alpha_1^2)}{\beta_2 - \beta_0 - 2(\alpha_2 \alpha_0 - \alpha_0^2 + \alpha_1^2)} \|\text{diag}(P)\|_F^2 \\
+ \frac{\beta_2 - \beta_0 - 2(\alpha_2 \alpha_0 - \alpha_0^2 + \alpha_1^2)}{\beta_2 - \beta_0 - 2(\alpha_2 \alpha_0 - \alpha_0^2 + \alpha_1^2)} \|\text{diag}(P)\|_F^2 \\
= \frac{\beta_2 - \beta_0 \|\text{diag}(P)\|_F^2 + \beta_0 \|\text{diag}(P)\|_F^2 + 2(\alpha_0 \alpha_2 - \alpha_0^2) \langle P \cdot 1 - \beta_0 \text{diag}(P)\rangle \|\text{diag}(P)\|_F^2} {\alpha_1^2} \\
+ \frac{\alpha_1^2}{\alpha_1^2} \left( \|\text{diag}(P)^\top \cdot 1\|^2 - \|\text{diag}(P)\|^2 + \langle P, P^\top \rangle - \|\text{diag}(P)\|^2 \right) + \alpha_1^2 \|\text{diag}(P)\|^2 \\
= A + B_1 + B_2 + B_3 \tag{1}
$$

where the second step follows by $\langle P, \text{diag}(\text{diag}(P))\rangle = \|\text{diag}(P)\|^2$ and $\|\text{diag}(\text{diag}(P))\|_F^2 = \|\text{diag}(P)\|^2$.

### D.2.2 Lower Bound on the Eigenvalues of Hessian for Non-orthogonal Case

First we show the lower bound of the eigenvalues. The main idea is to reduce the problem to a $k$-by-$k$ problem and then lower bound the eigenvalues using orthogonal weight matrices.

**Lemma D.6** (Lower bound). If $\phi(z)$ satisfies Property 3.1,3.2 and 3.3, we have the following property for the second derivative of function $f_D(W)$ at $W^*$,

$$
\Omega(v_{\min}^2(\mathcal{D}_k)/(\kappa^2 \lambda)) I \preceq \nabla^2 f_D(W^*).
$$

33
Proof. Let $a \in \mathbb{R}^{dk}$ denote vector $[a_1^\top \ a_2^\top \ \cdots \ a_k^\top]^\top$, let $b \in \mathbb{R}^{dk}$ denote vector $[b_1^\top \ b_2^\top \ \cdots \ b_k^\top]^\top$, and let $c \in \mathbb{R}^{dk}$ denote vector $[c_1^\top \ c_2^\top \ \cdots \ c_k^\top]^\top$. The smallest eigenvalue of the Hessian can be calculated by

$$
\nabla^2 f(W^*) \geq \min_{\|a\|=1} \mathbb{E}_{x \sim D_d} \left[ \left( \sum_{i=1}^k v_i^\top a_i x \cdot \phi'(w_i^\top x) \right)^2 \right] I_{dk} \quad (9)
$$

Note that

$$
\min_{\|a\|=1} \mathbb{E}_{x \sim D_d} \left[ \left( \sum_{i=1}^k v_i^\top a_i x \cdot \phi'(w_i^\top x) \right)^2 \right] = \min_{\sum_i \|b_i/v_i^*\|^2 \neq 0} \mathbb{E}_{x \sim D_d} \left[ \left( \sum_{i=1}^k b_i^\top x \cdot \phi'(w_i^\top x) \right)^2 \right] / \sum_i \|b_i/v_i^*\|^2
$$

by $a_i = b_i/v_i^*$

$$
= \min_{\sum_i \|b_i\|^2 \neq 0} \mathbb{E}_{x \sim D_d} \left[ \left( \sum_{i=1}^k b_i^\top x \cdot \phi'(w_i^\top x) \right)^2 \right] / \sum_i \|b_i\|^2
$$

by $v_{\min} = \min_i |v_i^*|$

$$
= v_{\min}^2 \min_{\|a\|=1} \mathbb{E}_{x \sim D_d} \left[ \left( \sum_{i=1}^k a_i^\top x \cdot \phi'(w_i^\top x) \right)^2 \right] \quad (10)
$$

Let $U \in \mathbb{R}^{d \times k}$ be the orthonormal basis of $W^*$ and let $V = [v_1 \ v_2 \ \cdots \ v_k] = U^\top W^* \in \mathbb{R}^{k \times k}$. Also note that $V$ and $W^*$ have same singular values and $W^* = UV$. We use $U_\perp \in \mathbb{R}^{d \times (d-k)}$ to denote the complement of $U$. For any vector $a_j \in \mathbb{R}^d$, there exist two vectors $b_j \in \mathbb{R}^k$ and $c_j \in \mathbb{R}^{d-k}$ such that

$$
a_j = Ub_j + U_\perp c_j.
$$

We use $D_d$ to denote Gaussian distribution $\mathcal{N}(0, I_d)$, $D_{d-k}$ to denote Gaussian distribution $\mathcal{N}(0, I_{d-k})$, and $D_k$ to denote Gaussian distribution $\mathcal{N}(0, I_k)$. Then we can rewrite formulation (10) (removing $v_{\min}^2$) as

$$
\mathbb{E}_{x \sim D_d} \left[ \left( \sum_{i=1}^k a_i^\top x \cdot \phi'(w_i^\top x) \right)^2 \right] = \mathbb{E}_{x \sim D_d} \left[ \left( \sum_{i=1}^k (b_i^\top U^\top + c_i^\top U_\perp) x \cdot \phi'(w_i^\top x) \right)^2 \right] = A + B + C
$$

34
where

\[
A = \mathbb{E}_{x \sim D_d} \left[ \left( \sum_{i=1}^{k} b_i^\top U^\top x \cdot \phi'(w_i^* x) \right)^2 \right],
\]

\[
B = \mathbb{E}_{x \sim D_d} \left[ \left( \sum_{i=1}^{k} c_i^\top U^\top x \cdot \phi'(w_i^* x) \right)^2 \right],
\]

\[
C = \mathbb{E}_{x \sim D_d} \left[ 2 \left( \sum_{i=1}^{k} b_i^\top U^\top x \cdot \phi'(w_i^* x) \right) \cdot \left( \sum_{i=1}^{k} c_i^\top U^\top x \cdot \phi'(w_i^* x) \right) \right].
\]

We calculate \(A, B, C\) separately. First, we can show

\[
A = \mathbb{E}_{x \sim D_d} \left[ \left( \sum_{i=1}^{k} b_i^\top U^\top x \cdot \phi'(w_i^* x) \right)^2 \right] = \mathbb{E}_{z \sim D_k} \left[ \left( \sum_{i=1}^{k} b_i^\top z \cdot \phi'(v_i^* z) \right)^2 \right].
\]

Second, we can show

\[
B = \mathbb{E}_{x \sim D_d} \left[ \left( \sum_{i=1}^{k} c_i^\top U^\top x \cdot \phi'(w_i^* x) \right)^2 \right] = \mathbb{E}_{s \sim D_{d-k}, z \sim D_k} \left[ \left( \sum_{i=1}^{k} c_i^\top s \cdot \phi'(v_i^* z) \right)^2 \right] 
\]

by defining \(y = \sum_{i=1}^{k} \phi'(v_i^* z) c_i \in \mathbb{R}^{d-k}\)

\[
= \mathbb{E}_{z \sim D_k} \left[ s^\top s \right] \quad \text{by definition of } y
\]

by \(\mathbb{E}[s_j s_j'] = 0\)

\[
= \mathbb{E}_{z \sim D_k} \left[ \sum_{j=1}^{d-k} s_j^2 y_j^2 \right] \quad \text{by } s_j \sim \mathcal{N}(0, 1)
\]

by definition of \(y\)

Third, we have \(C = 0\) since \(U_{\perp}^\top x\) is independent of \(w_i^* x\) and \(U^\top x\). Thus, putting them all together,

\[
\mathbb{E}_{x \sim D_d} \left[ \sum_{i=1}^{k} a_i^\top x \cdot \phi'(w_i^* x) \right] = \mathbb{E}_{z \sim D_k} \left[ \left( \sum_{i=1}^{k} b_i^\top z \cdot \phi'(v_i^* z) \right)^2 \right] + \mathbb{E}_{z \sim D_k} \left[ \left( \sum_{i=1}^{k} \phi'(v_i^* z) c_i \right)^2 \right]
\]
Let us lower bound $A$,

$$A = \mathbb{E}_{z \sim \mathcal{D}_k} \left[ \left( \sum_{i=1}^{k} b_i^T z \cdot \phi'(v_i^T z) \right)^2 \right]$$

$$= \int (2\pi)^{-k/2} \left( \sum_{i=1}^{k} b_i^T z \cdot \phi'(v_i^T z) \right)^2 e^{-\|z\|^2/2} dz$$

$$\geq \xi_1 \int (2\pi)^{-k/2} \left( \sum_{i=1}^{k} b_i^T V'^T s \cdot \phi'(s_i) \right)^2 e^{-\|V'^T s\|^2/2} \cdot |\det(V')| ds$$

$$\geq \xi_2 \int (2\pi)^{-k/2} \left( \sum_{i=1}^{k} b_i^T V'^T u/\sigma_1(V') \cdot \phi'(u_i/\sigma_1(V')) \right)^2 e^{-\|u\|^2/2} \cdot |\det(V')|/\sigma^k_1(V') du$$

$$= \int (2\pi)^{-k/2} \left( \sum_{i=1}^{k} p_i^T u \cdot \phi'(\sigma_k \cdot u_i) \right)^2 e^{-\|u\|^2/2} \frac{1}{\lambda} du$$

$$= \frac{1}{\lambda} \mathbb{E}_{u \sim \mathcal{D}_k} \left[ \left( \sum_{i=1}^{k} p_i^T u \cdot \phi'(\sigma_k \cdot u_i) \right)^2 \right]$$

where $V^\dagger \in \mathbb{R}^{k \times k}$ is the inverse of $V \in \mathbb{R}^{k \times k}$, i.e., $V^\dagger V = I$, $p_i^\dagger = b_i^T V'^T/\sigma_1(V')$ and $\sigma_k = \sigma_k(W^*)$. $\xi_1$ replaces $z$ by $z = V'^T s$, so $v_i^T z = s_i$. $\xi_2$ uses the fact $\|V'^T s\| \leq \sigma_1(V')\|s\|$. $\xi_3$ replaces $s$ by $s = u/\sigma_1(V')$. Note that $\phi'(\sigma_k \cdot u_i)$'s are independent of each other, so we can simplify the analysis. In particular, Lemma D.4 gives a lower bound in this case in terms of $p_i$. Note that $\|p_i\| \geq \|b_i\|/\kappa$. Therefore,

$$\mathbb{E}_{z \sim \mathcal{D}_k} \left[ \left( \sum_{i=1}^{k} b_i^T z \cdot \phi'(v_i^T z) \right)^2 \right] \geq \rho(\sigma_k) \frac{1}{\kappa^2 \lambda} \|b\|^2.$$
For $B$, similar to the proof of Lemma D.4, we have,

$$B = \mathbb{E}_{z \sim D_k} \left[ \left\| \sum_{i=1}^{k} \phi'(v_i^T z) c_i \right\|^2 \right]$$

$$= \int (2\pi)^{-k/2} \left\| \sum_{i=1}^{k} \phi'(v_i^T z) c_i \right\|^2 e^{-\|z\|^2/2} dz$$

$$= \int (2\pi)^{-k/2} \left\| \sum_{i=1}^{k} \phi'(\sigma_k \cdot u_i) c_i \right\|^2 e^{-\|V^T u/\sigma_1(V^t)\|^2/2} \cdot \det(V^T/\sigma_1(V^t)) du$$

$$= \int (2\pi)^{-k/2} \left\| \sum_{i=1}^{k} \phi'(\sigma_k \cdot u_i) c_i \right\|^2 e^{-\|u\|^2/2} \cdot \frac{1}{\lambda} du$$

$$\geq \int (2\pi)^{-k/2} \left\| \sum_{i=1}^{k} \phi'(\sigma_k \cdot u_i) c_i \right\|^2 e^{-\|u\|^2/2} \cdot \frac{1}{\lambda} du$$

$$= \frac{1}{\lambda} \mathbb{E}_{u \sim D_k} \left[ \left\| \sum_{i=1}^{k} \phi'(\sigma_k \cdot u_i) c_i \right\|^2 \right]$$

$$= \frac{1}{\lambda} \left( \sum_{i=1}^{k} \mathbb{E}_{u \sim D_k} [\phi'(\sigma_k \cdot u_i)^2] \sum_{i=1}^{k} \left\| c_i \right\|^2 + \sum_{i \neq l} \mathbb{E}_{u \sim D_k} [\phi'(\sigma_k \cdot u_i)\phi'(\sigma_k \cdot u_l)c_i^T c_l] \right)$$

$$= \frac{1}{\lambda} \left( \mathbb{E}_{z \sim D_1} [\phi'(\sigma_k \cdot z)^2] \sum_{i=1}^{k} \left\| c_i \right\|^2 + \mathbb{E}_{z \sim D_1} [\phi'(\sigma_k \cdot z)^2] \sum_{i \neq l} c_i^T c_l \right)$$

$$= \frac{1}{\lambda} \left( \left( \mathbb{E}_{z \sim D_1} [\phi'(\sigma_k \cdot z)^2] \right)^2 \sum_{i=1}^{k} \left\| c_i \right\|^2 + \left( \mathbb{E}_{z \sim D_1} [\phi'(\sigma_k \cdot z)^2] \right)^2 \left\| c \right\|^2 \right)$$

$$\geq \frac{1}{\lambda} \left( \mathbb{E}_{z \sim D_1} [\phi'(\sigma_k \cdot z)^2] - \mathbb{E}_{z \sim D_1} [\phi'(\sigma_k \cdot z)] \right)^2 \left\| c \right\|^2$$

$$\geq \rho(\sigma_k) \frac{1}{\lambda} \|c\|^2,$$

where the first step follows by definition of Gaussian distribution, the second step follows by replacing $z$ by $z = V^T u/\sigma_1(V^t)$, and then $v_i^T z = u_i/\sigma_1(V^t) = u_i \sigma_k(W^*)$, the third step follows by $\|u\|^2 \geq \frac{1}{\sigma_1(V)} V^T u \|u\|^2$, the fourth step follows by $\det(V^T/\sigma_1(V^t)) = \det(V^T)/\sigma_1(V^t) = 1/\lambda$, the fifth step follows by definition of Gaussian distribution, the ninth step follows by $x^2 \geq 0$ for any $x \in \mathbb{R}$, and the last step follows by Property 3.2.

Note that $1 = \|a\|^2 = \|b\|^2 + \|c\|^2$. Thus, we finish the proof for the lower bound.

\[\square\]

### D.2.3 Upper Bound on the Eigenvalues of Hessian for Non-orthogonal Case

**Lemma D.7** (Upper bound). If $\phi(z)$ satisfies Property 3.1,3.2 and 3.3, we have the following property for the second derivative of function $f_D(W)$ at $W^*$,

$$\nabla^2 f_D(W^*) \leq O(k v_{\text{max}}^2 \sigma_1^{2p}) I.$$
Proof. Similarly, we can calculate the upper bound of the eigenvalues by

\[ \| \nabla^2 f(W^*) \| \]

\[ = \max_{\|a\|=1} | a^\top \nabla^2 f(W^*) a | \]

\[ = v^2_{\max} \max_{\|a\|=1} \mathbb{E}_{x \sim \mathcal{D}_k} \left[ \left( \sum_{i=1}^{k} a_i^\top x \cdot \phi'(w_i^* x) \right)^2 \right] \]

\[ = v^2_{\max} \max_{\|a\|=1} \sum_{i=1}^{k} \sum_{l=1}^{k} \mathbb{E}_{x \sim \mathcal{D}_k} \left[ a_i^\top x \cdot \phi'(w_i^* x) \cdot a_l^\top x \cdot \phi'(w_l^* x) \right] \]

\[ \leq v^2_{\max} \max_{\|a\|=1} \sum_{i=1}^{k} \sum_{l=1}^{k} \left( \mathbb{E}_{x \sim \mathcal{D}_k} [(a_i^\top x)^4] \cdot \mathbb{E}_{x \sim \mathcal{D}_k} [(\phi'(w_i^* x))^4] \cdot \mathbb{E}_{x \sim \mathcal{D}_k} [(a_l^\top x)^4] \cdot \mathbb{E}_{x \sim \mathcal{D}_k} [(\phi'(w_l^* x))^4] \right)^{1/4} \]

\[ \leq v^2_{\max} \max_{\|a\|=1} \sum_{i=1}^{k} \sum_{l=1}^{k} \|a_i\| \cdot \|a_l\| \cdot \|w_i^*\|^p \cdot \|w_l^*\|^p \]

\[ \leq v^2_{\max} \max_{\|a\|=1} \sum_{i=1}^{k} \sum_{l=1}^{k} \|a_i\| \cdot \|a_l\| \cdot \sigma_1^{2p} \]

\[ \leq kv_{\max}^2 \sigma_1^{2p} , \]

where the first inequality follows Hölder’s inequality, the second inequality follows by Property 3.1, the third inequality follows by \( \|w_i^*\| \leq \sigma_1(W^*) \), and the last inequality by Cauchy-Schwarz inequality.

D.3 Error Bound of Hessians near the Ground Truth for Smooth Activations

The goal of this Section is to prove Lemma D.8

**Lemma D.8 (Error bound of Hessians near the ground truth for smooth activations).** Let \( \phi(z) \) satisfy Property 3.1, Property 3.2 and Property 3.3(a). Let \( W \) satisfy \( \| W - W^* \| \leq \sigma_k/2 \). Let \( S \) denote a set of i.i.d. samples from the distribution defined in (1). Then for any \( t \geq 1 \) and \( 0 < \epsilon < 1/2 \), if

\[ |S| \geq \epsilon^{-2d\kappa^2 \tau} \cdot \text{poly}(d, t) \]

then we have, with probability at least \( 1 - 1/d^{D(t)} \),

\[ \| \nabla^2 f_S(W) - \nabla^2 f_{\mathcal{D}}(W^*) \| \leq v^2_{\max} k \sigma_1^p (\epsilon \sigma_1^p + \| W - W^* \|) . \]

**Proof.** This follows by combining Lemma D.10 and Lemma D.11 directly.

D.3.1 Second-order Smoothness near the Ground Truth for Smooth Activations

The goal of this Section is to prove Lemma D.10.

**Fact D.9.** Let \( w_i \) denote the \( i \)-th column of \( W \in \mathbb{R}^{d \times k} \), and \( w_i^* \) denote the \( i \)-th column of \( W^* \in \mathbb{R}^{d \times k} \). If \( \| W - W^* \| \leq \sigma_k(W^*)/2 \), then for all \( i \in [k] \),

\[ \frac{1}{2} \| w_i^* \| \leq \| w_i \| \leq \frac{3}{2} \| w_i^* \| . \]
Proof. Note that if \( \| W - W^* \| \leq \sigma_k(W^*)/2 \), we have \( \sigma_k(W^*)/2 \leq \sigma_i(W) \leq \frac{3}{2} \sigma_1(W^*) \) for all \( i \in [k] \) by Weyl's inequality. By definition of singular value, we have \( \sigma_k(W^*) \leq \| w_i^* \| \leq \sigma_1(W^*) \). By definition of spectral norm, we have \( \| w_i - w_i^* \| \leq \| W - W^* \| \). Thus, we can lower bound \( \| w_i \| \),

\[
\| w_i \| \leq \| w_i^* \| + \| w_i - w_i^* \| \leq \| w_i^* \| + \| W - W^* \| \leq \| w_i^* \| + \sigma_k/2 \leq \frac{3}{2} \| w_i^* \|.
\]

Similarly, we have \( \| w_i \| \geq \frac{1}{2} \| w_i^* \| \).

\( \square \)

Lemma D.10 (Second-order smoothness near the ground truth for smooth activations). If \( \phi(z) \) satisfies Property 3.1, Property 3.2 and Property 3.3(a), then for any \( W \) with \( \| W - W^* \| \leq \sigma_k/2 \), we have

\[
\| \nabla^2 f_D(W) - \nabla^2 f_D(W^*) \| \lesssim k^2 v^2_{\max} \sigma^p \| W - W^* \|.
\]

Proof. Let \( \Delta = \nabla^2 f_D(W) - \nabla^2 f_D(W^*) \). For each \( (i, l) \in [k] \times [k] \), let \( \Delta_{i,l} \in \mathbb{R}^{d \times d} \) denote the \( (i, l) \)-th block of \( \Delta \). Then, for \( i \neq l \), we have

\[
\Delta_{i,l} = \mathbb{E}_{x \sim D_d} \left[ v_i^* v_l^* \left( \phi'(w_i^\top x) \phi'(w_l^\top x) - \phi'(w_i^*\top x) \phi'(w_l^*\top x) \right) xx^\top \right],
\]

and for \( i = l \), we have

\[
\Delta_{i,i} = \mathbb{E}_{x \sim D_d} \left[ \left( \sum_{r=1}^k v_r^* \phi(w_r^\top x) - y \right) v_i^* v_i^* \phi''(w_i^\top x) xx^\top + v_i^* v_i^* \left( \phi'(w_i^\top x) - \phi'(w_i^*\top x) \right) xx^\top \right]
\]

\[
= \mathbb{E}_{x \sim D_d} \left[ \left( \sum_{r=1}^k v_r^* \phi(w_r^\top x) - y \right) v_i^* v_i^* \phi''(w_i^\top x) xx^\top \right] + \mathbb{E}_{x \sim D_d} \left[ v_i^* v_i^* \left( \phi'(w_i^\top x) - \phi'(w_i^*\top x) \right) xx^\top \right]. \tag{11}
\]

In the next a few paragraphs, we first show how to bound the off-diagonal term, and then show how to bound the diagonal term.
First, we consider off-diagonal terms.

\[ ||\Delta_{i,l}|| \]
\[ = \left| \left| \mathbb{E}_{x \sim D_d} \left[ v_i^* v_i \left( \phi'(w_{i}^T x) \phi'(w_{i}^T x) - \phi'(w_{i}^T x) \phi'(w_{i}^T x) \right) xx^T \right] \right| \right| \]
\[ \leq v_{\max}^2 \max_{\|a\|=1} \mathbb{E}_{x \sim D_d} \left[ \left| \phi'(w_{i}^T x) - \phi'(w_{i}^T x) \right| \cdot (x^T a)^2 \right] \]
\[ \leq v_{\max}^2 \max_{\|a\|=1} \mathbb{E}_{x \sim D_d} \left[ \left| \phi'(w_{i}^T x) - \phi'(w_{i}^T x) \right| \left| \phi'(w_{i}^T x) - \phi'(w_{i}^T x) \right| (x^T a)^2 \right] \]
\[ = v_{\max}^2 \max_{\|a\|=1} \left( \mathbb{E}_{x \sim D_d} \left[ \left| \phi'(w_{i}^T x) - \phi'(w_{i}^T x) \right| \left| \phi'(w_{i}^T x) - \phi'(w_{i}^T x) \right| (x^T a)^2 \right] \right) \]
\[ \leq v_{\max}^2 \max_{\|a\|=1} \left( \mathbb{E}_{x \sim D_d} \left[ L_2 \left| (w_i - w_i^*)^T x \right| \phi'(w_{i}^T x) (x^T a)^2 \right] \right) \]
\[ \leq v_{\max}^2 \max_{\|a\|=1} \left( \mathbb{E}_{x \sim D_d} \left[ L_1 \left| \left| (w_i - w_i^*)^T x \right| \phi'(w_{i}^T x) \right| \left| (x^T a)^2 \right| \right] \right) \]
\[ \leq v_{\max}^2 L_1 L_2 \max_{\|a\|=1} \left( \mathbb{E}_{x \sim D_d} \left[ \left| \left| w_i - w_i^* \right| \phi'(w_{i}^T x) \left| \phi'(w_{i}^T x) \right| \left| (x^T a)^2 \right| \right] \right) \]
\[ \leq v_{\max}^2 L_1 L_2 \max_{\|a\|=1} \left( \left| w_i - w_i^* \right| \left| w_i - w_i^* \right| \left| \phi'(w_{i}^T x) \right| \left| (x^T a)^2 \right| \right) \]
\[ \leq v_{\max}^2 L_1 L_2 \sigma_1(W^*) \| W - W^* \| \quad (12) \]

where the first step follows by definition of \( \Delta_{i,l} \), the second step follows by definition of spectral norm and \( v_i^* v_i^* \leq v_{\max}^2 \), the third step follows by triangle inequality, the fourth step follows by linearity of expectation, the fifth step follows by Property 3.3(a), i.e., \( |\phi'(w_{i}^T x) - \phi'(w_{i}^T x)| \leq L_2 \| (w_i - w_i^*)^T x \| \), the sixth step follows by Property 3.1, i.e., \( \phi'(z) \leq L_1 |z|^p \), the seventh step follows by Fact B.6, and the last step follows by \( \| a \|^2 = 1 \), \( \| w_i - w_i^* \| \leq \| W - W^* \| \), \( \| w_i \| \leq \frac{1}{2} \| w_i^* \| \), and \( \| w_i^* \| \leq \sigma_1(W^*) \).

Note that the proof for the off-diagonal terms also applies to bounding the second-term in the diagonal block Eq. (11). Thus we only need to show how to bound the first term in the diagonal
block Eq. (11).

\[
\| \mathbb{E}_{x \sim D_d} \left[ \left( \sum_{r=1}^{k} v^*_r \phi(w_r^\top x) - y \right) v^*_r \phi''(w_i^\top x)x x^\top \right] \| \\
= \| \mathbb{E}_{x \sim D_d} \left[ \left( \sum_{r=1}^{k} v^*_r (\phi(w_r^\top x) - \phi(w_{r_i}^\top x)) \right) v^*_r \phi''(w_i^\top x)x x^\top \right] \|
\]

\[\leq v^2_{\max} \sum_{r=1}^{k} \max_{||a||=1} \mathbb{E}_{x \sim D_d} \left[ |\phi(w_r^\top x) - \phi(w_{r_i}^\top x)|L_2(x^\top a)^2 \right] \]

\[\leq v^2_{\max} \sum_{r=1}^{k} \max_{||a||=1} \mathbb{E}_{x \sim D_d} \left[ \max_{z \in [w_r^\top x,w_{r_i}^\top x]} |\phi'(z)| \cdot |(w_r - w_{r_i})^\top x| \cdot (x^\top a)^2 \right] \]

\[\leq v^2_{\max} L_2 \sum_{r=1}^{k} \max_{||a||=1} \mathbb{E}_{x \sim D_d} \left[ \max_{z \in [w_r^\top x,w_{r_i}^\top x]} L_1 |z|^p \cdot |(w_r - w_{r_i})^\top x| \cdot (x^\top a)^2 \right] \]

\[\leq v^2_{\max} L_1 L_2 \sum_{r=1}^{k} \max_{||a||=1} \mathbb{E}_{x \sim D_d} \left[ (|w_r|^p + |w_{r_i}^\top x|^p) \cdot |(w_r - w_{r_i})^\top x| \cdot (x^\top a)^2 \right] \]

\[\leq k v^2_{\max} L_1 L_2 \sigma^p(W^*) \|W - W^*\|, \] (13)

where the first step follows by \( y = \sum_{r=1}^{k} v^*_r \phi(w_{r_i}^\top x) \), the second step follows by definition of spectral norm and \( v^*_r v^*_r \leq |v^*_{\max}|^2 \), the third step follows by Property 3.3(a), i.e., \( |\phi''(w_i^\top x)| \leq L_2 \), the fourth step follows by \( |\phi(w_r^\top x) - \phi(w_{r_i}^\top x)| \leq \max_{z \in [w_r^\top x,w_{r_i}^\top x]} |\phi'(z)| |(w_r - w_{r_i})^\top x| \), the fifth step follows Property 3.1, i.e., \( |\phi'(z)| \leq L_1 |z|^p \), the sixth step follows by \( \max_{z \in [w_r^\top x,w_{r_i}^\top x]} |z|^p \leq (|w_r^\top x|^p + |w_{r_i}^\top x|^p) \), the seventh step follows by Fact B.6.
Putting it all together, we can bound the error by
\[
\|\nabla^2 f_D(W) - \nabla^2 f_D(W^*)\|
\leq \max_{\|a\|=1} a^\top (\nabla^2 f_D(W) - \nabla^2 f_D(W^*)) a
\]
\[
= \max_{\|a\|=1} \sum_{i=1}^k \sum_{l=1}^k a_i^\top \Delta_i,a_l
\]
\[
= \max_{\|a\|=1} \left( \sum_{i=1}^k a_i^\top \Delta_i,a_i + \sum_{i \neq l} a_i^\top \Delta_i,a_l \right)
\]
\[
\leq \max_{\|a\|=1} \left( \sum_{i=1}^k \|\Delta_i,a_i\|a_i \|^2 + \sum_{i \neq l} \|\Delta_i,a_l\|\|a_i\|\|a_l\| \right)
\]
\[
\leq \max_{\|a\|=1} \left( 2 \sum_{i=1}^k C_1 \|a_i\|^2 + \sum_{i \neq l} C_2 \|a_i\|\|a_l\| \right)
\]
\[
= \max_{\|a\|=1} \left( 2 \sum_{i=1}^k \|a_i\|^2 + C_2 \left( \sum_{i=1}^k \|a_i\| \right)^2 - \sum_{i=1}^k \|a_i\|^2 \right) \]
\[
\leq \max_{\|a\|=1} \left( 2 \sum_{i=1}^k \|a_i\|^2 + C_2 \left( \sum_{i=1}^k \|a_i\|^2 - \sum_{i=1}^k \|a_i\|^2 \right) \right)
\]
\[
= \max_{\|a\|=1} \left( 2 \sum_{i=1}^k \|a_i\|^2 + C_2 \left( \sum_{i=1}^k \|a_i\|^2 - \sum_{i=1}^k \|a_i\|^2 \right) \right)
\]
\[
\leq C_1 + C_2 (k - 1)
\]
\[
\lesssim k^2 \sigma_{\max}^2 L_1 L_2 \sigma_1^p (W^*) \|W - W^*\|
\]
where the first step follows by definition of spectral norm and \(a\) denotes a vector \(\in \mathbb{R}^{dk}\), the first inequality follows by \(\|A\| = \max_{\|x\| \neq 0, \|y\| \neq 0} \frac{x^\top A y}{\|x\|\|y\|}\), the second inequality follows by \(\|\Delta_i,a_i\| \leq C_1\) and \(\|\Delta_i,a_l\| \leq C_2\), the third inequality follows by Cauchy-Schwarz inequality, the eighth step follows by \(\sum_{i=1}^k \|a_i\|^2 = 1\), and the last step follows by Eq (12) and (13).

D.3.2 Empirical and Population Difference for Smooth Activations

The goal of this Section is to prove Lemma D.11. For each \(i \in [k]\), let \(\sigma_i\) denote the \(i\)-th largest singular value of \(W^* \in \mathbb{R}^{d \times k}\).

Note that Bernstein inequality requires the spectral norm of each random matrix to be bounded almost surely. However, since we assume Gaussian distribution for \(x\), \(\|x\|^2\) is not bounded almost surely. The main idea is to do truncation and then use Matrix Bernstein inequality. Details can be found in Lemma B.7 and Corollary B.8.

**Lemma D.11** (Empirical and population difference for smooth activations). *Let \(\phi(z)\) satisfy Property 3.1,3.2 and 3.3(a). Let \(W\) satisfy \(\|W - W^*\| \leq \sigma_k/2\). Let \(S\) denote a set of i.i.d. samples from distribution \(D\) (defined in (1)). Then for any \(t \geq 1\) and \(0 < \epsilon < 1/2\), if*
\[|S| \geq \epsilon^{-2} dk^2 \tau \cdot \text{poly}(\log d, t)\]
*then we have, with probability at least \(1 - d^{-\Omega(t)}\),
\[\|\nabla^2 f_S(W) - \nabla^2 f_D(W)\| \lesssim t^2 \sigma_{\max}^2 \sigma_1^p (\epsilon \sigma_1^p + \|W - W^*\|).\]
Proof. Define $\Delta = \nabla^2 f_d(W) - \nabla^2 \hat{f}_S(W)$. Let’s first consider the diagonal blocks. Define

$$
\Delta_{i,i} = \mathbb{E}_{x \sim \mathcal{D}_d} \left[ \left( \sum_{r=1}^{k} v_r^* \phi(w_r^\top x) - y \right) v_r^* \phi''(w_r^\top x)xx^\top + v_r^* \phi'(w_r^\top x)x \right]
- \left( \frac{1}{n} \sum_{j=1}^{n} \left( \sum_{r=1}^{k} v_r^* \phi(w_r^\top x_j) - y \right) v_r^* \phi''(w_r^\top x_j)x_jx_j^\top \right).
$$

Let’s further decompose $\Delta_{i,i}$ into $\Delta_{i,i} = \Delta_{i,i}^{(1)} + \Delta_{i,i}^{(2)}$, where

$$
\Delta_{i,i}^{(1)} = \mathbb{E}_{x \sim \mathcal{D}_d} \left[ \left( \sum_{r=1}^{k} v_r^* \phi(w_r^\top x) - y \right) v_r^* \phi''(w_r^\top x)xx^\top \right] - \frac{1}{n} \sum_{j=1}^{n} \left( \sum_{r=1}^{k} v_r^* \phi(w_r^\top x_j) - y \right) v_r^* \phi''(w_r^\top x_j)x_jx_j^\top
= v_i^* \sum_{r=1}^{k} \mathbb{E}_{x \sim \mathcal{D}_d} \left[ \phi(w_r^\top x) - \phi(w_r^\top x) \phi''(w_r^\top x)xx^\top \right]
- \frac{1}{n} \sum_{j=1}^{n} \left( \phi(w_r^\top x_j) - \phi(w_r^\top x_j) \phi''(w_r^\top x_j)x_jx_j^\top \right),
$$

and

$$
\Delta_{i,i}^{(2)} = \mathbb{E}_{x \sim \mathcal{D}_d} \left[ v_r^* \phi''(w_r^\top x)x \right] - \frac{1}{n} \sum_{j=1}^{n} \left[ v_r^* \phi''(w_r^\top x_j)x_jx_j^\top \right].
$$

(14)

The off-diagonal block is

$$
\Delta_{i,l} = v_i^* v_l^* \left( \mathbb{E}_{x \sim \mathcal{D}_d} \left[ \phi'(w_i^\top x) \phi'(w_l^\top x)xx^\top \right] - \frac{1}{n} \sum_{j=1}^{n} \phi'(w_i^\top x_j) \phi'(w_l^\top x_j)x_jx_j^\top \right).
$$

Combining Claims. D.12, D.13 and D.14, and taking a union bound over $k^2$ different $\Delta_{i,j}$, we obtain if $n \geq \epsilon^{-2} \kappa^2(W^*) \tau d \cdot \text{poly}(t, \log d)$ for any $\epsilon \in (0,1/2)$, with probability at least $1 - 1/d^d$,

$$
\|\nabla^2 \hat{f}_S(W) - \nabla^2 f(W)\| \lesssim v^2_{\max} k \sigma^p(W^*)(\epsilon \sigma^p(W^*) + \|W - W^*\|)
$$

Claim D.12. For each $i \in [k]$, if $n \geq d \text{poly}(\log d, t)$

$$
\|\Delta_{i,i}^{(1)}\| \lesssim k v^2_{\max} \sigma^p(W^*)\|W - W^*\|
$$

holds with probability $1 - 1/d^{4t}$.

Proof. For each $r \in [k]$, we define function $\hat{B}_r : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$,

$$
\hat{B}_r(x) = L_1 L_2 \cdot (|w_r^\top x|^p + |w_r^\top x|^p) \cdot (w_r - w_r^*)^\top x \cdot x^\top.
$$
According to Properties 3.1, 3.2 and 3.3(a), we have for each \( x \in S \),
\[
-\hat{B}_r(x) \preceq (\phi(w_r^\top x) - \phi(w_r^* x))\phi'(w_i^\top x)xx^\top \preceq \hat{B}_r(x)
\]
Therefore,
\[
\Delta_i^{(1)} \preceq v_{\max}^2 \sum_{r=1}^k \left( \mathbb{E}_{x \sim \mathcal{D}_d}[\hat{B}_r(x)] + \frac{1}{|S|} \sum_{x \in S} \hat{B}_r(x) \right).
\]

Let \( h_r(x) = L_1 L_2 |w_r^\top x|^p \cdot (w_r - w_r^*)^\top x \). Let \( \overline{B}_r = \mathbb{E}_{x \sim \mathcal{D}_d}[h_r(x)xx^\top] \). Define function \( B_r(x) = h_r(x)xx^\top \).

(I) Bounding \(|h_r(x)|\).

According to Fact B.1, we have for any constant \( t \geq 1 \), with probability \( 1 - 1/(nd^4t) \),
\[
|h_r(x)| \leq L_1 L_2 |w_r^\top x|^p |(w_r - w_r^*)^\top x| \preceq L_1 L_2 |w_r|^p \|w_r - w_r^*\|(t \log n)^{(p+1)/2}.
\]

(II) Bounding \( \|\overline{B}_r\| \).

\[
\|\overline{B}_r\| \geq \mathbb{E}_{x \sim \mathcal{D}_d} \left[ L_1 L_2 |w_r^\top x|^p (w_r - w_r^*)^\top x \left( \frac{(w_r - w_r^*)^\top x}{\|w_r - w_r^*\|} \right)^2 \right] \geq L_1 L_2 |w_r|^p \|w_r - w_r^*\|,
\]

where the first step follows by definition of spectral norm, and last step follows by Fact B.6. Using Fact B.6, we can also prove an upper bound \( \|\overline{B}_r\| \), \( \|\overline{B}_r\| \preceq L_1 L_2 |w_r|^p \|w_r - w_r^*\| \).

(III) Bounding \( \left( \mathbb{E}_{x \sim \mathcal{D}_d}[h^4(x)] \right)^{1/4} \).

Using Fact B.6, we have
\[
\left( \mathbb{E}_{x \sim \mathcal{D}_d}[h^4(x)] \right)^{1/4} = L_1 L_2 \left( \mathbb{E}_{x \sim \mathcal{D}_d} \left[ (|w_r^\top x|^p (w_r - w_r^*)^\top x)^4 \right] \right)^{1/4} \preceq L_1 L_2 |w_r|^p \|w_r - w_r^*\|.
\]

By applying Corollary B.8, if \( n \geq \epsilon^{-2}d \text{poly}(\log d, t) \), then with probability \( 1 - 1/d^4t \),
\[
\left\| \mathbb{E}_{x \sim \mathcal{D}_d} \left[ w_r^\top x|^p \cdot (w_r - w_r^*)^\top x \cdot xx^\top \right] - \frac{1}{|S|} \sum_{x \in S} |w_r^\top x|^p \cdot (w_r - w_r^*)^\top x \cdot xx^\top \right\|
\]
\[
= \|\overline{B}_r - \frac{1}{|S|} \sum_{x \in S} B_r(x)\|
\]
\[
\leq \epsilon \|\overline{B}_r\|
\]
\[
\preceq \epsilon \|w_r|^p \|w_r - w_r^*\|.
\]

If \( \epsilon \leq 1/2 \), we have
\[
\|\Delta_i^{(1)}\| \preceq \sum_{i=1}^k v_{\max}^2 \|\overline{B}_r\| \preceq k v_{\max}^2 \sigma_1^p(W^*) \|W - W^*\|
\]

\[\square\]
Claim D.13. For each $i \in [k]$, if $n \geq \epsilon^{-2}d\tau \log(\log d, t)$, then
\[ \|\Delta_{i,i}^{(2)}\| \lesssim \epsilon v_{\max}^2\sigma_1^{2p} \]
holds with probability $1 - 1/d^{4t}$.

Proof. Recall the definition of $\Delta_{i,i}^{(2)}$,
\[ \Delta_{i,i}^{(2)} = \E_{s \sim D_d}[v_i^s \varphi^2(w_i^\top x)x] - \frac{1}{|S|} \sum_{x \in S} [v_i^s \varphi^2(w_i^\top x)x] \]
Let $h_i(x) = \varphi^2(w_i^\top x)$. Let $\overline{B}_i = \E_{x \sim D_d}[h_i(x)x\top]$ Define function $B_i(x) = h_i(x)x\top$.

(I) Bounding $|h_i(x)|$.
For any constant $t \geq 1$, $(\varphi'(w_i^\top x))^2 \leq L_i^2|w_i^\top x|^{2p} \leq L_i^2\|w_i\|^{2p}p^p \log^p(n)$ with probability $1 - 1/(nd^{4t})$ according to Fact B.1.

(II) Bounding $\|\overline{B}_i\|$.
\[
\left\| \E_{x \sim D_d} [\varphi^2(w_i^\top x)x]\right\| = \max_{\|a\|=1} \E_{x \sim D_d} [\varphi^2(w_i^\top x)(x\top a)^2] \\
= \max_{\|a\|=1} \E_{x \sim D_d} [\varphi^2(w_i^\top x)\left(\alpha w_i^\top x + \beta x\top v\right)^2] \\
= \max_{\alpha^2 + \beta^2 = 1, \|v\|=1} \E_{x \sim D_d} [\varphi^2(w_i^\top x)\left((\alpha w_i^\top x)^2 + (\beta x\top v)^2\right)] \\
= \max_{\alpha^2 + \beta^2 = 1} \left(\alpha^2 \E_{z \sim D_1} [\varphi^2(\|w_i\|z)^2] + \beta^2 \E_{z \sim D_1} [\varphi^2(\|w_i\|z)]\right) \\
= \max \left(\E_{z \sim D_1} [\varphi^2(\|w_i\|z)^2], \E_{z \sim D_1} [\varphi^2(\|w_i\|z)]\right)
\]
where $w_i = |w_i|$ and $v$ is a unit vector orthogonal to $w_i$ such that $a = \alpha w_i + \beta v$. Now from Property 3.2, we have
\[ \rho(\|w_i\|) \leq \left\| \E_{x \sim D_d} [\varphi^2(w_i^\top x)x]\right\| \lesssim L_i^2\|w_i\|^{2p}. \]

(III) Bounding $(\E_{x \sim D_d}[h_i^k(x)])^{1/4}$.
\[ \left( \E_{x \sim D_d}[h_i^k(x)] \right)^{1/4} = \left( \E_{x \sim D_d}[\varphi^2(w_i^\top x)] \right)^{1/4} \lesssim L_i^2\|w_i\|^{2p}. \]

By applying Corollary B.8, we have, for any $0 < \epsilon < 1$, if $n \geq \epsilon^{-2}d\|w_i\|^{4p}/\rho^2(\|w_i\|) \log(\log d, t)$ the following bound holds
\[ \left\| \overline{B}_i - \frac{1}{|S|} \sum_{x \in S} B_i(x) \right\| \leq \epsilon \|\overline{B}_i\|, \]
with probability at least $1 - 1/d^{4t}$.

Therefore, if $n \geq \epsilon^{-2}d\tau \log(\log d, t)$, where $\tau = \min_{\alpha \in [1/2, 3/2]} \rho^{3}(\alpha)$, we have with probability $1 - 1/d^{4t}$
\[ \|\Delta_{i,i}^{(2)}\| \lesssim \epsilon v_{\max}^2\sigma_1^{2p} \]
\qed
Claim D.14. For each $i \in [k], j \in [k], i \neq j$, if $n \geq \epsilon^{-2} \kappa^2 \tau d \log d (t)$, then

$$\|\Delta_{i,j}\| \leq \epsilon v^2_{\max} \sigma_1^2 (W^*)$$

holds with probability $1 - 1/d^4$.

Proof. Recall the definition of off-diagonal blocks $\Delta_{i,j}$,

$$\Delta_{i,j} = v_i^* v_j^* \left( \mathbb{E}_{x \sim \mathcal{D}_d} [\phi'(w_i^T x)\phi'(w_j^T x)xx^T] - \frac{1}{|S|} \sum_{x \in S} \phi'(w_i^T x)\phi'(w_j^T x)xx^T \right)$$

Let $h_{i,l}(x) = \phi'(w_i^T x)\phi'(w_l^T x)$. Define function $B_{i,l}(x) = h_{i,l}(x)xx^T$. Let $\mathcal{B}_{i,l} = \mathbb{E}_{x \sim \mathcal{D}_d} [h_{i,l}(x)xx^T]$.

(I) Bounding $|h_{i,l}(x)|$.

For any constant $t \geq 1$, we have with probability $1 - 1/(nd^4)$

$$|h_{i,l}(x)| = |\phi'(w_i^T x)\phi'(w_l^T x)|$$

$$\leq L_2^2 ||w_i^T||^p ||w_l^T||^p$$

$$\leq L_2^2 ||w_i||^p ||w_l||^p (t \log n)^p$$

$$\lesssim L_2^2 \sigma_1^{2p} (t \log n)^p$$

where the third step follows by Fact B.1.

(II) Bounding $\|\mathcal{B}_{i,l}\|$.

Let $U \in \mathbb{R}^{d \times 2}$ be the orthogonal basis of span $\{w_i, w_l\}$ and $U_\perp \in \mathbb{R}^{d \times (d-2)}$ be the complementary matrix of $U$. Let matrix $V \in \mathbb{R}^{2 \times 2}$ denote $U^\top [w_i \ w_l]$, then $UV = [w_i \ w_l] \in \mathbb{R}^{d \times 2}$. Given any vector $\alpha \in \mathbb{R}^d$, there exist vectors $b \in \mathbb{R}^2$ and $c \in \mathbb{R}^{d-2}$ such that $\alpha = Ub + U_\perp c$. We can simplify $\|\mathcal{B}_{i,l}\|$ in the following way,

$$\|\mathcal{B}_{i,l}\| = \left\| \mathbb{E}_{x \sim \mathcal{D}_d} [\phi'(w_i^T x)\phi'(w_l^T x)xx^T] \right\|$$

$$= \max_{||a||=1} \mathbb{E}_{x \sim \mathcal{D}_d} [\phi'(w_i^T x)\phi'(w_l^T x) (x^T a)^2]$$

$$= \max_{||b||^2 + ||c||^2 = 1} \mathbb{E}_{x \sim \mathcal{D}_d} [\phi'(w_i^T x)\phi'(w_l^T x) (b^T U^\top x + c^T U_\perp^\top x)^2]$$

$$= \max_{||b||^2 + ||c||^2 = 1} \mathbb{E}_{x \sim \mathcal{D}_d} [\phi'(w_i^T x)\phi'(w_l^T x) ((b^T U^\top x)^2 + (c^T U_\perp^\top x)^2)]$$

$$= \max_{||b||^2 + ||c||^2 = 1} \left( \mathbb{E}_{z \sim \mathcal{D}_2} [\phi'(v_1^T z)\phi'(v_2^T z) (b^T z)^2] + \mathbb{E}_{z \sim \mathcal{D}_2, s \sim \mathcal{D}_{d-2}} [\phi'(v_1^T z)\phi'(v_2^T z) (c^T s)^2] \right)$$

$$= \max_{||b||^2 + ||c||^2 = 1} \left( A_1 + A_2 \right)$$
We can lower bound the term $A_1$,

$$A_1 = \mathbb{E}_{z \sim D_2} \left[ \phi'(v_1^T z) \phi'(v_2^T z) (b^T z)^2 \right]$$

$$= \int (2\pi)^{-1} \phi'(v_1^T z) \phi'(v_2^T z) (b^T z)^2 e^{-\|z\|^2/2} dz$$

$$= \int (2\pi)^{-1} \phi'(s_1) \phi'(s_2) (b^T V^T s)^2 e^{-\|V^T s\|^2/2} \cdot |\det(V^T)| ds$$

$$\geq \int (2\pi)^{-1} \phi'(s_1) \phi'(s_2) (b^T V^T s)^2 e^{-\sigma_1^2(V^T)\|s\|^2/2} \cdot |\det(V^T)| ds$$

$$= \int (2\pi)^{-1} \phi'(u_1/\sigma_1(V^T)) \phi'(u_2/\sigma_1(V^T)) \cdot (b^T V^T u/\sigma_1(V^T))^2 e^{-\|u\|^2/2} |\det(V^T)|/\sigma_1^2(V^T) du$$

$$= \frac{\sigma_2(V)}{\sigma_1(V)} \mathbb{E}_{u \sim D_2} \left[ (p^T u)^2 \phi'(\sigma_2(V) \cdot u_1) \phi'(\sigma_2(V) \cdot u_2) \right]$$

$$= \frac{\sigma_2(V)}{\sigma_1(V)} \mathbb{E}_{u \sim D_2} \left[ (p^T u)^2 \phi'(\sigma_2(V) \cdot u_1) \phi'(\sigma_2(V) \cdot u_2) \right]$$

$$\geq \frac{\sigma_2(V)}{\sigma_1(V)} |b|^2 \mathbb{E}_{z \sim D_1} \left[ \phi'(\sigma_2(V) \cdot z) z^2 \right] \mathbb{E}_{z \sim D_1} \left[ \phi'(\sigma_2(V) \cdot z) \right]$$

For the term $A_2$, similarly we have

$$A_2 = \mathbb{E}_{z \sim D_2, s \sim D_{d-2}} \left[ \phi'(v_1^T z) \phi'(v_2^T z) (c^T s)^2 \right]$$

$$= \mathbb{E}_{z \sim D_2} \left[ \phi'(v_1^T z) \phi'(v_2^T z) \right] \mathbb{E}_{s \sim D_{d-2}} \left[ (c^T s)^2 \right]$$

$$= |c|^2 \mathbb{E}_{z \sim D_2} \left[ \phi'(v_1^T z) \phi'(v_2^T z) \right]$$

$$\geq |c|^2 \frac{\sigma_2(V)}{\sigma_1(V)} \left( \mathbb{E}_{z \sim D_1} \left[ \phi'(\sigma_2(V) \cdot z) \right] \right)^2$$

For simplicity, we just set $\|b\| = 1$ and $\|c\| = 0$ to lower bound,
where the second step is from Property 3.2 and the fact that \( \sigma_k/2 \leq \sigma_2(V) \leq \sigma_1(V) \leq 3\sigma_1/2. 
For the upper bound of \( \|E_{x \sim D}\left[\phi'(w_i^T x)\phi'(w_i^T x)xx^T\right]\| \), we have

\[
E_{z \sim D_2}\left[\phi'(v_1^T z)\phi'(v_2^T z)(b^T z)^2\right] \leq L_1^2 E_{z \sim D_2}\left[|v_1^T z|^p \cdot |v_2^T z|^p \cdot |b^T z|^2\right] \lesssim L_1^2\|v_1\|^p\|v_2\|^p\|b\|^2 \lesssim L_2^2\sigma_1^{2p}
\]

where the first step follows by Property 3.1, the second step follows by Fact B.6, and the last step follows by \( \|v_1\| \leq \sigma_1, \|v_2\| \leq \sigma_1 \) and \( \|b\| \leq 1 \). Similarly, we can upper bound,

\[
E_{z \sim D_2, s \sim D_{d-2}}\left[\phi'(v_1^T z)\phi'(v_2^T z)(\epsilon^T s)^2\right] = \|c\|^2 E_{z \sim D_2}\left[\phi'(v_1^T z)\phi'(v_2^T z)\right] \lesssim L_1^2\sigma_1^{2p}
\]

Thus, we have

\[
\left\|E_{x \sim D_d}\left[\phi'(w_i^T x)\phi'(w_i^T x)xx^T\right]\right\| \lesssim L_1^2\sigma_1^{2p} \lesssim \sigma_1^{2p}.
\]

(III) Bounding \( \left(\left(E_{x \sim D_d}[h^4_{i,t}(x)]\right)\right)^{1/4} \).

\[
\left(\left(E_{x \sim D_d}[h^4_{i,t}(x)]\right)\right)^{1/4} = \left(\left(E_{x \sim D_d}[\phi'^4(w_i^T x) \cdot \phi'^4(w_i^T x)]\right)\right)^{1/4} \lesssim L_2^2\|w_i\|^p\|w_i\|^p \lesssim L_2^2\sigma_1^{2p}.
\]

Therefore, applying Corollary B.8, we have, if \( n \geq \epsilon^{-2}\kappa^2(W^*)^{-d} poly(\log d, t) \), then

\[
\|\Delta_{i,j}\| \leq \epsilon v_{\max}^2\sigma_1^{2p}(W^*)
\]

holds with probability at least \( 1 - 1/d^4t \). \( \square \)

D.4 Error Bound of Hessians near the Ground Truth for Non-smooth Activations

The goal of this Section is to prove Lemma D.15,

**Lemma D.15** (Error bound of Hessians near the ground truth for non-smooth activations). Let \( \phi(z) \) satisfy Property 3.1,3.2 and 3.3(b). Let \( W \) satisfy \( \|W - W^*\| \leq \sigma_k/2 \). Let \( S \) denote a set of i.i.d. samples from the distribution defined in (1). Then for any \( t \geq 1 \) and \( 0 < \epsilon < 1/2 \), if

\[
|S| \geq \epsilon^{-2}\kappa^2d\log d, t
\]

with probability at least \( 1 - d^{-\Omega(t)} \),

\[
\|\nabla^2 f_S^*(W) - \nabla^2 f_D^*(W^*)\| \lesssim v_{\max}^2k\sigma_1^{2p}(\epsilon + (\sigma_k^{-1}\|W - W^*\|)^{1/2}).
\]

**Proof.** As we noted previously, when Property 3.3(b) holds, the diagonal blocks of the empirical Hessian can be written as, with probability 1,

\[
\frac{\partial^2 f_S(W)}{\partial w_i^2} = \frac{1}{|S|} \sum_{(x,y) \in S} (v_i^s \phi'(w_i^T x))^2 xx^T
\]

48
We construct a matrix $H \in \mathbb{R}^{dk \times dk}$ with $i, l$-th block as

$$H_{i,l} = v_i^* v_l^* \mathbb{E}_{x \sim D_d} \left[ \phi'(w_i^\top x) \phi'(w_l^\top x) x x^\top \right] \in \mathbb{R}^{d \times d}, \quad \forall i \in [k], l \in [k].$$

Note that $H \neq \nabla^2 f_D(W)$. However, we can still bound $\|H - \nabla^2 \hat{f}_S(W)\|$ and $\|H - \nabla^2 f_D(W^*)\|$ when we have enough samples and $\|W - W^*\|$ is small enough. The proof for $\|H - \nabla^2 \hat{f}_S(W)\|$ basically follows the proof of Lemma D.11 as $\Delta_{i,l}^{(2)}$ in Eq. (14) and $\Delta_{i,l}$ in Eq. (16) forms the blocks of $H - \nabla^2 \hat{f}_S(W)$ and we can bound them without smoothness of $\phi(\cdot)$.

Now we focus on $H - \nabla^2 f_D(W^*)$. We again consider each block.

$$\Delta_{i,l} = \mathbb{E}_{x \sim D_d} \left[ v_i^* v_l^* \left( \phi'(w_i^\top x) \phi'(w_l^\top x) - \phi'(w_i^* \top x) \phi'(w_l^* \top x) \right) x x^\top \right].$$

We used the boundedness of $\phi''(z)$ to prove Lemma D.10. Here we can’t use this condition. Without smoothness, we will stop at the following position.

$$\left\| \mathbb{E}_{x \sim D_d} \left[ v_i^* v_l^* \left( \phi'(w_i^\top x) \phi'(w_l^\top x) - \phi'(w_i^* \top x) \phi'(w_l^* \top x) \right) x x^\top \right] \right\|$$

$$\leq |v_i^* v_l^*| \max_{||a||=1} \mathbb{E}_{x \sim D_d} \left[ \phi'(w_i^\top x) \phi'(w_l^\top x) - \phi'(w_i^* \top x) \phi'(w_l^* \top x) \right] (x^\top a)^2$$

$$\leq |v_i^* v_l^*| \max_{||a||=1} \mathbb{E}_{x \sim D_d} \left[ |\phi'(w_i^\top x) - \phi'(w_i^* \top x)| \phi'(w_l^\top x) \right]$$

$$+ \left| \phi'(w_i^* \top x) \phi'(w_l^\top x) - \phi'(w_i^* \top x) \phi'(w_l^\top x) \right| (x^\top a)^2$$

$$= |v_i^* v_l^*| \max_{||a||=1} \left( \mathbb{E}_{x \sim D_d} \left[ |\phi'(w_i^\top x) - \phi'(w_i^* \top x)| \phi'(w_l^\top x) \right] (x^\top a)^2 \right)$$

$$+ \mathbb{E}_{x \sim D_d} \left[ |\phi'(w_i^* \top x) - \phi'(w_i^* \top x)| \phi'(w_l^\top x) \right] (x^\top a)^2 \right).$$

(17)

where the first step follows by definition of spectral norm, the second step follows by triangle inequality, and the last step follows by linearity of expectation.

Without loss of generality, we just bound the first term in the above formulation. Let $U$ be the orthogonal basis of span$(w_i, w_i^*, w_l)$. If $w_i, w_i^*, w_l$ are independent, $U$ is $d$ by rank$(\text{span}(w_i, w_i^*, w_l))$. Without loss of generality, we assume $U = \text{span}(w_i, w_i^*, w_l)$ is $d$ by rank.

Let $[v_i^* v_l^*] = U^\top [w_i^* w_i^* w_l] \in \mathbb{R}^{3 \times 3}$, and $[u_i^* u_i^* u_l] = U^\top [w_i^* w_i^* w_l] \in \mathbb{R}^{d \times d}$. Let $a = Ub + U_{\perp} c$, where $U_{\perp} \in \mathbb{R}^{d \times (d-3)}$ is the complementary matrix of $U$.

$$\mathbb{E}_{x \sim D_d} \left[ |\phi'(w_i^\top x) - \phi'(w_i^* \top x)| \phi'(w_l^\top x) \right] (x^\top a)^2 \right]$$

$$= \mathbb{E}_{x \sim D_d} \left[ |\phi'(w_i^\top x) - \phi'(w_i^* \top x)| \phi'(w_l^\top x) \right] (U^\top Ub)^2 + (x^\top U_{\perp} c)^2 \right]$$

$$\leq \mathbb{E}_{x \sim D_d} \left[ |\phi'(w_i^\top x) - \phi'(w_i^* \top x)| \phi'(w_l^\top x) \right] (U^\top Ub)^2 + (x^\top U_{\perp} c)^2 \right]$$

$$+ \mathbb{E}_{x \sim D_d} \left[ |\phi'(w_i^\top x) - \phi'(w_i^* \top x)| \phi'(w_l^\top x) \right] (x^\top U_{\perp} c)^2 \right]$$

$$= \mathbb{E}_{z \sim D_{d-3}} \left[ |\phi'(v_i^\top z) - \phi'(v_i^* \top z)| \phi'(v_l^\top z) \right] (z^\top b)^2 \right]$$

$$+ \mathbb{E}_{y \sim D_{d-3}} \left[ |\phi'(u_i^\top y) - \phi'(u_i^* \top y)| \phi'(u_l^\top y) \right] (y^\top c)^2 \right]$$

(18)
where the first step follows by $a = Ub + U_1c$, the last step follows by $(a + b)^2 \leq 2a^2 + 2b^2$.

By Property 3.3(b), we have $e$ exceptional points which have $\phi''(z) \neq 0$. Let these $e$ points be $p_1, p_2, \ldots, p_e$. Note that if $v_i^Tz$ and $v_i^*Tz$ are not separated by any of these exceptional points, i.e., there exists no $j \in [e]$ such that $v_i^Tz \leq p_j \leq v_i^*Tz$ or $v_i^*Tz \leq p_j \leq v_i^Tz$, then we have $\phi'(v_i^Tz) = \phi'(v_i^*Tz)$ since $\phi''(s)$ are zeros except for $\{p_j\}_{j=1,2,\ldots,e}$. So we consider the probability that $v_i^Tz, v_i^*Tz$ are separated by any exception point. We use $\xi_j$ to denote the event that $v_i^Tz, v_i^*Tz$ are separated by an exceptional point $p_j$. By union bound, $1 - \sum_{j=1}^e \Pr[\xi_j]$ is the probability that $v_i^Tz, v_i^*Tz$ are not separated by any exceptional point. The first term of Equation (18) can be bounded as,

$$
\mathbb{E}_{z \sim D_3} \left[ |\phi'(v_i^Tz) - \phi'(v_i^*Tz)||\phi'(v_i^Tz)|(z^Tb)^2 \right] 
= \mathbb{E}_{z \sim D_3} \left[ 1_{\{j=1, \ldots, e\}}|\phi'(v_i^Tz) + \phi'(v_i^*Tz)||\phi'(v_i^Tz)|(z^Tb)^2 \right]
\leq \left( \mathbb{E}_{z \sim D_3} \left[ 1_{\{j=1, \ldots, e\}} \right] \right)^{1/2} \left( \mathbb{E}_{z \sim D_3} \left[ (\phi'(v_i^Tz) + \phi'(v_i^*Tz))^2(\phi'(v_i^Tz)^2(z^Tb)^4) \right] \right)^{1/2}
\leq \left( \sum_{j=1}^e \Pr[\xi_j] \right)^{1/2} \left( \mathbb{E}_{z \sim D_3} \left[ (\phi'(v_i^Tz) + \phi'(v_i^*Tz))^2(\phi'(v_i^Tz)^2(z^Tb)^4) \right] \right)^{1/2}
\leq \left( \sum_{j=1}^e \Pr[\xi_j] \right)^{1/2} (||v_i^p|| + ||v_i^*||^p)||v_i||b||^2
$$

where the first step follows by if $v_i^Tz, v_i^*Tz$ are not separated by any exceptional point then $\phi'(v_i^Tz) = \phi'(v_i^*Tz)$ and the last step follows by Hölder’s inequality and Property 3.1.

It remains to upper bound $\Pr_{z \sim D_3}[\xi_j]$. First note that if $v_i^Tz, v_i^*Tz$ are separated by an exceptional point, $p_j$, then $|v_i^*Tz - p_j| \leq |v_i^Tz - v_i^*Tz| \leq ||v_i - v_i^*||z||$. Therefore,

$$
Pr_{z \sim D_3}[\xi_j] \leq Pr_{z \sim D_3} \left[ \frac{|v_i^Tz - p_j|}{||z||} \leq ||v_i - v_i^*|| \right].
$$

Note that $(\frac{v_i^Tz}{||z||} + 1)/2$ follows Beta(1,1) distribution which is uniform distribution on $[0,1]$.

$$
Pr_{z \sim D_3} \left[ \frac{|v_i^Tz - p_j|}{||z||} \leq \frac{||v_i - v_i^*||}{||z||} \right] \leq Pr_{z \sim D_3} \left[ \frac{||v_i^*Tz||}{||z||} \leq \frac{||v_i - v_i^*||}{||z||} \right] \leq \frac{||v_i - v_i^*||}{||v_i^*||} \leq \frac{||W - W^*||}{\sigma_k(W*)},
$$

where the first step is because we can view $\frac{v_i^Tz}{||z||}$ and $\frac{p_j}{||z||}$ as two independent random variables: the former is about the direction of $z$ and the later is related to the magnitude of $z$. Thus, we have

$$
\mathbb{E}_{z \sim D_3} \left[ |\phi'(v_i^Tz) - \phi'(v_i^*Tz)||\phi'(v_i^Tz)|(z^Tb)^2 \right] \lesssim (e||W - W^*||/\sigma_k(W*))^{1/2}\sigma_1^{2p}(W^*)||b||^2.
$$

Similarly we have

$$
\mathbb{E}_{y \sim D_{d-3}} \left[ |\phi'(u_i^Ty) - \phi'(u_i^*Ty)||\phi'(u_i^Ty)|(y^Tc)^2 \right] \lesssim (e||W - W^*||/\sigma_k(W*))^{1/2}\sigma_1^{2p}(W^*)||c||^2.
$$

Finally combining Eq. (17), Eq. (19) and Eq. (20), we have

$$
||H - \nabla^2 f_D(W^*)|| \lesssim kv_2^2(e||W - W^*||/\sigma_k(W*))^{1/2}\sigma_1^{2p}(W^*),
$$

which completes the proof. 

\square
D.5 Positive Definiteness for a Small Region

Here we introduce a lemma which shows that the Hessian of any $W$, which may be dependent on the samples but is very close to an anchor point, is close to the Hessian of this anchor point.

**Lemma D.16.** Let $S$ denote a set of samples from Distribution $\mathcal{D}$ defined in Eq. (1). Let $W^a \in \mathbb{R}^{d \times k}$ be a point (respect to function $\hat{f}_S(W)$), which is independent of the samples $S$, satisfying $\|W^a - W^*\| \leq \sigma_k/2$. Assume $\phi$ satisfies Property 3.1, 3.2 and 3.3(a). Then for any $t \geq 1$, if

$$|S| \geq d \text{ poly}(\log d, t),$$

with probability at least $1 - d^{-t}$, for any $W$ (which is not necessarily to be independent of samples) satisfying $\|W^a - W\| \leq \sigma_k/4$, we have

$$\|\nabla^2 \hat{f}_S(W) - \nabla^2 \hat{f}_S(W^a)\| \leq kv^2_{\text{max}} \sigma_4^{p} (\|W^a - W^*\| + \|W - W^a\|^{(p+1)/2}).$$

**Proof.** Let $\Delta = \nabla^2 \hat{f}_S(W) - \nabla^2 \hat{f}_S(W^a) \in \mathbb{R}^{d \times k \times k}$, then $\Delta$ can be thought of as $k^2$ blocks, and each block has size $d \times d$. The off-diagonal blocks are,

$$\Delta_{i,j} = v_i^* v_j^* \frac{1}{|S|} \sum_{x \in S} \left( \phi'(w_i^\top x)\phi'(w_j^\top x) - \phi'(w_i^{a^\top} x)\phi'(w_j^{a^\top} x) \right) xx^\top$$

For diagonal blocks,

$$\Delta_{i,i} = \frac{1}{|S|} \sum_{x \in S} \left( \sum_{q=1}^{k} v_q^* \phi(w_q^\top x) - y \right) v_i^* \phi''(w_i^\top x) xx^\top + v_i^* \phi'(w_i^\top x) xx^\top$$

$$- \frac{1}{|S|} \sum_{x \in S} \left( \sum_{q=1}^{k} v_q^* \phi(w_q^\top x) - y \right) v_i^* \phi''(w_i^{a^\top} x) xx^\top + v_i^* \phi'(w_i^{a^\top} x) xx^\top$$

We further decompose $\Delta_{i,i}$ into $\Delta_{i,i} = \Delta_{i,i}^{(1)} + \Delta_{i,i}^{(2)}$, where

$$\Delta_{i,i}^{(1)} = v_i^* \frac{1}{|S|} \sum_{x \in S} \left( \sum_{q=1}^{k} v_q^* \phi(w_q^\top x) - y \right) \phi''(w_i^\top x) - \left( \sum_{q=1}^{k} v_q^* \phi(w_q^{a^\top} x) - y \right) \phi''(w_i^{a^\top} x) xx^\top,$$

and

$$\Delta_{i,i}^{(2)} = v_i^* \frac{1}{|S|} \sum_{(x,y) \in S} \phi'(w_i^\top x) xx^\top - \phi'(w_i^{a^\top} x) xx^\top.$$ (21)
We can further decompose $\Delta_{i,i}^{(1)}$ into $\Delta_{i,i}^{(1,1)}$ and $\Delta_{i,i}^{(1,2)}$, 

\[
\Delta_{i,i}^{(1)} = v_i^* \frac{1}{|S|} \sum_{x \in S} \left( \sum_{q=1}^{k} v_q^* \phi(w_q^T x) - y \right) \phi''(w_i^T x) - \left( \sum_{q=1}^{k} v_q^* \phi(w_q^T x) - y \right) \phi''(w_i^T x) xx^T
\]

\[
= v_i^* \frac{1}{|S|} \sum_{x \in S} \left( \sum_{q=1}^{k} v_q^* \phi(w_q^T x) - \sum_{q=1}^{k} v_q^* \phi(w_q^T x) \right) \phi''(w_i^T x) xx^T
\]

\[
+ v_i^* \frac{1}{|S|} \sum_{x \in S} \sum_{q=1}^{k} v_q^* (\phi(w_q^T x) - \phi(w_q^T x)) \phi''(w_i^T x) xx^T
\]

\[
+ v_i^* \frac{1}{|S|} \sum_{x \in S} \sum_{q=1}^{k} v_q^* (\phi(w_q^T x) - \phi(w_q^T x)) (\phi''(w_i^T x) - \phi''(w_i^T x)) xx^T
\]

\[
= \Delta_{i,i}^{(1,1)} + \Delta_{i,i}^{(1,2)}.
\]

Combining Claim D.17 and Claim D.18, we have if

\[
n \geq d \text{ poly}(\log d, t)
\]

with probability at least $1 - 1/d^U$,

\[
\|\Delta_{i,i}^{(1)}\| \lesssim k v_{\max}^2 \sigma_1^p \|W - W^*\| + \|W^* - W\| d^{(p+1)/2}.
\]

(22)

Therefore, combining Eq. (22), Claim D.19 and Claim D.20, we complete the proof. \qed

**Claim D.17.** For each $i \in [k]$, if $n \geq d \text{ poly}(\log d, t)$, then

\[
\|\Delta_{i,i}^{(1,1)}\| \lesssim k v_{\max}^2 \sigma_1^p \|W - W^*\| d^{(p+1)/2}
\]

**Proof.** Define function $h_1(x) = \|x\|^{p+1}$ and $h_2(x) = |w_q^T x|^p |(w_q^*)^T w_q^T x|$. Note that $h_1$ and $h_2$ don’t contain $W$ which may depend on the samples. Therefore, we can use the modified matrix Bernstein inequality (Corollary B.8) to bound $\Delta_{i,i}^{(1)}$.

(I) Bounding $|h_1(x)|$.

By Fact B.2, we have $h_1(x) \lesssim (td \log n)^{(p+1)/2}$ with probability at least $1 - 1/(nd^U)$.

(II) Bounding $\|\mathbb{E}_{x \sim D_d} \|x\|^{p+1} xx^T\|$.
Let \( g(x) = (2\pi)^{-d/2}e^{-\|x\|^2/2} \). Note that \( xg(x)dx = -dg(x) \).

\[
\begin{align*}
\mathbb{E}_{x \sim D_d}[\|x\|^{p+1}xx^\top] &= \int \|x\|^{p+1}g(x)xx^\top dx \\
&= -\int \|x\|^{p+1}d(g(x))x^\top \\
&= -\int \|x\|^{p+1}d(g(x)x^\top) + \int \|x\|^{p+1}g(x)I_d dx \\
&= \int d(\|x\|^{p+1})g(x)x^\top + \int \|x\|^{p+1}g(x)I_d dx \\
&= \int (p+1)\|x\|^{p-1}g(x)xx^\top dx + \int \|x\|^{p+1}g(x)I_d dx \\
&\geq \int \|x\|^{p+1}g(x)I_d dx \\
&= \mathbb{E}_{x \sim D_d}[\|x\|^{p+1}]I_d.
\end{align*}
\]

Since \( \|x\|^2 \) follows \( \chi^2 \) distribution with degree \( d \), \( \mathbb{E}_{x \sim D_d}[\|x\|^q] = 2^{q/2}(d+q)/\Gamma((d+q)/2) \) for any \( q \geq 0 \). So, \( d^{q/2} \leq \mathbb{E}_{x \sim D_d}[\|x\|^q] \leq d^{q/2} \). Hence, \( \| \mathbb{E}_{x \sim D_d}[h_1(x)xx^\top] \| \geq d^{(p+1)/2} \). Also

\[
\left\| \mathbb{E}_{x \sim D_d}[h_1(x)xx^\top] \right\| \leq \max_{\|a\|=1} \mathbb{E}_{x \sim D_d}[h_1(x)(x^\top a)^2] \\
\leq \max_{\|a\|=1} \left( \mathbb{E}_{x \sim D_d}[h_1^2(x)] \right)^{1/2} \left( \mathbb{E}_{x \sim D_d}[(x^\top a)^4] \right)^{1/2} \\
\leq d^{(p+1)/2}.
\]

(III) Bounding \( \mathbb{E}_{x \sim D_d}[h_1^4(x)] \)\(^{1/4} \).

\[
\left( \mathbb{E}_{x \sim D_d}[h_1^4(x)] \right)^{1/4} \leq d^{(p+1)/2}.
\]

Define function \( B(x) = h(x)xx^\top \in \mathbb{R}^{d \times d}, \forall i \in [n] \). Let \( \mathcal{B} = \mathbb{E}_{x \sim D_d}[h(x)xx^\top] \). Therefore by applying Corollary B.8, we obtain for any \( 0 < \epsilon < 1 \), if

\[
n \geq \epsilon^{-2}d \text{poly}(\log d, t)
\]

with probability at least \( 1 - 1/d^t \),

\[
\left\| \frac{1}{|S|} \sum_{x \in S} \|x\|^{p+1}xx^\top - \mathbb{E}_{x \sim D_d}[\|x\|^{p+1}xx^\top] \right\| \leq \delta d^{(p+1)/2}.
\]

Therefore we have with probability at least \( 1 - 1/d^t \),

\[
\left\| \frac{1}{|S|} \sum_{x \in S} \|x\|^{p+1}xx^\top \right\| \leq d^{(p+1)/2}.
\] \( \square \)
Claim D.18. For each $i \in [k]$, if $n \geq d \poly(\log d, t)$, then
\[
\|\Delta_{i,i}^{(1,2)}\| \lesssim k \sigma_i^p \|W^a - W^*\|,
\]
holds with probability at least $1 - 1/d^4t$.

Proof. Recall the definition of $\Delta_{i,i}^{(1,2)}$,\[
\Delta_{i,i}^{(1,2)} = v_i^* \frac{1}{|S|} \sum_{x \in S} \sum_{q=1}^k v_q^*(\phi(w_q^*tx) - \phi(w_q^{q^*}tx))(\phi''(w_t^x) - \phi''(w_t^{q^*}tx))xx^T.
\]
In order to upper bound the $\|\Delta_{i,i}^{(1,2)}\|$, it suffices to upper bound the spectral norm of this quantity,
\[
\frac{1}{|S|} \sum_{x \in S} (\phi(w_q^*tx) - \phi(w_q^{q^*}tx))(\phi''(w_t^x) - \phi''(w_t^{q^*}tx))xx^T.
\]
By Property 3.1, we have
\[
|\phi(w_q^*tx) - \phi(w_q^{q^*}tx)| \lesssim L_1(|w_q^*tx|^p + |w_q^{q^*}tx|)(w_q^* - w_q^{q^*})^T x.|\]
By Property 3.3, we have $|\phi''(w_t^x) - \phi''(w_t^{q^*}tx)| \leq 2L_2$.
For the second part $|w_q^*tx|^p(w_q^* - w_q^{q^*})^T x|xx^T$, according to Eq. (15), we have with probability $1 - d^{-t}$, if $n \geq d \poly(\log d, t)$,
\[
\left\| \mathbb{E}_{x \sim D_d} \left[ |w_q^*tx|^p(w_q^* - w_q^{q^*})^T x|xx^T \right] - \frac{1}{|S|} \sum_{x \in S} |w_q^*tx|^p(w_q^* - w_q^{q^*})^T x|xx^T \right\| \lesssim \delta^p \|w_q^*\|^p \|w_q^* - w_q^{q^*}\|.
\]
Also, note that
\[
\left\| \mathbb{E}_{x \sim D_d} \left[ |w_q^*tx|^p(w_q^* - w_q^{q^*})^T x|xx^T \right] \right\| \lesssim \|w_q^*\|^p \|w_q^* - w_q^{q^*}\|.
\]
Thus, we obtain
\[
\left\| \frac{1}{|S|} \sum_{x \in S} |w_q^*tx|^p(w_q^* - w_q^{q^*})^T x|xx^T \right\| \lesssim \|W^a - W^*\| \sigma_i^p. \tag{24}
\]

Claim D.19. For each $i \in [k]$, if $n \geq d \poly(\log d, t)$, then
\[
\|\Delta_{i,i}^{(2)}\| \lesssim k 1\max_i \sigma_i^p \|W - W^a\| d^{(p+1)/2}
\]
holds with probability $1 - 1/d^4t$.

Proof. We have
\[
\|\Delta_{i,i}^{(2)}\| \leq v_i^2 \left\| \frac{1}{|S|} \sum_{x \in S} \left( (\phi'(w_i^*tx) - \phi'(w_i^{q^*}tx)) \cdot (\phi'(w_i^*tx) + \phi'(w_i^{q^*}tx)) \right) xx^T \right\|
\leq v_i^2 \left\| \frac{1}{|S|} \sum_{x \in S} \left( L_2(|w_i^*tx|^p + |w_i^{q^*}tx|^p) \right) xx^T \right\|
\leq v_i^2 \|W - W^a\| \left\| \frac{1}{|S|} \sum_{x \in S} \left( L_2(|w_i^*tx|^p + |w_i^{q^*}tx|^p) \right) xx^T \right\|.
\]
Applying Corollary B.8 finishes the proof.
Claim D.20. For each $i \in [k], j \in [k], i \neq j$, if $n \geq d \text{poly}(\log d, t)$, then

$$\|\Delta_{i,t}\| \lesssim v_{\max}^2 \|W - W\|$$

holds with probability $1 - d^4t$.

Proof. Recall the definition of $\Delta_{i,t}$,

$$\Delta_{i,t} = v_i^*v_i^* \frac{1}{|S|} \sum_{x \in S} \left( \phi'(w_i^Tx)\phi'(w_i^Tx) - \phi'(w_i^Tx)\phi'(w_i^Tx) 
+ \phi'(w_i^Tx)\phi'(w_i^Tx) - \phi'(w_i^Tx)\phi'(w_i^Tx) \right) xx^T$$

$$= v_i^*v_i^* \frac{1}{|S|} \sum_{x \in S} \left( \phi'(w_i^Tx)\phi'(w_i^Tx) - \phi'(w_i^Tx)\phi'(w_i^Tx) 
+ v_i^*v_i^* \frac{1}{|S|} \sum_{x \in S} \left( \phi'(w_i^Tx)\phi'(w_i^Tx) - \phi'(w_i^Tx)\phi'(w_i^Tx) \right) xx^T$$

$$\lesssim |v_i^*v_i^*| \frac{1}{|S|} \sum_{x \in S} \left( L_1\|w_i\|^p L_2\|w_i - w_i^q\||x||x||p+1 + L_2\|w_i - w_i^q\||x||L_1\|w_i^q\| x^p \right) xx^T$$

Applying Corollary B.8 completes the proof. \hfill \square

E Tensor Methods

E.1 Tensor Initialization Algorithm

We describe the details of each procedure in Algorithm 1 in this section.

a) Compute the subspace estimation from $\hat{P}_2$ (Algorithm 3). Note that the eigenvalues of $P_2$ and $\hat{P}_2$ are not necessarily nonnegative. However, only $k$ of the eigenvalues will have large magnitude. So we can first compute the top-$k$ eigenvectors/eigenvalues of both $C \cdot I + \hat{P}_2$ and $C \cdot I - \hat{P}_2$, where $C$ is large enough such that $C \geq 2\|P_2\|$. Then from the $2k$ eigen-pairs, we pick the top-$k$ eigenvectors with the largest eigenvalues in magnitude, which is executed in TopK in Algorithm 3. For the outputs of TopK, $k_1, k_2$ are the numbers of picked largest eigenvalues from $C \cdot I + \hat{P}_2$ and $C \cdot I - \hat{P}_2$ respectively and $\pi_1(i)$ returns the original index of $i$-th largest eigenvalue from $C \cdot I + \hat{P}_2$ and similarly $\pi_2$ is for $C \cdot I - \hat{P}_2$. Finally orthogonalizing the picked eigenvectors leads to an estimation of the subspace spanned by $\{w_1^*, w_2^*, \cdots, w_k^*\}$. Also note that forming $\hat{P}_2$ takes $O(n \cdot d^2)$ time and each step of the power method doing a multiplication between a $d \times d$ matrix and a $d \times k$ matrix takes $k \cdot d^3$ time by a naive implementation. Here we reduce this complexity from $O((k + n)d^2)$ to $O(knd)$. The idea is to compute each step of the power method without explicitly forming $\hat{P}_2$. We take $P_2 = M_2$ as an example; other cases are similar. In Algorithm 3, let the step $\hat{P}_2V$ be calculated as $\hat{P}_2V = \frac{1}{|S|} \sum_{(x,y) \in S} y(x(x^TV) - V)$. Now each iteration only needs $O(knd)$ time. Furthermore, the number of iterations required will be a small number, since the power method has a linear convergence rate and as an initialization method we don’t need a very accurate solution. The detailed algorithm is shown in Algorithm 3. The approximation error bound of $\hat{P}_2$ to $P_2$ is provided in Lemma E.5. Lemma E.6 provides the theoretical bound for Algorithm 3.

b) Form and decompose the 3rd-moment $\hat{R}_3$ (Algorithm 1 in [KCL15]). We apply the non-orthogonal tensor factorization algorithm, Algorithm 1 in [KCL15], to decompose $\hat{R}_3$. According to Theorem 3 in [KCL15], when $\hat{R}_3$ is close enough to $R_3$, the output of the algorithm, $\hat{u}_i$ will
close enough to $s_i V^T \overline{w}_i$, where $s_i$ is an unknown sign. Lemma E.10 provides the error bound for $\| \hat{R}_3 - R_3 \|$.  

**c) Recover the magnitude of $w_i^*$ and the signs $s_i,v_i^*$ (Algorithm 4).** For Algorithm 4, we only consider homogeneous functions. Hence we can assume $v_i^* \in \{-1,1\}$ and there exist some universal constants $c_j$ such that $m_{j,i} = c_j \| w_i^* \|^{p+1}$ for $j = 1,2,3,4$, where $p + 1$ is the degree of homogeneity. Note that different activations have different situations even under Assumption 5.3. In particular, if $M_4 = M_2 = 0$, $\phi(\cdot)$ is an odd function and we only need to know $s_i v_i^*$. If $M_3 = M_1 = 0$, $\phi(\cdot)$ is an even function, so we don’t care about what $s_i$ is.

Let’s describe the details for Algorithm 4. First define two quantities $Q_1$ and $Q_2$,

\[
Q_1 = M_1 \left( I, \alpha, \ldots, \alpha \right) = \sum_{i=1}^{k} v_i^* c_i \| w_i^* \|^{p+1} (\alpha^T \overline{w}_i)^{l_1-1} \overline{w}_i^*,
\]

\[
Q_2 = M_2 \left( V, V, \alpha, \ldots, \alpha \right) = \sum_{i=1}^{k} v_i^* c_{12} \| w_i^* \|^{p+1} (\alpha^T \overline{w}_i)^{l_2-2} (V^T \overline{w}_i) (V^T \overline{w}_i)^T,
\]

where $l_1 \geq 1$ such that $M_1 \neq 0$ and $l_2 \geq 2$ such that $M_2 \neq 0$. There are possibly multiple choices for $l_1$ and $l_2$. We will discuss later on how to choose them. Now we solve two linear systems.

\[
z^* = \underset{z \in \mathbb{R}^k}{\text{argmin}} \left\| \sum_{i=1}^{k} z_i s_i \overline{w}_i - Q_1 \right\|, \text{ and } r^* = \underset{r \in \mathbb{R}^k}{\text{argmin}} \left\| \sum_{i=1}^{k} r_i V^T \overline{w}_i (V^T \overline{w}_i)^T - Q_2 \right\|.
\]

The solutions of the above linear systems are

\[
z_i^* = v_i^* s_i^l \overline{c}_i \| w_i^* \|^{p+1} (\alpha^T \overline{w}_i)^{l_1-1}, \text{ and } r_i^* = v_i^* s_i^l \overline{c}_{12} \| w_i^* \|^{p+1} (\alpha^T \overline{w}_i)^{l_2-2}.
\]

We can approximate $s_i \overline{w}_i$ by $V \hat{u}_i$ and approximate $Q_1$ and $Q_2$ by their empirical versions $\hat{Q}_1$ and $\hat{Q}_2$ respectively. Hence, in practice, we solve

\[
\hat{z} = \underset{z \in \mathbb{R}^k}{\text{argmin}} \left\| \sum_{i=1}^{k} z_i V \hat{u}_i - \hat{Q}_1 \right\|, \text{ and } \hat{r} = \underset{r \in \mathbb{R}^k}{\text{argmin}} \left\| \sum_{i=1}^{k} r_i \hat{u}_i \hat{u}_i^T - \hat{Q}_2 \right\|.
\]

So we have the following approximations,

\[
\hat{z}_i \approx v_i^* s_i^l \overline{c}_i \| w_i^* \|^{p+1} (\alpha^T V \hat{u}_i)^{l_1-1}, \text{ and } \hat{r}_i \approx v_i^* s_i^l \overline{c}_{12} \| w_i^* \|^{p+1} (\alpha^T V \hat{u}_i)^{l_2-2}, \forall i \in [k].
\]

In Lemma E.13 and Lemma E.14, we provide robustness of the above two linear systems, i.e., the solution errors, $\| \hat{z} - z^* \|$ and $\| \hat{r} - r^* \|$, are bounded under small perturbations of $\overline{w}_i, Q_1$ and $Q_2$. Recall that the final goal is to approximate $\| w_i^* \|$ and the signs $v_i^*, s_i$. Now we can approximate $\| w_i^* \|$ by $(|\hat{z}_i|/(c_1 (\alpha^T V \hat{u}_i)^{l_1-1}))^{1/(p+1)}$. To recover $v_i^*, s_i$, we need to note that if $l_1$ and $l_2$ are both odd or both even, we can’t recover both $v_i^*$ and $s_i$. So we consider the following situations,

1. If $M_1 = M_3 = 0$, we choose $l_1 = l_2 = \min \{ j \in \{2,4\} | M_j \neq 0 \}$. Return $v_i^{(0)} = \text{sign}(\hat{r}_i \overline{c}_{12})$ and $s_i^{(0)}$ being $-1$ or $1$.

2. If $M_2 = M_4 = 0$, we choose $l_1 = \min \{ j \in \{1,3\} | M_j \neq 0 \}$, $l_2 = 3$. Return $v_i^{(0)}$ being $-1$ or $1$ and $s_i^{(0)} = \text{sign}(v_i^{(0)} \hat{z}_i \overline{c}_i)$. 

56
Theorem 5.6. Let the activation function be homogeneous satisfying Assumption 5.3. For any $0 < \epsilon < 1$ and $t \geq 1$, if $|S| \geq \epsilon^{-2} \cdot d \cdot \text{poly}(t, k, \kappa, \log d)$, then there exists an algorithm (Algorithm 1) that takes $|S|k \cdot O(d)$ time and outputs a matrix $W(0) \in \mathbb{R}^{d \times k}$ and a vector $v(0) \in \mathbb{R}^k$ such that, with
The second is the estimation of the magnitude of $V$, where the first step follows by triangle inequality, the second step follows by the normalized $w$. The success of Algorithm 1 depends on two approximations. The first is the estimation of $\Pr$.

Proof. The success of Algorithm 1 depends on two approximations. The first is the estimation of the normalized $w_i^*$ up to some unknown sign flip, i.e., the error $\|\bar{w}_i^* - s_i V \hat{u}_i\|$ for some $s_i \in \{-1, 1\}$. The second is the estimation of the magnitude of $w_i^*$ and the signs $v_i^*, s_i$ which is conducted in Algorithm 4.

For the first one,\
\[
\|\bar{w}_i^* - s_i V \hat{u}_i\| \leq \|VV^T \bar{w}_i^* - \bar{w}_i^*\| + \|VV^T \bar{w}_i^* - V s_i \hat{u}_i\| = \|VV^T \bar{w}_i^* - \bar{w}_i^*\| + \|V^T \bar{w}_i^* - s_i \hat{u}_i\|, 
\]
where the first step follows by triangle inequality, the second step follows by $V^TV = I$.

We can upper bound $\|VV^T \bar{w}_i^* - \bar{w}_i^*\|$,\
\[
\|VV^T \bar{w}_i^* - \bar{w}_i^*\| \leq (\|\bar{P}_2 - P_2\|/\sigma_k(P_2) + \epsilon) \\
\leq (\text{poly}(k, \kappa)\|\bar{P}_2 - P_2\| + \epsilon) \\
\leq \text{poly}(k, \kappa)\epsilon, 
\]
where the first step follows by Lemma E.6, the second step follows by $\sigma_k(P_2) \geq 1/\text{poly}(k, \kappa)$, and the last step follows by $\|\bar{P}_2 - P_2\| \leq \epsilon \text{poly}(k, \kappa)$ if the number of samples is proportional to $O(d/\epsilon^2)$ as shown in Lemma E.5.

We can upper bound $\|V^T \bar{w}_i^* - s_i \hat{u}_i\|$,\
\[
\|V^T \bar{w}_i^* - s_i \hat{u}_i\| \leq \text{poly}(k, \kappa)\|\bar{R}_3 - R_3\| \leq \epsilon \text{poly}(k, \kappa), 
\]
where the first step follows by Theorem 3 in [KCL15], and the last step follows by \( \| \hat{R}_3 - R_3 \| \leq \epsilon \text{poly}(k, \kappa) \) if the number of samples is proportional to \( \bar{O}(k^2/\epsilon^2) \) as shown in Lemma E.10.

Combining Eq. (29), (30) and (31) together,

\[
\| \hat{w}_i^* - s_i V \hat{u}_i \| \leq \epsilon \text{poly}(k, \kappa).
\]

For the second one, we can bound the error of the estimation of moments, \( Q_1 \) and \( Q_2 \), using number of samples proportional to \( \bar{O}(d) \) by Lemma E.12 and Lemma E.5 respectively. The error of the solutions of the linear systems Eq.(28) can be bounded by \( \| Q_1 - \hat{Q}_1 \|, \| Q_2 - \hat{Q}_2 \|, \| \hat{u}_i - V^T \hat{w}_i \| \) and \( \|(I - V V^T)\hat{w}_i\| \) according to Lemma E.13 and Lemma E.14. Then we can bound the error of the output of Algorithm 4. Furthermore, since \( v_i \)'s are discrete values, they can be exactly recovered. All the sample complexities mentioned in the above lemmata are linear in dimension and polynomial in other factors to achieve a constant error. So accumulating all these errors we complete our proof.

\[\square\]

**Remark E.1.** The proofs of these lemmata for Theorem 1 can be found in the following sections. Note that these lemmata also hold for any activations satisfying Property 3.1 and Assumption 5.3. However, since we are unclear how to implement the last step of Algorithm 1 (Algorithm 4) for general non-homogeneous activations, we restrict our theorem to homogeneous activations only.

### E.3 Error Bound for the Subspace Spanned by the Weight Matrix

#### E.3.1 Error Bound for the Second-order Moment in Different Cases

**Lemma E.2.** Let \( M_2 \) be defined as in Definition 5.1. Let \( \hat{M}_2 \) be the empirical version of \( M_2 \), i.e.,

\[
\hat{M}_2 = \frac{1}{|S|} \sum_{(x,y) \in S} y \cdot (x \otimes x - I_d),
\]

where \( S \) denote a set of samples from Distribution \( D \) defined in Eq. (1). Assume \( M_2 \neq 0 \), i.e., \( m_i(2) \neq 0 \) for any \( i \). Then for any \( 0 < \epsilon < 1, t \geq 1 \), if

\[
|S| \geq \max_{i \in [k]} (\|w_i^*\|^{p+1}/|m_{2,i}| + 1) \cdot \epsilon^{-2} d \text{poly}(\log d, t)
\]

with probability at least \( 1 - d^{-t} \),

\[
\|M_2 - \hat{M}_2\| \leq \epsilon \sum_{i=1}^{k} |v_i^* m_{2,j}|.
\]

**Proof.** Recall that, for each sample \( (x, y) \), \( y = \sum_{i=1}^{k} v_i^* \phi(w_i^* x) \). We consider each component \( i \in [k] \). Define function \( B_i(x) : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d} \) such that

\[
B_i(x) = \phi(w_i^* x) \cdot (x \otimes x - I_d).
\]

Define \( g(z) = \phi(z) - \phi(0) \), then \( |g(z)| = | \int_0^z \phi'(s) ds | \leq L_1/(p+1)|z|^{p+1} \), which follows Property 3.1.

(I) Bounding \( \|B_i(x)\| \).

\[
\|B_i(x)\| \lesssim \left( \frac{L_1}{p+1} |w_i^* x|^{p+1} + |\phi(0)| \right)(\|x_j\|^2 + 1)
\]

\[
\lesssim \left( \frac{L_1}{p+1} |w_i^*|^{p+1} + |\phi(0)| \right)d \text{poly}(\log d, t)
\]
where the last step follows by Fact B.1 and Fact B.2.

(II) Bounding \( \| E_{x \sim D_d}[B_i(x)] \| \).

Note that \( E_{x \sim D_d}[B_i(x)] = m_{2,i} \bar{w}_i^* \bar{w}_i^{*\top} \). Therefore, \( \| E_{x \sim D_d}[B_i(x)] \| = |m_{2,i}|. \)

(III) Bounding \( \max \{ E_{x \sim D_d}[B_i(x) B_i(x)^\top], E_{x \sim D_d}[B_i(x)^\top B_i(x)] \} \).

Note that \( B_i(x) \) is a symmetric matrix, thus it suffices to consider one of them.

\[
\left\| \mathbb{E}_{x \sim D_d}[B_i^2(x)] \right\| \leq \left( \mathbb{E}_{x \sim D_d}[|\phi(w_i^{*\top} x)|^4] \right)^{1/2} \left( \mathbb{E}_{x \sim D_d}[||x||^4] \right)^{1/2} \leq \frac{L_1}{p + 1} \| w_i^* \|^{p + 1} + |\phi(0)|^2 d.
\]

(IV) Bounding \( \max_{|a| = |b| = 1} \{ E_{x \sim D_d}[(a^\top B_i(x)b)^2] \} \).

Note that \( B_i(x) \) is a symmetric matrix, thus it suffices to consider the case where \( a = b \).

\[
\max_{|a| = 1} \left( \mathbb{E}_{x \sim D_d}[(a^\top B_i(x)a)^2] \right)^{1/2} \leq \left( \mathbb{E}_{x \sim D_d}[|\phi^4(w_i^{*\top} x)|] \right)^{1/4} \leq \frac{L_1}{p + 1} \| w_i^* \|^{p + 1} + |\phi(0)|.
\]

Define \( L = \| w_i^* \|^{p + 1} + |\phi(0)|. \) Then we have for any \( 0 < \epsilon < 1, \)
\[
n > \frac{L^2 d + |m_{2,i}|^2 + L |m_{2,i}| d \cdot \text{poly}(\log d, t) \epsilon}{\epsilon^2 |m_{2,i}|^2}
\]
with probability at least \( 1 - 1/d^t \),
\[
\left\| \mathbb{E}_{x \sim D_d}[B_i(x)] - \frac{1}{|S|} \sum_{x \in S} B_i(x) \right\| \leq \epsilon |m_{2,i}|.
\]

\( \square \)

**Lemma E.3.** Let \( M_3 \) be defined as in Definition 5.1. Let \( \widehat{M}_3 \) be the empirical version of \( M_3 \), i.e.,
\[
\widehat{M}_3 = \frac{1}{|S|} \sum_{(x,y) \in S} y \cdot (x^{\otimes 3} - x \otimes I),
\]
where \( S \) denote a set of samples (each sample is i.i.d. sampled from Distribution \( D \) defined in Eq. (1)). Assume \( M_3 \neq 0 \), i.e., \( m_{3,i} \neq 0 \) for any \( i \). Let \( \alpha \) be a fixed unit vector. Then for any \( 0 < \epsilon < 1, t \geq 1, \)
\[
|S| \geq \max_{i \in [k]} (|w_i^*|^{p+1}/|m_{3,i}|(\bar{w}_i^{*\top} \alpha)^2 + 1) \cdot \epsilon^{-2} d \cdot \text{poly}(\log d, t)
\]
with probability at least \( 1 - 1/d^t \),
\[
\| M_3(I, I, \alpha) - \widehat{M}_3(I, I, \alpha) \| \leq \epsilon \sum_{i=1}^{k} |v_i^* m_{3,i}(\bar{w}_i^{*\top} \alpha)|.
\]

**Proof.** Since \( y = \sum_{i=1}^{k} v_i^* \phi(w_i^{*\top} x) \). We consider each component \( i \in [k] \).

Define function \( B_i(x) : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d} \) such that
\[
B_i(x) = [\phi(w_i^{*\top} x) \cdot (x^{\otimes 3} - x \otimes I)](I, I, \alpha) = \phi(w_i^{*\top} x) \cdot ((x^\top \alpha)x^{\otimes 2} - \alpha^\top x I - \alpha x^\top - x \alpha^\top).
\]

Define \( g(z) = \phi(z) - \phi(0) \), then \( |g(z)| = | \int_0^z \phi'(s) ds | \leq \frac{L_1}{p + 1} |z|^{p + 1} \leq |z|^{p + 1}, \) which follows Property 3.1. In order to apply Lemma B.7, we need to bound the following four quantities,
(I) Bounding $\|B_i(x)\|$. 
\[
\|B_i(x)\| = \|\phi(w_i^\top x) \cdot ((x^\top \alpha)x^\otimes 2 - \alpha^\top xI_d - \alpha x^\top - x\alpha^\top)\| 
\leq |\phi(w_i^\top x)| \cdot \|(x^\top \alpha)x^\otimes 2 - \alpha^\top xI - \alpha x^\top - x\alpha^\top\| 
\lesssim (|w_i^\top x|^{p+1} + |\phi(0)|)\|(x^\top \alpha)x^\otimes 2 - \alpha^\top xI - \alpha x^\top - x\alpha^\top\| 
\lesssim (|w_i^\top x|^{p+1} + |\phi(0)|)(|x^\top \alpha||x|^{2} + 3|x\alpha^\top|),
\]
where the third step follows by definition of $g(z)$, and last step follows by definition of spectral norm and triangle inequality.

Using Fact B.1 and Fact B.2, we have for any constant $t \geq 1$, with probability $1 - 1/(nd^4t)$,
\[
\|B_i(x)\| \lesssim (|w_i^*|^{p+1} + |\phi(0)|)d\text{poly}(\log d, t).
\]

(II) Bounding $\|E_{x \sim D_d}[B_i(x)]\|$. 
Note that $E_{x \sim D_d}[B_i(x)] = m_{3,i}(\bar{w}_i^\top \alpha)\bar{w}_i^\top \bar{w}_i^\top$. Therefore, $\|E_{x \sim D_d}[B_i(x)]\| = |m_{3,i}(\bar{w}_i^\top \alpha)|$.

(III) Bounding $\max\{\|E_{x \sim D_d}[B_i(x)B_i(x)^\top]\|, \|E_{x \sim D_d}[B_i(x)^\top B_i(x)]\|\}$. 
Because matrix $B_i(x)$ is symmetric, it suffices to bound one of them,
\[
\bigg\| \mathbb{E}_{x \sim D_d}[B_i^2(x)] \bigg\| \lesssim \left( \mathbb{E}_{x \sim D_d}\left[ \phi(w_i^\top x)^4 \right] \right)^{1/2} \left( \mathbb{E}_{x \sim D_d}\left[ (x^\top \alpha)^4 \right] \right)^{1/2} \left( \mathbb{E}_{x \sim D_d}\left[ |x|^4 \right] \right)^{1/2} 
\lesssim (|w_i^*|^{p+1} + |\phi(0)|)^2d.
\]

(IV) Bounding $\max_{\|a\|=\|b\|=1}(\mathbb{E}_{x \sim D_d}[(a^\top B_i(x)b)^2])^{1/2}$. 
Define $L = |w_i^*|^{p+1} + |\phi(0)|$. Then we have for any $0 < \epsilon < 1$, if
\[
|S| \underset{\text{with probability at least } 1 - d^{-t}}{\geq} \frac{L^2d + |m_{3,i}(\bar{w}_i^\top \alpha)|^2 + L|m_{3,i}(\bar{w}_i^\top \alpha)|d \cdot \text{poly}(\log d, t)\epsilon \cdot t \log d}{\epsilon^2|m_{3,i}(\bar{w}_i^\top \alpha)|^2}
\]

\[ \mathbb{E}_{x \sim D_d}[B_i(x)] - \frac{1}{|S|} \sum_{x \in S} B_i(x) \leq \epsilon|m_{3,i}(\bar{w}_i^\top \alpha)|. \]

\[ \square \]

**Lemma E.4.** Let $M_4$ be defined as in Definition 5.1. Let $\hat{M}_4$ be the empirical version of $M_4$, i.e.,
\[
\hat{M}_4 = \frac{1}{|S|} \sum_{(x,y) \in S} y \cdot (x^\otimes 4 - (x \otimes x) \otimes I + I \otimes I),
\]
where $S$ denote a set of samples (where each sample is i.i.d. sampled are sampled from Distribution $D$ defined in Eq. (1)). Assume $M_4 \neq 0$, i.e., $m_{4,i} \neq 0$ for any $i$. Let $\alpha$ be a fixed unit vector. Then for any $0 < \epsilon < 1, t \geq 1$, if
\[
|S| \geq \max_{i \in [k]}(|w_i^*|^{p+1}/|m_{4,i}(\bar{w}_i^\top \alpha)^2 + 1)^2 \cdot \epsilon^{-2} \cdot d \cdot \text{poly}(\log d, t)
\]

61
with probability at least \(1 - 1/d^t\),

\[
\|M_d(I, I, \alpha, \alpha) - \widetilde{M}_d(I, I, \alpha, \alpha)\| \leq \epsilon \sum_{i=1}^{k} |v_i^* m_{4,i}|(\overrightarrow{w}_i^* \alpha)^2.
\]

**Proof.** Since \(y = \sum_{i=1}^{k} v_i^* \phi(w_i^* x)\). We consider each component \(i \in [k]\).

Define function \(B_i(x) : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}\) such that

\[
B_i(x) = [\phi(w_i^* x) \cdot ((x^\top \alpha)^2 x^2 - (\alpha^\top x)^2 I - 2(\alpha^\top x)(x^\top x + \alpha x^\top) - xx^\top + 2\alpha \alpha^\top + I)].
\]

Define \(g(z) = \phi(z) - \phi(0)\), then \(|g(z)| = |\int_0^z \phi'(s)ds| \leq L_1/(p + 1)|z|^{p+1} \lesssim |z|^{p+1}\), which follows Property 3.1.

(I) Bounding \(\|B_i(x)\|\).

\[
\|B_i(x)\| \lesssim \|\phi(w_i^* x)\| \cdot ((x^\top \alpha)^2 \|x\|^2 + 1 + \|x\|^2 + (\alpha^\top x)^2)
\]

\[
\lesssim (\|w_i^* x\|^{p+1} + |\phi(0)|) \cdot ((x^\top \alpha)^2 \|x\|^2 + 1 + \|x\|^2 + (\alpha^\top x)^2)
\]

Using Fact B.1 and Fact B.2, we have for any constant \(t \geq 1\), with probability \(1 - 1/(nd^{4t})\),

\[
\|B_i(x)\| \lesssim (\|w_i^* \|^p + |\phi(0)|) d \text{poly}(d, t).
\]

(II) Bounding \(\|E_{x \sim D_d}[B_i(x)]\|\).

Note that \(E_{x \sim D_d}[B_i(x)] = m_{4,i}(\overrightarrow{w}_i^* \alpha)^2 \overrightarrow{w}_i^* \overrightarrow{w}_i\). Therefore, \(\|E_{x \sim D_d}[B_i(x)]\| = |m_{4,i}|(\overrightarrow{w}_i^* \alpha)^2\).

(III) Bounding \(\max(\|E_{x \sim D_d}[B_i(x)]B_i(x)^\top\|, \|E_{x \sim D_d}[B_i(x)^\top B_i(x)]\|)\).

\[
\left\| E_{x \sim D_d}[B_i(x)^2] \right\| \lesssim \left( E_{x \sim D_d}[\phi(w_i^* x)^4] \right)^{1/2} \left( E_{x \sim D_d}[(x^\top \alpha)^8] \right)^{1/2} \left( E_{x \sim D_d}[\|x\|^4] \right)^{1/2}
\]

\[
\lesssim (\|w_i^* \|^p + |\phi(0)|)^2 d.
\]

(IV) Bounding \(\max_{\|a\|_2 = \|b\|_2 = 1}(E_{x \sim D_d}[(a^\top B_i(x)b)^2])^{1/2}\).

\[
\max_{\|a\|_2 = \|b\|_2 = 1} \left( E_{x \sim D_d}[(a^\top B_i(x)a)^2] \right)^{1/2} \lesssim \left( E_{x \sim D_d}[\phi^4(w_i^* x)^2] \right)^{1/4} \lesssim \|w_i^* \|^p + |\phi(0)|.
\]

Define \(L = \|w_i^* \|^p + |\phi(0)|\). Then we have for any \(0 < \epsilon < 1\), if

\[
n \gtrsim \frac{L^2 d + |m_{4,i}|^2(\overrightarrow{w}_i^* \alpha)^4 + L|m_{4,i}|(\overrightarrow{w}_i^* \alpha)^2 d \text{poly}(d, t) \epsilon}{\epsilon^2|m_{4,i}|^2(\overrightarrow{w}_i^* \alpha)^4} \cdot t \log d
\]

with probability at least \(1 - d^{-t}\),

\[
\left\| E_{x \sim D_d}[B_i(x)] - \frac{1}{|S|} \sum_{x \in S} B_i(x) \right\| \leq \epsilon |m_{4,i}|(\overrightarrow{w}_i^* \alpha)^2.
\]

\qed
E.3.2 Error Bound for the Second-order Moment

The goal of this section is to prove Lemma E.5, which shows we can approximate the second order moments up to some precision by using linear sample complexity in $d$.

**Lemma E.5** (Estimation of the second order moment). Let $P_2$ and $j_2$ be defined in Definition 5.4. Let $S$ denote a set of i.i.d. samples generated from distribution $D$ (defined in (1)). Let $\hat{P}_2$ be the empirical version of $P_2$ using dataset $S$, i.e., $\hat{P}_2 = \mathbb{E}_S[P_2]$. Assume the activation function satisfies Property 3.1 and Assumption 5.3. Then for any $0 < \epsilon < 1$ and $t \geq 1$, and $m_0 = \min_{i \in [k]} \{ |m_{j_2, i}|^2 (\overline{w}_i^\top \alpha)^{2(j_2 - 2)} \}$, if

$$|S| \geq \sigma_1^{2p+2} \cdot d \cdot \text{poly}(t, \log d) / (\epsilon^2 m_0),$$

then with probability at least $1 - d^{-\Omega(t)}$,

$$\|P_2 - \hat{P}_2\| \leq \epsilon \sum_{i=1}^k |v_i^* m_{j_2, i}(\overline{w}_i^\top \alpha)^{j_2 - 2}|.$$

**Proof.** We have shown the bound for $j_2 = 2, 3, 4$ in Lemma E.2, Lemma E.3 and Lemma E.4 respectively. To summarize, for any $0 < \epsilon < 1$ we have if

$$|S| \geq \max_{i \in [k]} \left\{ \left( \frac{\|w_i^*\|^{p+1} + |\phi(0)| + |m_{j_2, i}(\overline{w}_i^\top \alpha)^{j_2 - 2}|^2}{|m_{j_2, i}|^2 (\overline{w}_i^\top \alpha)^{2(j_2 - 2)}} \right) \cdot \epsilon^{-2} d \cdot \text{poly}(\log d, t) \right\} \cdot \text{poly}(\log d, t),$$

with probability at least $1 - d^{-t},$

$$\|P_2 - \hat{P}_2\| \leq \epsilon \sum_{i=1}^k |v_i^* m_{j_2, i}(\overline{w}_i^\top \alpha)^{j_2 - 2}|.$$

E.3.3 Subspace Estimation Using Power Method

Lemma E.6 shows a small number of power iterations can estimate the subspace of $\{w_i^*\}_{i \in [k]}$ to some precision, which provides guarantees for Algorithm 3.

**Lemma E.6** (Bound on subspace estimation). Let $P_2$ be defined as in Definition 5.4 and $\hat{P}_2$ be its empirical version. Let $U \in \mathbb{R}^{d \times k}$ be the orthogonal column span of $W^* \in \mathbb{R}^{d \times k}$. Assume $\|\hat{P}_2 - P_2\| \leq \sigma_k(P_2)/10$. Let $C$ be a large enough positive number such that $C > 2\|P_2\|$. Then after $T = O(\log(1/\epsilon))$ iterations, the output of Algorithm 3, $V \in \mathbb{R}^{d \times k}$, will satisfy

$$\|UU^\top - VV^\top\| \lesssim \|\hat{P}_2 - P_2\|/\sigma_k(P_2) + \epsilon,$$

which implies

$$\|(I - VV^\top)w_i^*\| \lesssim \left( \frac{\|\hat{P}_2 - P_2\|/\sigma_k(P_2) + \epsilon}{\sigma_i(P_2)} \right) \|w_i^*\|.$$

**Proof.** According to Weyl’s inequality, we are able to pick up the correct numbers of positive eigenvalues and negative eigenvalues in Algorithm 3 as long as $\hat{P}_2$ and $P_2$ are close enough.

Let $U = [U_1 \ U_2] \in \mathbb{R}^{d \times k}$ be the eigenspace of span$\{w_1^* \ w_2^* \ \cdots \ w_k^*\}$, where $U_1 \in \mathbb{R}^{d \times k_1}$ corresponds to positive eigenvalues of $P_2$ and $U_2 \in \mathbb{R}^{d \times k_2}$ is for negatives.
Let \( \hat{V}_1 \) be the top-\( k_1 \) eigenvectors of \( CI + \hat{P}_2 \). Let \( \hat{V}_2 \) be the top-\( k_2 \) eigenvectors of \( CI - \hat{P}_2 \). Let \( \hat{V} = [\hat{V}_1 \ \hat{V}_2] \in \mathbb{R}^{d \times k} \).

According to Lemma 9 in [HK13], we have \( \|I - U_1U_1^T\hat{V}_1\| \lesssim \|\hat{P}_2 - P_2\|/\sigma_k(P_2), \|I - U_2U_2^T\hat{V}_2\| \lesssim \|\hat{P}_2 - P_2\|/\sigma_k(P_2) \). Using Fact B.4, we have \( \|I - UU^T\| = \|UU^T - \hat{V}\hat{V}^\top\| \).

Let \( \epsilon \) be the precision we want to achieve using power method. Let \( V_1 \) be the top-\( k_1 \) eigenvectors returned after \( O(\log(1/\epsilon)) \) iterations of power methods on \( CI + \hat{P}_2 \) and \( V_2 \in \mathbb{R}^{d \times k_2} \) for \( CI - \hat{P}_2 \) similarly.

According to Theorem 7.2 in [AKZ12], we have \( \|\hat{V}_1V_1^T - V_1V_1^T\| \leq \epsilon \) and \( \|\hat{V}_2V_2^T - V_2V_2^T\| \leq \epsilon \). Let \( U_\perp \) be the complementary matrix of \( U \). Then we have,

\[
\| (I - U_1U_1^T)\hat{V}_1 \| = \max_{\|a\|=1} \| (I - U_1U_1^T)\hat{V}_1 a \|
= \max_{\|a\|=1} \| (U_1U_1^T + U_2U_2^T)\hat{V}_1 a \|
= \max_{\|a\|=1} \sqrt{\|U_1U_1^T\hat{V}_1 a \|^2 + \|U_2U_2^T\hat{V}_1 a \|^2}
\geq \max_{\|a\|=1} \|U_2U_2^T\hat{V}_1 a \|
= \|U_2U_2^T\hat{V}_1 \|,
\]

where the first step follows by definition of spectral norm, the second step follows by \( I = U_1U_1^T + U_2U_2^T + U_\perp U_\perp^T \), the third step follows by \( U_2^T U_\perp = 0 \), and last step follows by definition of spectral norm.

We can upper bound \( \| (I - UU^T)\hat{V} \| \),

\[
\| (I - UU^T)\hat{V} \|
\leq \| (I - U_1U_1^T)\hat{V}_1 \| + \| (I - U_2U_2^T)\hat{V}_2 \| + \|U_2U_2^T\hat{V}_1\| + \|U_1U_1^T\hat{V}_2\|
\leq 2\| (I - U_1U_1^T)\hat{V}_1 \| + \| (I - U_2U_2^T)\hat{V}_2 \|
\leq \| \hat{P}_2 - P_2 \|/\sigma_k(P_2),
\]

where the first step follows by triangle inequality, the second step follows by Eq. (32), and the last step follows by Lemma 9 in [HK13].

We define matrix \( R \) such that \( \hat{V}_2 R = (I - V_1V_1^\top)V_2 \) is the QR decomposition of \( (I - V_1V_1^\top)V_2 \), then we have

\[
\| (I - \hat{V}_2\hat{V}_2^\top)\hat{V}_2 \|
= \| (I - \hat{V}_2\hat{V}_2^\top)(I - V_1V_1^\top)V_2 R^{-1} \|
= \| (I - \hat{V}_2\hat{V}_2^\top)(I - \hat{V}_1\hat{V}_1^\top) + \hat{V}_1\hat{V}_1^\top - V_1V_1^\top \| V_2 R^{-1} \|
\leq \| (I - \hat{V}_2\hat{V}_2^\top)(I - \hat{V}_1\hat{V}_1^\top) \| R^{-1} \| \hat{V}_1\hat{V}_1^\top - V_1V_1^\top \|
\alpha \beta
\]

where the first step follows by \( \hat{V}_2 = (I - V_1V_1^\top)V_2 R^{-1} \), and the last step follows by triangle inequality.
Furthermore, we have,

\[ \alpha + \beta \]
\[ = \| (I - \nabla^2_2) V_2 - \nabla_1 V_1^T \| V_2 R^{-1} + (I - \nabla^2_2) \| R^{-1} \| \nabla_1 V_1^T - V_1 V_1^T \| \]
\[ \leq \| (I - \nabla^2_2) V_2 \| V_2 R^{-1} + \| \nabla_1 V_1^T V_2 \| R^{-1} + (I - \nabla^2_2) \| R^{-1} \| \nabla_1 V_1^T - V_1 V_1^T \| \]
\[ \leq \| \nabla^2_2 V_2 \| V_2 V_2^T \| R^{-1} \| + \| V_1 V_1^T V_2 \| V_2 \| R^{-1} \| + (I - \nabla^2_2) \| R^{-1} \| \nabla_1 V_1^T - V_1 V_1^T \| \]
\[ = \| \nabla^2_2 V_2 \| V_2 V_2^T \| R^{-1} \| + \| \nabla_1 V_1^T V_2 \| V_2 \| R^{-1} \| + (I - \nabla^2_2) \| R^{-1} \| \nabla_1 V_1^T - V_1 V_1^T \| \]
\[ \leq \| \nabla^2_2 V_2 \| V_2 V_2^T \| R^{-1} \| + \| (I - \nabla^2_2) ) V_2 \| V_2 \| R^{-1} \| + (I - \nabla^2_2) \| R^{-1} \| \nabla_1 V_1^T - V_1 V_1^T \| \]
\[ \leq (2 \| \nabla^2_2 V_2 \| V_2 V_2^T \| + \| \nabla_1 V_1^T - V_1 V_1^T \| ) \| R^{-1} \| \]
\[ \leq 3 \| R^{-1} \| \]
\[ \leq 6 \epsilon, \]

where the first step follows by definition of \( \alpha, \beta \), the second step follows by triangle inequality, the third step follows by \( \| AB \| \leq \| A \| \| B \| \), the fourth step follows by \( \| (I - \nabla^2_2) \| = 1 \), the fifth step follows by Eq. (32), the sixth step follows by Fact B.4, the seventh step follows by \( \| \nabla_1 V_1^T - V_1 V_1^T \| \leq \epsilon \) and \( \| \nabla^2_2 V_2 \| V_2 V_2^T \| \leq \epsilon \), and the last step follows by \( \| R^{-1} \| \leq 2 \) (Claim E.7).

Finally,

\[ \| UU^T - VV^T \| \leq \| UU^T - \nabla \nabla^T \| + \| \nabla \nabla^T - VV^T \| \]
\[ = \| (I - UU^T) \| \nabla \| + \| \nabla \nabla^T - VV^T \| \]
\[ \leq \| \tilde{P}_2 - P_2 \| / \sigma_k (P_2) + \| \nabla \nabla^T - VV^T \| \]
\[ \leq \| \tilde{P}_2 - P_2 \| / \sigma_k (P_2) + \| \nabla_1 V_1^T - V_1 V_1^T \| + \| \nabla_2 V_2^T - V_2 V_2^T \| \]
\[ \leq \| \tilde{P}_2 - P_2 \| / \sigma_k (P_2) + 2 \epsilon, \]

where the first step follows by triangle inequality, the second step follows by Fact B.4, the third step follows by Eq. (33), the fourth step follows by triangle inequality, and the last step follows by \( \| \nabla_1 V_1^T - V_1 V_1^T \| \leq \epsilon \) and \( \| \nabla_2 V_2^T - V_2 V_2^T \| \leq \epsilon \).

Therefore we finish the proof. \( \square \)

It remains to prove Claim E.7.

Claim E.7. \( \sigma_k (R) \geq 1/2 \).

Proof. First, we can rewrite \( R^T R \) in the follow way,

\[ R^T R = V_2^T (I - V_1 V_1^T) V_2 = I - V_2^T V_1 V_1^T V_2 \]
Second, we can upper bound $\|V_2^T V_1\|$ by $1/4$,

$$\|V_2^T V_1\| = \|V_2 V_2^T V_1\|
\leq \|(V_2 V_2^T - V_2 V_2^T) V_1\| + \|V_2 V_2^T V_1\|
\leq \|(V_2 V_2^T - V_2 V_2^T) V_1\| + \|V_2 (V_1 V_1^T - V_1 V_1^T)\| + \|V_2^T V_1 V_1^T\|
= \|(V_2 V_2^T - V_2 V_2^T) V_1\| + \|V_2 (V_1 V_1^T - V_1 V_1^T)\|
\leq \|V_2 V_2^T - V_2 V_2^T\| \cdot \|V_1\| + \|V_2\| \cdot \|V_1^T - V_1 V_1^T\|
\leq \epsilon + \epsilon
\leq 1/4,$$

where the first step follows by $V_2^T V_2 = I$, the second step follows by triangle inequality, the third step follows by triangle inequality, the fourth step follows by $\|V_2 V_2^T V_1\| = 0$, the fifth step follows by $\|A B\| \leq \|A\| \cdot \|B\|$, and the last step follows by $\|V_1 V_1^T - V_1 V_1^T\| \leq \epsilon$, $\|V_1\| = 1$, $\|V_2 V_2^T - V_2 V_2^T\| \leq \epsilon$, and $\|V_2\| = 1$, and the last step follows by $\epsilon < 1/8$.

Thus, we can lower bound $\sigma_k^2(R)$,

$$\sigma_k^2(R) = \lambda_{\min}(R^T R)
= \min_{\|a\|=1} a^T R^T R a
= \min_{\|a\|=1} a^T I a - \|V_2 V_1 a\|^2
= 1 - \max_{\|a\|=1} \|V_2 V_1 a\|^2
= 1 - \|V_2^T V_1\|^2
\geq 3/4,$$

which implies $\sigma_k(R) \geq 1/2$. \hfill \Box

### E.4 Error Bound for the Reduced Third-order Moment

#### E.4.1 Error Bound for the Reduced Third-order Moment in Different Cases

**Lemma E.8.** Let $M_3$ be defined as in Definition 5.1. Let $\tilde{M}_3$ be the empirical version of $M_3$, i.e.,

$$\tilde{M}_3 = \frac{1}{|S|} \sum_{(x,y) \in S} y \cdot (x \otimes^3 - x \otimes I),$$

where $S$ denote a set of samples (where each sample is i.i.d. sampled from Distribution $D$ defined in Eq. (1)). Assume $M_3 \neq 0$, i.e., $m_{3,i} \neq 0$ for any $i$. Let $V \in \mathbb{R}^{d \times k}$ be an orthogonal matrix satisfying $\|U U^T - V V^T\| \leq 1/4$, where $U$ is the orthogonal basis of $\text{span}\{w_1^*, w_2^*, \ldots, w_k^*\}$. Then for any $0 < \epsilon < 1$, $t \geq 1$, if

$$|S| \geq \max_{i \in [k]} (\|w_i^*\|^p + 1) / |m_{3,i}|^2 + 1) \cdot \epsilon^{-2} \cdot k^2 \cdot \text{poly}(\log d, t)$$

with probability at least $1 - 1/d^t$,

$$\|M_3(V, V, V) - \tilde{M}_3(V, V, V)\| \leq \epsilon \sum_{i=1}^k |v_i^* m_{3,i}|.$$
Proof. Since \( y = \sum_{i=1}^{k} v_i^* \phi(w_i^T x) \). We consider each component \( i \in [k] \). We define function \( T_i(x) : \mathbb{R}^d \rightarrow \mathbb{R}^{k \times k} \) such that,

\[
T_i(x) = \phi(w_i^T x) \cdot ((V^T x)^{\otimes 3} - (V^T x)^{\otimes 1} I). 
\]

We flatten tensor \( T_i(x) \) along the first dimension into matrix \( B_i(x) \in \mathbb{R}^{k \times k} \). Define \( g(z) = \phi(z) - \phi(0) \), then \( |g(z)| = \int_0^z \phi'(s) ds \leq L_1/(p + 1)|z|^{p + 1} \), which follows Property 3.1.

(I) Bounding \( \|B_i(x)\| \).

\[
\|B_i(x)\| \lesssim \|\phi(w_i^T x)\| \cdot (\|V^T x\|^3 + 3k\|V^T x\|) 
\lesssim (\|w_i^T x\|^{p+1} + |\phi(0)|) \cdot (\|V^T x\|^3 + 3k\|V^T x\|) 
\]

Note that \( V^T x \sim \mathcal{N}(0, I_k) \). According to Fact B.1 and Fact B.2, we have for any constant \( t \geq 1 \), with probability \( 1 - 1/(nd^t) \),

\[
\|B_i(x)\| \lesssim (\|w_i^*\|^{p+1} + |\phi(0)|)k^{3/2}\text{poly}(\log d, t) 
\]

(II) Bounding \( \|\mathbb{E}_{x \sim D_d}[B_i(x)]\| \).

Note that \( \mathbb{E}_{x \sim D_d}[B_i(x)] = m_{3,i}(V^T \bar{w}_i)\text{vec}((V^T \bar{w}_i)^3(V^T \bar{w}_i)^\top)^\top \). Therefore, \( \|\mathbb{E}_{x \sim D_d}[B_i(x)]\| = |m_{3,i}|\|V^T \bar{w}_i\|^3 \|. Since \( \|VV^T - UU^T\| \leq 1/4, \|VV^T \bar{w}_i - \bar{w}_i\| \leq 1/4 \) and \( 3/4 \leq \|V^T \bar{w}_i\| \leq 5/4 \). So \( \frac{1}{4}m_{3,i} \leq \|B_i\| \leq 2|m_{3,i}| \).

(III) Bounding \( \max_{x \sim D_d}(\|B_i(x)B_i(x)^\top\|, \mathbb{E}_{x \sim D_d}[\|B_i(x)^\top B_i(x)\|]) \).

\[
\left\| \mathbb{E}_{x \sim D_d}[B_i(x)B_i(x)^\top] \right\| \lesssim \left( \mathbb{E}_{x \sim D_d}[\|\phi(w_i^T x)^4\|] \right)^{1/2} \left( \mathbb{E}_{x \sim D_d}[\|V^T x\|^6] \right)^{1/2} 
\lesssim (\|w_i^*\|^{p+1} + |\phi(0)|)^2k^{3/2}. 
\]

(IV) Bounding \( \max_{\|a\| = \|b\| = 1}(\mathbb{E}_{x \sim D_d}[(a^T B_i(x)(b)^2)]) \).

\[
\max_{\|a\| = \|b\| = 1} \left( \mathbb{E}_{x \sim D_d}[(a^T B_i(x)(b)^2)] \right)^{1/2} 
\lesssim \left( \mathbb{E}_{x \sim D_d}[\|\phi(w_i^T x)^4\|] \right)^{1/4} \max_{\|a\| = 1} \left( \mathbb{E}_{x \sim D_d}[(a^T V^T x)^4] \right)^{1/2} \max_{\|A\| = 1} \left( \mathbb{E}_{x \sim D_d}[(A, (V^T x)(V^T x)^\top)^4] \right)^{1/4} 
\lesssim (\|w_i^*\|^{p+1} + |\phi(0)|)k 
\]

Define \( L = \|w_i^*\|^{p+1} + |\phi(0)| \). Then we have for any \( 0 < \epsilon < 1 \), if

\[
|S| \gtrsim \frac{L^2k^2 + m_{3,i}^2 + k^{3/2}\text{poly}(\log d, t)|m_{3,i}|\epsilon}{\epsilon^2|m_{3,i}|^2} t \log(k) 
\]

67
with probability at least $1 - k^{-t}$,

$$
\left\| \mathbb{E}_{x \sim \mathcal{D}_d} [B_i(x)] - \frac{1}{|S|} \sum_{x \in S} B_i(x) \right\| \leq \epsilon|m_{3,i}|.
$$

We can set $t = T \log(d)/\log(k)$, then if

$$
|S| \geq \epsilon^{-2}(1 + 1/|m_{3,i}|^2)\text{poly}(T, \log d)
$$

with probability at least $1 - d^{-T}$,

$$
\left\| \mathbb{E}_{x \sim \mathcal{D}_d} [B_i(x)] - \frac{1}{|S|} \sum_{x \in S} B_i(x) \right\| \leq \epsilon|m_{3,i}|.
$$

Also note that for any symmetric $3$rd-order tensor $R$, the operator norm of $R$,

$$
\|R\| = \max_{\|a\|=1} |R(a,a,a)| \leq \max_{\|a\|=1} \|R(a,I,I)\|_F = \|R^{(1)}\|.
$$

\[
\square
\]

**Lemma E.9.** Let $M_4$ be defined as in Definition 5.1. Let $\widehat{M}_4$ be the empirical version of $M_4$, i.e.,

$$
\widehat{M}_4 = \frac{1}{|S|} \sum_{(x,y) \in S} y \cdot (x \otimes - (x \otimes x) \otimes I + I \otimes I),
$$

where $S$ is a set of samples (where each sample is i.i.d. sampled from Distribution $\mathcal{D}$ defined in Eq. (1)). Assume $M_4 \neq 0$, i.e., $m_{4,i} \neq 0$ for any $i$. Let $\alpha$ be a fixed unit vector. Let $V \in \mathbb{R}^{d \times k}$ be an orthogonal matrix satisfying $\|UU^\top - VV^\top\| \leq 1/4$, where $U$ is the orthogonal basis of span\{w_1^*, w_2^*, \cdots, w_k^*\}. Then for any $0 < \epsilon < 1, t \geq 1$, if

$$
|S| \geq \max_{i \in [k]} (1 + ||w_i^*||^{p+1}/|m_{4,i}(\alpha^\top \overline{w}_i^*)|^2) \cdot \epsilon^{-2} \cdot k^2 \cdot \text{poly}(\log d, t)
$$

with probability at least $1 - d^{-t}$,

$$
\|M_4(V, V, V, \alpha) - \widehat{M}_4(V, V, V, \alpha)\| \leq \epsilon \sum_{i=1}^{k} |v_i^* m_{4,i}(\alpha^\top \overline{w}_i^*)|.
$$

**Proof.** Recall that for each $(x,y) \in S$, we have $y = \sum_{i=1}^{k} v_i^* \phi(w_i^\top x)$. We consider each component $i \in [k]$. We define function $r(x) : \mathbb{R}^d \to \mathbb{R}^k$ such that

$$
r(x) = V^\top x.
$$

Define function $T_i(x) : \mathbb{R}^d \to \mathbb{R}^{k \times k \times k}$ such that

$$
T_i(x) = \phi(w_i^\top x) \left( x^\top \alpha \cdot r(x) \otimes r(x) \otimes r(x) - (V^\top \alpha) \otimes (r(x) \otimes r(x)) - \alpha^\top x \cdot r(x) \otimes I + (V^\top \alpha) \otimes I \right).
$$

We flatten $T_i(x) : \mathbb{R}^d \to \mathbb{R}^{k \times k \times k}$ along the first dimension to obtain function $B_i(x) : \mathbb{R}^d \to \mathbb{R}^{k \times k^2}$. Define $g(z) = \phi(z) - \phi(0)$, then $|g(z)| = \int_0^z |\phi'(s)ds| \leq L_1/(p+1)|z|^{p+1}$, which follows Property 3.1.
(I) Bounding \( \|B_i(x)\| \).
\[
\|B_i(x)\| \lesssim (|u_i^* x|^{p+1} + |\phi(0)|) \cdot (\|x^\top \alpha\|\|V^\top x\|^3 + 3\|V^\top \alpha\|\|V^\top x\|^2 \\
+ 3\|x^\top \alpha\|\|V^\top x_j\|\sqrt{k} + 3\|V^\top \alpha\|\sqrt{k})
\]
Note that \( V^\top x \sim \mathcal{N}(0, I_k) \). According to Fact B.1 and Fact B.2, we have for any constant \( t \geq 1 \), with probability \( 1 - 1/(nd^t) \),
\[
\|B_i(x)\| \lesssim (|u_i^*|^{p+1} + |\phi(0)|)k^{3/2} \text{poly}(\log d, t)
\]

(II) Bounding \( \|E_{x \sim D_d}[B_i(x)]\| \).
Note that \( E_{x \sim D_d}[B_i(x)] = m_{4,i}(\alpha^\top \overline{w}_i^*) (V^\top \overline{w}_i^*) \text{vec}((V^\top \overline{w}_i^*) (V^\top \overline{w}_i^*)^\top)^\top \). Therefore,
\[
\left\| E_{x \sim D_d}[B_i(x)] \right\| = |m_{4,i}(\alpha^\top \overline{w}_i^*)| \|V^\top \overline{w}_i^*\|^3.
\]
Since \( \|VV^\top - UU^\top\| \leq 1/4, \|VV^\top \overline{w}_i^* - \overline{w}_i^*\| \leq 1/4 \) and \( 3/4 \leq \|V^\top \overline{w}_i^*\| \leq 5/4 \). So \( \frac{1}{4}|m_{4,i}(\alpha^\top \overline{w}_i^*)| \leq \|E_{x \sim D_d}[B_i(x)]\| \leq 2|m_{4,i}(\alpha^\top \overline{w}_i^*)| \).

(III) Bounding \( \max(E_{x \sim D_d}[B_i(x) B_i(x)^\top], E_{x \sim D_d}[B_i(x)^\top B_i(x)]) \).
\[
\left\| E_{x \sim D_d}[B_i(x) B_i(x)^\top] \right\| \lesssim \left( E_{x \sim D_d}[\phi(w_i^* x)^4]\right)^{1/2} \left( E_{x \sim D_d}[(\alpha^\top x)^4]\right)^{1/2} \left( E_{x \sim D_d}[\|V^\top x\|^4]\right)^{1/2}
\lesssim (|u_i^*|^{p+1} + |\phi(0)|)^2 k^{3/2}.
\]
\[
\left\| E_{x \sim D_d}[B_i(x)^\top B_i(x)] \right\|
\lesssim \left( E_{x \sim D_d}[\phi(w_i^* x)^4]\right)^{1/2} \left( E_{x \sim D_d}[(\alpha^\top x)^4]\right)^{1/2} \left( E_{x \sim D_d}[\|V^\top x\|^4]\right)^{1/2}
\cdot \left( \max_{\|A\|_{\ell^1} = 1} E_{x \sim D_d}[\langle A, (V^\top x)(V^\top x)^\top\rangle^4]\right)^{1/2}
\lesssim (|u_i^*|^{p+1} + |\phi(0)|)^2 k^2.
\]

(IV) Bounding \( \max_{\|a\| = \|b\| = 1} (E_{x \sim D_d}[(a^\top B_i(x) b)^2])^{1/2} \).
\[
\max_{\|a\| = \|b\| = 1} \left( E_{x \sim D_d}[(a^\top B_i(x) b)^2]\right)^{1/2}
\lesssim \left( E_{x \sim D_d}[\phi^4(w_i^* x)]\right)^{1/4} \left( E_{x \sim D_d}[(\alpha^\top x)^4]\right)^{1/4} \max_{\|a\| = 1} \left( E_{x \sim D_d}[(a^\top V^\top x)^4]\right)^{1/2}
\cdot \left( \max_{\|A\|_{\ell^1} = 1} E_{x \sim D_d}[\langle A, (V^\top x)(V^\top x)^\top\rangle^4]\right)^{1/2}
\lesssim (|u_i^*|^{p+1} + |\phi(0)|) k.
\]
Define \( L = |u_i^*|^{p+1} + |\phi(0)| \). Then we have for any \( 0 < \epsilon < 1 \), if
\[
|S| \geq \frac{L^2k^2 + |m_{4,i}(\alpha^\top \overline{w}_i^*)|^2 + k^{3/2} \text{poly}(t, \log d)|m_{4,i}(\alpha^\top \overline{w}_i^*)|\epsilon}{\epsilon^2(m_{4,i}(\alpha^\top \overline{w}_i^*)^2)^2} \cdot t \log k
\]
with probability at least \(1 - k^{-t}\),
\[
\left\| \frac{1}{|S|} \sum_{x \in S} B_i(x) - \mathbb{E}_{x \sim D_d} [B_i(x)] \right\| \leq \epsilon |m_{4,i}(\alpha^\top \bar{w}_i^*)|.
\] (34)

We can set \(t = T \log(d) / \log(k)\), then if
\[
|S| \geq \frac{(L + |m_{4,i}(\alpha^\top \bar{w}_i^*)|)^2 k^2 \log(T, \log d) \cdot T \log^2 d}{\epsilon^2 |m_{4,i}(\alpha^\top \bar{w}_i^*)|^2}
\]
with probability at least \(1 - d^{-T}\), Eq. (34) holds. Also note that for any symmetric 3rd-order tensor \(R\), the operator norm of \(R\),
\[
\|R\| = \max_{\|a\|=1} \|R(a, a, a)\| \leq \max_{\|a\|=1} \|R(a, I, I)\|_F = \|R^{(1)}\|.
\]

### E.4.2 Final Error Bound for the Reduced Third-order Moment

Lemma E.10 shows \(\hat{R}_3\) can approximate \(R_3\) to some small precision with \(\text{poly}(k)\) samples.

**Lemma E.10** (Estimation of the reduced third order moment). Let \(U \in \mathbb{R}^{d \times k}\) denote the orthogonal column span of \(W^*\). Let \(\alpha\) be a fixed unit vector and \(V \in \mathbb{R}^{d \times k}\) denote an orthogonal matrix satisfying \(\|VV^\top - UU^\top\| \leq 1/4\). Define \(R_3 := P_3(V, V, V)\), where \(P_3\) is defined as in Definition 5.4 using \(\alpha\). Let \(\hat{R}_3\) be the empirical version of \(R_3\) using dataset \(S\), where each sample of \(S\) is i.i.d. sampled from distribution \(D\) (defined in (1)). Assume the activation function satisfies Property 3.1 and Assumption 5.3. Then for any \(0 < \epsilon < 1\) and \(t \geq 1\), define \(j_3 = \min\{j \geq 3 | M_j \neq 0\}\) and \(m_0 = \min_{i \in [k]} \{m_i^{(j_3)}(\alpha^\top \bar{w}_i^*)^{j_3 - 3}\}\), if
\[
|S| \geq \sigma_1^{2p+2} \cdot k^2 \cdot \text{poly}(\log d, t) / (\epsilon^2 m_0)
\]
then we have,
\[
\|R_3 - \hat{R}_3\| \leq \epsilon \sum_{i=1}^k |v_i^* m_{j_3,i}(\bar{w}_i^* \alpha)^{j_3 - 3}|,
\]
holds with probability at least \(1 - d^{-\Omega(t)}\).

**Proof.** The main idea is to use matrix Bernstein bound after matricizing the third-order tensor. Similar to the proof of Lemma E.5, we consider each node component individually and then sum up the errors and apply union bound.

We have shown the bound for \(j_3 = 3, 4\) in Lemma E.8 and Lemma E.9 respectively. To summarize, for any \(0 < \epsilon < 1\) we have if
\[
|S| \geq \max_{i \in [k]} \left(1 + \|w_i^*\|^{p+1} / |m_{j_3,i}(\bar{w}_i^* \alpha)^{j_3 - 3}|^2\right) \cdot \epsilon^{-2} \cdot k^2 \text{poly}(\log d, t)
\]
with probability at least \(1 - d^{-t}\),
\[
\|R_3 - \hat{R}_3\| \leq \epsilon \sum_{i=1}^k |v_i^* m_{j_3,i}(\bar{w}_i^* \alpha)^{j_3 - 3}|.
\]
E.5 Error Bound for the Magnitude and Sign of the Weight Vectors

The lemmata in this section together with Lemma E.5 provide guarantees for Algorithm 4. In particular, Lemma E.12 shows with linear sample complexity in $d$, we can approximate the 1st-order moment to some precision. And Lemma E.13 and Lemma E.14 provide the error bounds of linear systems Eq. (28) under some perturbations.

E.5.1 Robustness for Solving Linear Systems

**Lemma E.11** (Robustness of linear system). Given two matrices $A, \tilde{A} \in \mathbb{R}^{d \times k}$, and two vectors $b, \tilde{b} \in \mathbb{R}^d$. Let $z^* = \arg\min_{z \in \mathbb{R}^k} \|Az - b\|$ and $\tilde{z} = \arg\min_{z \in \mathbb{R}^k} \|(A + \tilde{A})z - (b + \tilde{b})\|$. If $\|\tilde{A}\| \leq \frac{1}{4\kappa} \sigma_k(A)$ and $\|\tilde{b}\| \leq \frac{1}{4}\|b\|$, then, we have
\[
\|z^* - \tilde{z}\| \lesssim (\sigma_k^{-4}(A)\sigma_2^2(A) + \sigma_k^{-2}(A))\|b\|\|\tilde{A}\| + \sigma_k^{-2}(A)\sigma_1(A)\|\tilde{b}\|.
\]

**Proof.** By definition of $z$ and $\tilde{z}$, we can rewrite $z$ and $\tilde{z}$,
\[
z = A^\dagger b = (A^\top A)^{-1} A^\top b
\]
\[
\tilde{z} = (A + \tilde{A})^\dagger (b + \tilde{b}) = ((A + \tilde{A})^\top (A + \tilde{A}))^{-1} (A + \tilde{A})^\top (b + \tilde{b}).
\]
As $\|\tilde{A}\| \leq \frac{1}{4\kappa} \sigma_k(A)$, we have $\|(A^\top A + A^\top \tilde{A})\|(A^\top A)^{-1}\| \leq 1/4$. Together with $\|\tilde{b}\| \leq \frac{1}{4}\|b\|$, we can ignore the high-order errors. So we have
\[
\|\tilde{z} - z^*\|
\lesssim \|(A^\top A)^{-1}(A^\top b + A^\top \tilde{b}) + (A^\top A)^{-1}(A^\top \tilde{A} + \tilde{A}^\top A)(A^\top A)^{-1} A^\top b\|
\lesssim \|(A^\top A)^{-1}(\|\tilde{A}\|\|b\| + \|A\|\|\tilde{b}\|) + ||(A^\top A)^{-2}\| \cdot \|A\|\|\tilde{A}\|\|A\|\|b\|
\lesssim \sigma_k^{-2}(A)(\|\tilde{A}\|\|b\| + \sigma_1(A)\|\tilde{b}\|) + \sigma_k^{-4}(A)\cdot \sigma_1^2(A)\|\tilde{A}\|\|\tilde{b}\|.
\]


E.5.2 Error Bound for the First-order Moment

**Lemma E.12** (Error bound for the first-order moment). Let $Q_1$ be defined as in Eq. (25) and $\hat{Q}_1$ be the empirical version of $Q_1$ using dataset $S$, where each sample of $S$ is i.i.d. sampled from distribution $D$ (defined in (1)). Assume the activation function satisfies Property 3.1 and Assumption 5.3. Then for any $0 < \epsilon < 1$ and $t \geq 1$, define $j_1 = \min\{j \geq 1| M_j \neq 0\}$ and $m_0 = \min_{i \in [k]}(m_{j_1,i}(\omega_i^\top \alpha)^{j_1 - 1})^2$ if
\[
|S| \geq \sigma_1^{2p+2}d\text{poly}(t, \log d)/(\epsilon^2 m_0)
\]
we have with probability at least $1 - d^{-\Theta(t)}$,
\[
\|Q_1 - \hat{Q}_1\| \leq \epsilon \sum_{i=1}^{k} |v_i^* m_{j_1,i}(\omega_i^\top \alpha)^{j_1 - 1}|.
\]

**Proof.** We consider the case when $l_1 = 3$, i.e.,
\[
Q_1 = M_3(I, \alpha, \alpha) = \sum_{i=1}^{k} v_i^* m_{3,i}(\alpha^\top \omega_i)^3 \omega_i^*.
\]
And other cases are similar.

Since \( y = \sum_{i=1}^{k} v_i^T \phi(w_i^T x) \). We consider each component \( i \in [k] \).
Define function \( B_i(x) : \mathbb{R}^d \to \mathbb{R}^d \) such that
\[
B_i(x) = [\phi(w_i^T x) \cdot (x^{\otimes 3} - x \otimes I)](I, \alpha, \alpha) = \phi(w_i^T x) \cdot ((x^\top \alpha)^2 x - 2(x^\top \alpha)x - x).
\]

Define \( g(z) = \phi(z) - \phi(0) \), then \( |g(z)| = |\int_0^z \phi'(s)ds| \leq L_1/(p+1)|z|^{p+1} \), which follows Property 3.1.

(I) Bounding \( \|B_i(x)\| \).
\[
\|B_i(x)\| \leq |\phi(w_i^T x)| \cdot \|((x^\top \alpha)^2 x - 2\alpha^\top x x - x)\| \\
\leq \|w_i^T x\|^{p+1} + |\phi(0)|((|x|^2 + 1)||x| + 2|\alpha^\top x|)
\]

According to Fact B.1 and Fact B.2, we have for any constant \( t \geq 1 \), with probability \( 1 - 1/(nd^t) \),
\[
\|B_i(x)\| \lesssim (\|w_i^*\|^{p+1} + |\phi(0)|)\sqrt{d} \text{poly}(\log d, t)
\]

(II) Bounding \( \|E_{x \sim D_d}[B_i(x)]\| \).
Note that \( E_{x \sim D_d}[B_i(x)] = m_{3,i}(w_i^\top \alpha)^2 \). Therefore, \( \|E_{x \sim D_d}[B_i(x)]\| = |m_{3,i}(w_i^\top \alpha)^2| \).

(III) Bounding \( \max \|E_{x \sim D_d}[B_i(x)]\|, \|E_{x \sim D_d}[B_i(x)^\top B_i(x)]\| \).
\[
\left\| \mathbb{E}_{x \sim D_d} \left[ B_i(x)^\top B_i(x) \right] \right\| \lesssim \left( \mathbb{E}_{x \sim D_d} \left[ \phi(w_i^T x)^4 \right] \right)^{1/2} \left( \mathbb{E}_{x \sim D_d} \left[ (x^\top \alpha)^8 \right] \right)^{1/2} \left( \mathbb{E}_{x \sim D_d} \left[ \|x\|^4 \right] \right)^{1/2} \\
\lesssim (\|w_i^*\|^{p+1} + |\phi(0)|)^2 d.
\]

(IV) Bounding \( \max_{\|a\|=\|b\|=1} \|E_{x \sim D_d}[(a^\top B_i(x)b)^2]\| \).
\[
\max_{\|a\|=\|b\|=1} \left( \mathbb{E}_{x \sim D_d} \left[ (a^\top B_i(x)a)^2 \right] \right)^{1/2} \lesssim \left( \mathbb{E}_{x \sim D_d} \left[ \phi^4(w_i^\top x) \right] \right)^{1/4} \lesssim \|w_i^*\|^{p+1} + |\phi(0)|.
\]

Define \( L = \|w_i^*\|^{p+1} + |\phi(0)| \). Then we have for any \( 0 < \epsilon < 1 \), if
\[
|S| \geq \frac{\epsilon^2 d + m_{3,i}(w_i^\top \alpha)^2 + L|m_{3,i}(w_i^\top \alpha)^2|\sqrt{d} \text{poly}(\log d, t)\epsilon}{\epsilon^2 m_{3,i}(w_i^\top \alpha)^2} \cdot t\log d
\]
with probability at least \( 1 - 1/d^t \),
\[
\left\| \mathbb{E}_{x \sim D_d} [B_i(x)] - \frac{1}{|S|} \sum_{x \sim S} B_i(x) \right\| \leq \epsilon|m_{3,i}(w_i^\top \alpha)^2|.
\]

Summing up all \( k \) components, we obtain if
\[
|S| \geq \max_{i \in [k]} \left\{ \left( \frac{\|w_i^*\|^{p+1} + |\phi(0)| + |m_{3,i}(w_i^\top \alpha)^2|}{m_{3,i}(w_i^\top \alpha)^2} \right) \cdot \epsilon^2 d \text{poly}(\log d, t) \right\}
\]
72
with probability at least $1 - 1/d'$,

$$\|M_3(I, \alpha, \alpha) - \hat{M}_3(I, \alpha, \alpha)\| \leq \epsilon \sum_{i=1}^{k} |v_i^* m_{3,i}(\tilde{w}_i^T \alpha)^2|.$$  

Other cases ($j_1 = 1, 2, 4$) are similar, so we complete the proof. \hfill \Box

### E.5.3 Linear System for the First-order Moment

The following lemma provides estimation error bound for the first linear system in Eq. (28).

**Lemma E.13** (Solution of linear system for the first order moment). Let $U \in \mathbb{R}^{d \times k}$ be the orthogonal column span of $W^*$. Let $V \in \mathbb{R}^{d \times k}$ denote an orthogonal matrix satisfying that $\|VV^T - UU^T\| \leq \hat{\delta}_2 \leq 1/(\kappa^2 \sqrt{k})$. For each $i \in [k]$, let $\hat{u}_i$ denote the vector satisfying $\|\hat{u}_i - V^T \tilde{w}_i^*\| \leq \hat{\delta}_3 \leq 1/(\kappa^2 \sqrt{k})$. Let $Q_1$ be defined as in Eq. (25) and $\hat{Q}_1$ be the empirical version of $Q_1$ such that $\|Q_1 - \hat{Q}_1\| \leq \hat{\delta}_4\|Q_1\| \leq \frac{1}{4}\|Q_1\|$. Let $z^* \in \mathbb{R}^k$ and $\tilde{z} \in \mathbb{R}^k$ be defined as in Eq. (27) and Eq. (28). Then

$$|\tilde{z}_i - z_i^*| \leq (\kappa^4 k^{3/2}(\hat{\delta}_2 + \hat{\delta}_3) + \kappa^2 k^{1/2}\hat{\delta}_4)|z^*_i|.$$  

**Proof.** Let $A \in \mathbb{R}^{k \times k}$ denote the matrix where the $i$-th column is $s_i \tilde{w}_i^*$. Let $\bar{A} \in \mathbb{R}^{k \times k}$ denote the matrix where the $i$-th column is $V \hat{u}_i$. Let $b \in \mathbb{R}^k$ denote the vector $Q_1$, let $\bar{b}$ denote the vector $\bar{Q}_1 - Q_1$. Then we have

$$\|A\| \leq \sqrt{k}.$$  

Using Fact B.3, we can lower bound $\sigma_k(A)$,

$$\sigma_k(A) \geq 1/k.$$  

We can upper bound $\|\bar{A}\|$ in the following way,

$$\|\bar{A}\| \leq \sqrt{k} \max_{i \in [k]} \{\|V \hat{u}_i - s_i \tilde{w}_i^*\|\}$$  

$$\leq \sqrt{k} \max_{i \in [k]} \{\|V \hat{u}_i - s_i VV^T \tilde{w}_i^* + s_i VV^T \tilde{w}_i^* - s_i UU^T \tilde{w}_i^*\|\}$$  

$$\leq \sqrt{k}(\hat{\delta}_3 + \hat{\delta}_2).$$  

We can upper bound $\|b\|$ and $\|\bar{b}\|$,

$$\|b\| = \|Q_1\| \leq k \sum_{i=1}^{k} |z_i^*|,$$  

and $\|\bar{b}\| \leq \hat{\delta}_4\|Q_1\|$.  

To apply Lemma E.11, we need $\hat{\delta}_4 \leq 1/4$ and $\hat{\delta}_2 \leq 1/(\sqrt{k}\kappa^2)$, $\hat{\delta}_3 \leq 1/(\sqrt{k}\kappa^2)$. So we have,

$$|\tilde{z}_i - z_i^*| \leq (\kappa^4 k^{3/2}(\hat{\delta}_2 + \hat{\delta}_3) + \kappa^2 k^{1/2}\hat{\delta}_4)\|Q_1\|$$  

$$\leq (\kappa^4 k^{3/2}(\hat{\delta}_2 + \hat{\delta}_3) + \kappa^2 k^{1/2}\hat{\delta}_4) \sum_{i=1}^{k} |z_i^*|.$$  

\hfill \Box
E.5.4 Linear System for the Second-order Moment

The following lemma provides estimation error bound for the second linear system in Eq. (28).

**Lemma E.14** (Solution of linear system for the second order moment). Let $U \in \mathbb{R}^{d \times k}$ be the orthogonal column span of $W^*$ denote an orthogonal matrix satisfying that $\|VV^T - UU^T\| \leq \tilde{\delta}_2 \leq 1/(\kappa \sqrt{k})$. For each $i \in [k]$, let $\tilde{u}_i$ denote the vector satisfying $\|\tilde{u}_i - V^T w^*_i\| \leq \tilde{\delta}_3 \leq 1/(\sqrt{k} \kappa^3)$. Let $Q_2$ be defined as in Eq. (26) and $\hat{Q}_2$ be the estimation of $Q_2$ such that $\|Q_2 - \hat{Q}_2\|_F \leq \tilde{\delta}_4 \|Q_2\|_F \leq \frac{1}{4} \|Q_2\|_F$. Let $r^* \in \mathbb{R}^k$ and $\hat{r} \in \mathbb{R}^k$ be defined as in Eq. (27) and Eq. (28). Then

$$|\hat{r}_i - r^*_i| \leq (k^3 \kappa^8 \tilde{\delta}_3 + \kappa^2 k^3) \|r^*\|_F.$$

**Proof.** For each $i \in [k]$, let $u_i = V^T w^*_i$. Let $A \in \mathbb{R}^{k^2 \times k}$ denote the matrix where the $i$-th column is $\text{vec}(u_i u_i^\top)$. Let $\hat{A} \in \mathbb{R}^{k^2 \times k}$ denote the matrix where the $i$-th column is $\text{vec}(u_i u_i^\top - \tilde{u}_i \tilde{u}_i^\top)$. Let $b \in \mathbb{R}^{k^2}$ denote the vector $\text{vec}(Q_2)$, let $\tilde{b} \in \mathbb{R}^{k^2}$ denote the vector $\text{vec}(Q_2 - \hat{Q}_2)$.

Let $\circ$ be the element-wise matrix product (a.k.a. Hadamard product), $\tilde{W} = [\tilde{w}_1 \tilde{w}_2 \cdots \tilde{w}_k]$ and $U = [u_1 \ u_2 \cdots \ u_k] = V^T \tilde{W}$. We can upper bound $\|A\|$ and $\|\hat{A}\|$ as follows,

$$\|A\| = \max_{\|x\| = 1} \left\| \sum_{i=1}^k x_i \text{vec}(u_i u_i^\top) \right\| = \max_{\|x\| = 1} \|U \text{diag}(x) U^\top\|_F \leq \|U\|^2 \leq \sigma_k^2 (V^T \tilde{W}),$$

and

$$\|\hat{A}\| = \sqrt{k} \max_{i \in [k]} \|\hat{A}_i\| \leq \sqrt{k} \max_{i \in [k]} \|u_i u_i^\top - \tilde{u}_i \tilde{u}_i^\top\|_F \leq \sqrt{k} \max_{i \in [k]} 2 \|u_i - \tilde{u}_i\|_2 \leq 2 \sqrt{k} \tilde{\delta}_3.$$

We can lower bound $\sigma_k(A),$

$$\sigma_k(A) = \sqrt{\sigma_k(A^\top A)} = \sqrt{\sigma_k((U^\top U) \circ (U^\top U))} = \min_{\|x\| = 1} \sqrt{x^\top (U^\top U) \circ (U^\top U)x} = \min_{\|x\| = 1} \|(U^\top U)^{1/2} \text{diag}(x)(U^\top U)^{1/2}\|_F \geq \sigma_k^2 (V^T \tilde{W})$$

where fourth step follows Schur product theorem, the last step follows by the fact that $\|CB\|_F \geq \sigma_{\min}(C) \|B\|_F$ and $\circ$ is the element-wise multiplication of two matrices.
We can upper bound $\|b\|$ and $\|\tilde{b}\|$, 

$$\|b\| \leq \|Q_2\|_F \leq \|r^*\|,$$

$$\|\tilde{b}\| = \|Q_2 - \hat{Q}_2\|_F \leq \hat{\delta}_4 \|r^*\|.$$ 

Since $\|VV^TW - \hat{W}\| \leq \sqrt{k}\hat{\delta}_2$, we have for any $x \in \mathbb{R}^k$, 

$$\|VV^TWx\| \geq \|Wx\| - \|(VV^TW - \hat{W})x\| \\
\geq \sigma_k(W)\|x\| - \hat{\delta}_2 \sqrt{k}\|x\|$$

Note that according to Fact B.3, $\sigma_k(W) \geq 1/\kappa$. Therefore, if $\hat{\delta}_2 \leq 1/(2\kappa\sqrt{k})$, we will have $\sigma_k(V^TW) \geq 1/(2\kappa)$. Similarly, we have $\sigma_1(V^TW) \leq \|V\||\hat{W}\| \leq \sqrt{k}$. Then applying Lemma E.11 and setting $\hat{\delta}_2 \lesssim \frac{1}{\sqrt{\kappa k^3}}$, we complete the proof.

\[ \square \]

**F Acknowledgments**

The authors would like to thank Surbhi Goel, Adam Klivans, Qi Lei, Eric Price, David P. Woodruff, Peilin Zhong, Hongyang Zhang and Jiong Zhang for useful discussions.