WEIGHTED REGULARIZATION METHOD FOR EFFICIENT NEURAL NETWORK COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Regularization is widely applied to model complexity reduction and neural network compression. Existing L_1 and nuclear norm regularizations can achieve favorable results, but these methods treat all parameters equally and ignore the importance of the parameters. Taking the trained parameters as prior information to construct weights, a weighted regularization method is proposed in this paper. Theoretically, we establish the bounds on the estimation errors for values of the global minimum for a fully connected single hidden layer neural network. Further we prove the estimates generated from the weighted L_1 regularization and the weighted nuclear norm regularization can recover the sparsity and the low rank structure of a global minimum of the neural network with a high probability, respectively. The effectiveness of the algorithm is validated by conducting a numerical simulation and experiments with popular neural networks on public datasets from real-world applications.

1 INTRODUCTION

Deep neural networks are often over-parameterized (Belkin et al., 2019). How to delete the redundant parameters of the network and keep the parameters that really work is a key concern of network compression. In network compression, parameter pruning, parameter quantization, low-rank approximation, and knowledge distillation are popular methods. Regularization is broadly applied to these methods to reduce the model complexity and alleviate the overfitting problem, and is now also widely used to prune unimportant parameters of the neural network to reduce computation and storage overhead.

There are many variants of regularization methods by choosing different regularization terms (Wen et al., 2016; He et al., 2017; Alvarez & Salzmann, 2017; Li et al., 2020; Liu et al., 2020), where L_1 regularization and nuclear norm regularization are two commonly used regularization methods. Specifically, L_1 regularization focuses more on promoting the sparsity of connections between neurons, namely unstructured pruning. Meanwhile, the nuclear norm regularization pays more attention to whether the parameter matrix can be decomposed into a more lightweight two-matrix product, that is, the low-rank approximation. Both regularization methods have the neural network compression effect (Jaderberg et al., 2014; Xu et al., 2019; Chen et al., 2020; Papadimitriou & Jain, 2021). However, existing regularization-based neural network compression techniques select the same regularization coefficient for each element in the parameter without distinguishing the importance of the elements, which tends to make the estimates of the really important elements small.

The weighted regularization idea is widely applied in compressed sensing (Daubechies et al., 2008; Chartrand & Yin, 2008; Candés et al., 2008; Wipf & Nagarajan, 2010; Ba et al., 2014) and image processing (Gu et al., 2014; Xu et al., 2017; Yair & Michaeli, 2018; Huang et al., 2020). Inspired by these, we use the inverse of the trained parameters estimates as weights applied to the regularization term to achieve adaptive penalty strength. Specifically, the elements with smaller true values correspond to larger weighted values, and the elements with larger true values correspond to relatively smaller weighted values. In this way, the purpose of pruning the redundant elements while protecting the truly critical elements is achieved. However, linear problems are usually considered in the above fields, and their theoretical analysis is not applicable to neural network compression. We theoretically prove that the weighted regularization method can accurately recover the ground truth parametric sparse structure and low-rank structure. Our simulation experiments demonstrate

that the sparsity and rank of the weighted regularization method are closer to the ground truth parameters than the general regularization method. In experiments with public datasets, the weighted regularization method can achieve higher accuracy with the same element sparsity or rank sparsity.

Organizations. The paper is divided into seven sections. Section 3 introduces the network model used in theoretical analysis. Section 4 provides the problem formulation. Section 5 presents the design idea and flow of the algorithm. Section 6 gives the assumptions and theoretical analysis. The last section shows experiments on synthetic data and real datasets.

Contributions. Our work makes the following contributions:

- For network compression, we establish weighted L_1 regularization method and weighted nuclear norm regularization methods respectively.
- We provide error upper bounds between the estimates obtained by two weighted regularization methods and some global minimum point of the expected risk with a high probability.
- We prove the zero element and the zero singular value can be correctly identified by selecting the appropriate regularization coefficient with only limited amount of data respectively.
- Experiments on synthetic and real datasets demonstrate that the weighted regularization methods outperform non-weighted counterparts, supporting the theoretical analysis.

2 Related work

Regularization-based network compression. According to the selection of regularization terms, the pruning methods by imposing regularization terms (Tang et al., 2022) can be divided into structured pruning (Wen et al., 2016; He et al., 2017; Scardapane et al., 2017; Li et al., 2019; Mitsuno & Kurita, 2021; Bui et al., 2021) and unstructured pruning (Louizos et al., 2017; Alvarez & Salzmann, 2017; Srinivas et al., 2017; Ma et al., 2019; Liu et al., 2020; Chen et al., 2020; Tartaglione et al., 2021; Pandit et al., 2021; Idelbayev & Carreira-Perpiñán, 2022), and the objects used by regularization terms can be divided into mask regularization and parameter regularization. Wen et al. (2016) proposed a structured sparsity learning (SSL) method whose main idea is that different regularization terms are applied to achieve different fine-grained structured pruning. In He et al. (2017), a hierarchical channel pruning method is obtained by setting a mask for each channel and applying L_1 regularization to the mask. Both nuclear norm regularization and grouped regularization are considered to promote the low-rank and group sparsity of the parameter matrix in Alvarez & Salzmann (2017). Structured pruning is realized by sparse optimization in Chen et al. (2020), namely L_1 regularization method, and an iterative algorithm is given. To the best of our knowledge, there is no weighted regularization method for network compression. In this paper, the proposed weighted L_1 regularization method is applied to unstructured pruning. Innovation from regularization term, Ma et al. (2019); Tartaglione et al. (2021); Pandit et al. (2021); Idelbayev & Carreira-Perpiñán (2022) introduce different regularization forms for network compression. Orthogonally, this paper considers how to improve the performance of the method for the fixed regularization term.

Low-rank approximation based network compression. The low-rank approximation (Jaderberg et al., 2014; Tai et al., 2015; Xu et al., 2019; Papadimitriou & Jain, 2021) usually decomposes a parameter matrix into a product of two matrices of smaller dimensions. In Jaderberg et al. (2014); Xu et al. (2019); Papadimitriou & Jain (2021), low-rank approximations of the original parameters are obtained by minimizing the nuclear norm. In Tai et al. (2015), low-rank approximation is achieved by adding low-rank constraints in the training process. To obtain a higher compression rate, Swaminathan et al. (2020) propose the sparse low rank (SLR) method which sparsifies the parameters while ensuring the low rank of the parameters. Determining the optimal rank, the key of low-rank approximation, Idelbayev & Carreira-Perpiñán (2020) regard it as a hyperparameter per layer and Kim et al. (2019) considers the optimal rank selection problem for the whole network. Yu et al. (2017) considers feature map reconstruction by setting parameter matrix as the sum of low-rank matrix and sparse matrix, to establish a unified framework of low-rank and sparse matrix. The proposed weighted idea is an orthogonal direction and can be applied to above nuclear norm regularization method to improve the low-rank approximation.

3 RESEARCH MODEL

For theoretical analysis, we consider a d-dimensional input and single output network with one hidden layer (Zhong et al., 2017; Oymak, 2018; Fu et al., 2020) which has K neurons, however, our experimental results show that the weighted regularization method also has significant advantages in deep networks. Similar to Zhong et al. (2017); Fu et al. (2020); Dinh & Ho (2020), assume there exists a lightweight underlying model, i.e., the "true" parameter $W^* = [w_1^*, \cdots, w_K^*] \in \mathbb{R}^{d\times K}$ which means the number of nonzero elements of W^* such that $\mathrm{supp}(W^*) < dK$ and $\mathrm{rank}(W^*) < \min\{d, K\}$. The corresponding "true" distribution of input and output over $\mathbb{R}^d \times \mathbb{R}$ is

$$\mathcal{D}: \ x \sim \mathcal{N}(0, I), \ y = \sum_{k=1}^{K} \phi\left(w_k^{*\top} x\right) \triangleq h(W^*; x) \tag{1}$$

where $\phi(z)$ is the activation function and $h(\cdot)$ is the network architecture.

The training dataset $\{x_i, y_i\}_{i=1}^N$ which are independent and identically distributed (i.i.d.) samples generated from $\mathcal{D}$ satisfies $y_i = h(W^*; x_i)$.

The *Empirical Risk* is defined as

$$f_N(W) = \frac{1}{2N} \sum_{i=1}^{N} \left(\sum_{k=1}^{K} \phi(w_k^\top x_i) - y_i \right)^2 \triangleq \frac{1}{N} \sum_{i=1}^{N} \ell\left(W; x_i, y_i\right) \tag{2}$$

where $\ell\left(W; x, y\right)$ is the mean square loss function, i.e., $\ell\left(W; x, y\right) \triangleq \frac{1}{2}\left(\sum_{k=1}^{K} \phi(w_k^\top x) - y\right)^2$.

The *Expected Risk* is defined as

$$f(W) = \frac{1}{2} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\left(\sum_{k=1}^{K} \phi(w_k^\top x) - y \right)^2 \right] \triangleq \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\ell\left(W; x, y\right) \right]. \tag{3}$$

4 PROBLEM FORMULATION

For the expected risk, denote the set of its global minimum points as $\mathcal{H}^* \triangleq \{W : f(W) = f(W^*)\}$, that is to say,

$$\mathcal{H}^* = \mathop{\mathrm{argmin}}_{W\in\mathbb{R}^{d\times K}} f(W). \tag{4}$$

By the proof of Lemma 3.1 in Dinh & Ho (2020),i.e., Lemma A.1 in Appendix A.1, all the global minima of the expected risk can be regarded as the ground truth parameter for generating data in neural networks. For any global minimum point of the empirical risk $f_N(W)$ denoted as

$$\widehat{W}_N \in \mathop{\mathrm{argmin}}_{W\in\mathbb{R}^{d\times K}} f_N(W). \tag{5}$$

Define the point in $\mathcal{H}^*$ closest to $\widehat{W}_N$ as $W_{\mathcal{H}^*} \triangleq \mathop{\mathrm{argmin}}_{W\in\mathcal{H}^*} \|W - \widehat{W}_N\|_F$.

In this work, we consider sparse and low-rank recovery of $W_{\mathcal{H}^*}$ by L_1 and nuclear regularization respectively. Assume $W_{\mathcal{H}^*} \in \mathbb{R}^{d\times K}$ is sparse and low-rank which means the number of nonzero elements of $W_{\mathcal{H}^*}$ such that $\mathrm{supp}(W_{\mathcal{H}^*}) = h < dK$ and $\mathrm{rank}(W_{\mathcal{H}^*}) = r < \min\{d, K\}$.

5 ALGORITHM DESIGN

To improve model generalization, reduce computation and storage consumption, we consider the regularization method. Existing regularization-based neural network compression techniques select the same regularization coefficient for each element in the parameter matrix without distinguishing the importance of the elements, which is unfair and will make the estimation value of the really important elements small. Thus the weighted regularization is introduced.

Weighted L_1 norm regularization. For $\forall W,\ A \in \mathbb{R}^{d\times K}$, if matrix A is used as the weight, the weighted L_1 norm of matrix W is defined as $\|W\|_{A,1} \triangleq \sum_{s=1}^{d}\sum_{t=1}^{K}\frac{|W(s,t)|}{|A(s,t)|}$.

Weighted nuclear norm regularization. For $\forall W,\ A \in \mathbb{R}^{d\times K}$ and A is full rank, if matrix A is used as the weight, the weighted nuclear norm of matrix W is defined as $\|W\|_{A,*} \triangleq \sum_{l=1}^{\min\{d,K\}}\frac{\sigma_l(W)}{\sigma_l(A)}$ where $\sigma_l(W)$ denotes the l-th singular value of the matrix W singular value in descending order, i.e., $\sigma_1(W) \geq \cdots \sigma_l(W) \geq \cdots \geq \sigma_{\min\{d,K\}}(W)$. The singular values of matrix A are defined the same way.

Firstly we obtain the initial estimate $\widehat{W}_N$ by minimizing the empirical risk, i.e., solving the optimization problem (5). Then use the initial estimate $\widehat{W}_N$ to construct weights according to different ways to form weighted L_1 norm regularization and nuclear norm regularization. Denote the Frobenius norm of matrix as $\|\cdot\|_F$. The square term $\|W-\widehat{W}_N\|_F^2$ in the optimization criterion (6) and (7) is to strengthen the convexity of the objective function at $W_{\mathcal{H}^*}$.

$$\min_{W\in\mathcal{B}(W_{\mathcal{H}^*},R)} J_{N,1}(W) \triangleq f_N(W) + \lambda_N\|W-\widehat{W}_N\|_F^2 + \gamma_N\|W\|_{\widehat{W}_N,1} \tag{6}$$

$$\min_{W\in\mathcal{B}(W_{\mathcal{H}^*},R)} J_{N,*}(W) \triangleq f_N(W) + \lambda_N\|W-\widehat{W}_N\|_F^2 + \gamma_N\|W\|_{\widehat{W}_N,*} \tag{7}$$

where R is the neighborhood radius centered on $W_{\mathcal{H}^*}$ that satisfies local convexity. The specific sparse recovery and low-rank recovery algorithm process are displayed in Algorithm 1 and 2 respectively.

Algorithm 1 Sparse recovery algorithm

Input: Training data $\{x_i,y_i\}_{i=1}^N$; quadratic term parameter $\lambda_N>0$; regularization term parameter $\gamma_N>0$;
Output: The sparse estimates $W_{N,1}$
1: compute initial estimates

$$\widehat{W}_N \in \underset{W\in\mathbb{R}^{d\times K}}{\operatorname{argmin}} f_N(W); \tag{8}$$

2: compute sparse estimates

$$W_{N,1} \triangleq \underset{W\in\mathcal{B}(\widehat{W}_N,\widehat{R})}{\operatorname{argmin}} f_N(W) + \lambda_N\|W-\widehat{W}_N\|_F^2 + \gamma_N\|W\|_{\widehat{W}_N^w,1}. \tag{9}$$

Algorithm 2 Low-rank recovery algorithm

Input: Training data $\{x_i,y_i\}_{i=1}^N$; quadratic term parameter $\lambda_N>0$; regularization term parameter $\gamma_N>0$;
Output: The low-rank estimates $W_{N,*}$
1: compute initial estimates

$$\widehat{W}_N \in \underset{W\in\mathbb{R}^{d\times K}}{\operatorname{argmin}} f_N(W); \tag{10}$$

2: compute low-rank estimates

$$W_{N,*} \triangleq \underset{W\in\mathcal{B}(\widehat{W}_N,\widehat{R})}{\operatorname{argmin}} f_N(W) + \lambda_N\|W-\widehat{W}_N\|_F^2 + \gamma_N\|W\|_{\widehat{W}_N^w,*}. \tag{11}$$

Remark 5.1 *Noting that the feasible domain in optimization criterion (6) and (7) is related to $W_{\mathcal{H}^*}$ which is not available, we consider replacing $W_{\mathcal{H}^*}$ with $\widehat{W}_N$ at the cost of making the radius of the feasible domain smaller. By Remark A.1 in Appendix A.2, we can choose $\widehat{R} \triangleq R - C_{\delta_1}\cdot\left(\frac{\log N}{\sqrt{N}}\right)^{1/\nu} > 0$ where $R = \min\left\{\sqrt{\frac{(2L_1+L_1L_2)^2}{2L_1^2L_3^2dK}}, \frac{\lambda_N}{K^{7/2}}\right\}$.*

Remark 5.2 *Noting that the adaptive weights in the regularization terms appear in the denominator, in order that the algorithms are well-defined, we set* $\widehat{W}_N^w(s,t) = \widehat{W}_N(s,t) + \operatorname{sgn}(\widehat{W}_N(s,t))\left(\frac{\log N}{\sqrt{N}}\right)^{1/\nu}$ *and* $\sigma_l^w(\widehat{W}_N) = \sigma_l(\widehat{W}_N) + \operatorname{sgn}(\sigma_l(\widehat{W}_N)\left(\frac{\log N}{\sqrt{N}}\right)^{1/\nu}$ *where* $\operatorname{sgn}(x) = 1$*, if* $x \geq 0$*;* $\operatorname{sgn}(x) = -1$*, if* $x < 0$*. By this way,* $C_1\left(\frac{\log N}{\sqrt{N}}\right)^{1/\nu} \leq |\widehat{W}_N^w(s,t)| \leq C_2\left(\frac{\log N}{\sqrt{N}}\right)^{1/\nu}$ *and* $C_3\left(\frac{\log N}{\sqrt{N}}\right)^{1/\nu} \leq \sigma_l^w(\widehat{W}_N) \leq C_4\left(\frac{\log N}{\sqrt{N}}\right)^{1/\nu}$ *where* $C_1,\ C_2,\ C_3$ *and* C_4 *are some constants.*

6 THE THEORETICAL RESULTS

We first make the following assumptions about the activation function and coefficient selection of quadratic term and regular term, i.e., λ_N and γ_N.

Assumption 6.1 *The activation function* $\phi(z)$ *is analytic. The first derivative* $\phi'(z)$ *is non-negative and bounded, i.e.,* $0 \leq \phi'(z) \leq L_1$ *for some constants* $L_1 > 0$*. The second derivative* $\phi''(z)$ *and the third derivative* $\phi'''(z)$ *are bounded, i.e.,* $|\phi''(z)| \leq L_2$ *and* $|\phi'''(z)| \leq L_3$ *for some constant* L_2 *and* L_3*.*

Assumption 6.2 *Let*

$$\alpha_q(\sigma) = \mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[\phi'(\sigma\cdot z)z^q\right], \forall q \in \{0,1,2\},$$
$$\beta_q(\sigma) = \mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[\phi'^2(\sigma\cdot z)z^q\right], \forall q \in \{0,2\},$$
$$\rho(\sigma) \triangleq \min\left\{\beta_0(\sigma) - \alpha_0^2(\sigma) - \alpha_1^2(\sigma),\ \beta_2(\sigma) - \alpha_1^2(\sigma) - \alpha_2^2(\sigma),\ \alpha_0(\sigma)\cdot\alpha_2(\sigma) - \alpha_1^2(\sigma)\right\}.$$

The first derivative $\phi'(z)$ *satisfies that, for all* $\sigma > 0$*, we have* $\rho(\sigma) > 0$*.*

Remark 6.1 *Some assumptions are similar to the assumption in Zhong et al. (2017); Oymak (2018). Generally speaking, they guarantee the local convexity of the neural network.*

It is direct to check that Sigmoid activation function $\phi(z) = \frac{1}{1+e^{-z}}$ *satisfies the above assumptions.*

Assumption 6.3 *For the coefficient of quadratic term* λ_N*, it satisfies* $\lambda_N > \max\left\{C_{\delta_1}\cdot\left(\frac{\log N}{\sqrt{N}}\right)^{1/\nu},\ C_{\delta_2}\sqrt{\frac{\log N}{N}}\right\}$ *and* $\lambda_N \to 0$ *as* $N \to \infty$*. For the coefficient of regularization term* γ_N*, it satisfies* $\frac{\gamma_N}{\left(\frac{\log N}{\sqrt{N}}\right)^{1/\nu}} \to \infty$ *and* $\gamma_N \to 0$ *as* $N \to \infty$ *where* ν *is some positive constant and* $C_{\delta_1},\ C_{\delta_2}$ *are positive constants depending on* δ_1 *and* δ_2 *respectively.*

6.1 THEORETICAL RESULTS OF SPARSE AND LOW-RANK RECOVERY ALGORITHM

This section includes two parts of theoretical analysis. The first part is to provide the upper bounds of the estimation errors obtained by the two weighted regularization methods respectively. The second part is to prove the sparse consistency and low-rank consistency.

6.1.1 ERROR BOUNDS FOR SPARSE AND LOW-RANK ESTIMATORS

Theorem 6.1 *(The error bound of the estimates by minimizing* $f_N(W)$*). Assume the activation function is analytic. Then for* $\forall\delta_1 > 0$*, there exist* $\nu > 0$*,* $C_{\delta_1} > 0$ *and* $N_0(\delta_1) > 0$ *for* $\forall N \geq N_0(\delta_1)$ *such that*

$$\|\widehat{W}_N - W_{\mathcal{H}^*}\|_F \leq C_{\delta_1}\left(\frac{\log N}{\sqrt{N}}\right)^{1/\nu} \tag{12}$$

with probability at least $1 - \delta_1$*.*

Proof *The proof is based on the corollary of Lojasewicz's inequality and generalization bound in Dinh & Ho (2020). The detailed proof is given in Appendix A.2.*

Lemma 6.1 *(The empirical Hessian is close to the expected Hessian). Assume Assumption 6.1 and 6.2 hold. The empirical Hessian converges uniformly to the expected Hessian. Namely,* $\forall\delta_2 > 0$*, there exist* $C_{\delta_2} > 0$ *and* $N_0(\delta_2) > 0$*, for* $\forall N \geq N_0(\delta_2)$ *we have*

$$\mathbb{P}\left(\sup_{W\in\mathcal{B}(W_{\mathcal{H}^*},R)}\left\|\nabla^2 f_N(W) - \nabla^2 f(W)\right\| \leq C_{\delta_2}\sqrt{\frac{\log N}{N}}\right) \geq 1 - \delta_2. \tag{13}$$

Proof *The proof is based on the covering number theory and Bernstein inequality inspired by Mei et al. (2018); Fu et al. (2020). The detailed proof is given in Appendix A.3.*

Lemma 6.2 *(Local uniform strong convexity and smoothness of $F_N(W)$). Denote $F_N(W) \triangleq f_N(W) + \lambda_N \|W - \widehat{W}_N\|_F^2$. Choose $\lambda_N > C_{\delta_2}\sqrt{\frac{\log N}{N}}$, for $\forall W \in \mathcal{B}(W_{\mathcal{H}^*}, R)$ where $R = \min\left\{\sqrt{\frac{(2L_1+L_1L_2)^2}{2L_1^2L_3^2dK}}, \frac{\lambda_N}{K^{7/2}}\right\}$, we have*

$$l_N \cdot I \preceq \nabla^2 F_N(W) \preceq L_N \cdot I. \tag{14}$$

with probability at least $1-\delta_2$, where $l_N \triangleq \lambda_N - C_{\delta_2}\sqrt{\frac{\log N}{N}} > 0$ and $L_N \triangleq C_{\delta_2}\sqrt{\frac{\log N}{N}} + L_f + 2\lambda_N$.

Proof *The proof is based on the locally convexity of expected risk $f(W)$ and this good property can be transferred to the empirical risk $f_N(W)$ by Lemma 6.1. Adding the square term, for $f_N(W)$, local convexity is strengthened to strong local convexity. The detailed proof is given in Appendix A.4.*

Theorem 6.2 *(Error bounds of $W_{N,1}$ and $W_{N,*}$). Let the activation function satisfy assumptions 6.1 and 6.2. Then for any $W \in \mathcal{B}(W_{\mathcal{H}^*}, R)$, $J_{N,1}(W)$ and $J_{N,*}(W)$ are strongly convex with probability at least $1-\delta$. Choose $\lambda_N > C_{\delta_2}\sqrt{\frac{\log N}{N}}$. Then there exist $N_0(\delta) > 0$, for $\forall N \geq N_0(\delta)$ we have $\|W_{N,1} - W_{\mathcal{H}^*}\|_F = O\left(\sqrt{\frac{\log N}{N}} + \lambda_N\left(\frac{\log N}{\sqrt{N}}\right)^{1/\nu} + \gamma_N\right)$ and $\|W_{N,*} - W_{\mathcal{H}^*}\|_F = O\left(\sqrt{\frac{\log N}{N}} + \lambda_N\left(\frac{\log N}{\sqrt{N}}\right)^{1/\nu} + \gamma_N\right)$ with probability at least $1-\delta$.*

Proof *The proof is based on the above strong local convexity, we use the theory of convex optimization to analyze the error upper bounds of the estimates obtained by the two weighted regularization methods. The detailed proof is given in Appendix A.6.*

Remark 6.2 *As long as λ_N and γ_N satisfy Assumption 6.3, we can deduce $W_{N,1} \to W_{\mathcal{H}^*}$ and $W_{N,*} \to W_{\mathcal{H}^*}$ as $N \to \infty$ with probability at least $1-\delta$.*

6.1.2 Sparse and low-rank recovery

For the ground truth parameter $W_{\mathcal{H}^*}$, define the zero element index set $\mathcal{A}_{\mathcal{H}^*}$ and zero singular value index set $\mathcal{B}_{\mathcal{H}^*}$ respectively.

$$\mathcal{A}_{\mathcal{H}^*} = \{(s,t) : W_{\mathcal{H}^*}(s,t) = 0,\ s = 1, \cdots, d;\ t = 1, \cdots, K\} \tag{15}$$

$$\mathcal{B}_{\mathcal{H}^*} = \{l : \sigma_l(W_{\mathcal{H}^*}) = 0,\ l = 1, \cdots, \min\{d, K\}\}. \tag{16}$$

Theorem 6.3 *(Sparse and low-rank selection consistency). Define*

$$\mathcal{A}_N = \{(s,t) : W_{N,1}(s,t) = 0,\ s = 1, \cdots, d;\ t = 1, \cdots, K\} \tag{17}$$

$$\mathcal{B}_N = \{l : \sigma_l(W_{N,*}) = 0,\ l = 1, \cdots, \min\{d, K\}\}. \tag{18}$$

Choose $\frac{\gamma_N}{\left(\frac{\log N}{\sqrt{N}}\right)^{1/\nu}} \to \infty$ as $N \to \infty$. There exist $N_1(\delta) > 0$, for $\forall N \geq N_1(\delta)$, we have $\mathcal{A}_N = \mathcal{A}_{\mathcal{H}^*}$ and $\mathcal{B}_N = \mathcal{B}_{\mathcal{H}^*}$ with probability at least $1-\delta$.

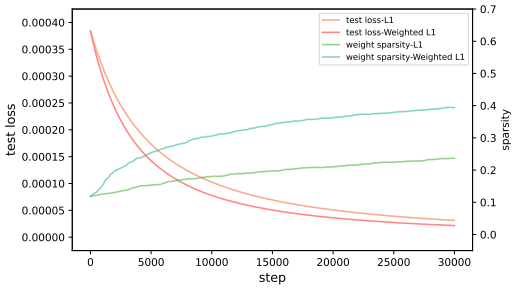
Proof *The proof is based on the optimality of $W_{N,1}$, $W_{N,*}$ and the selection of regularization coefficient γ_N in Assumption 6.3. The detailed proof is given in Appendix A.7.*

Remark 6.3 *Theorem 6.3 shows that the estimates obtained by the weighted regularization methods can correctly prune redundant elements and redundant singular values.*

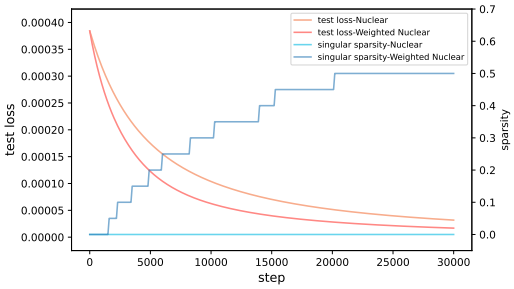
7 Experiment

7.1 Results on Synthetic Dataset

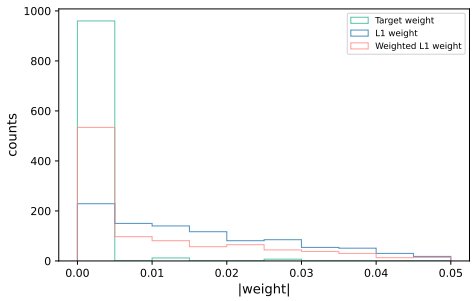
Settings. We consider synthetic simulations to support the theoretical analysis. The designed network is from equation (1) with the input dimension $d = 20$, hidden dimension $K = 80$ and activation function $\phi(z) = \mathrm{Sigmoid}(z)$. The ground truth parameter W^* is sparse and low-rank



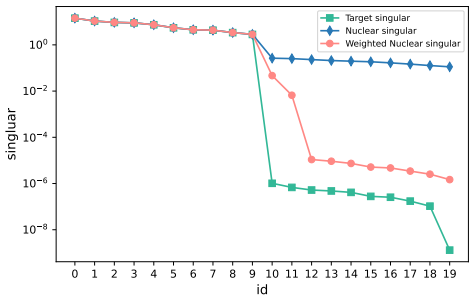
(a) L1 vs. Weighted L1



(b) Nuclear vs. Weighted Nuclear



(c) $w^*, \hat{w}_{l1}, \hat{w}_{weightedl1}$



(d) $s^*, \hat{s}_{nuc}, \hat{s}_{weightnuc}$

Figure 1: (a) The test loss and weight sparsity of L_1 and Weighted L_1 methods. (b) The test loss and singular sparsity of Nuclear and Weighted Nuclear methods. (c) The weight histogram of W^*, the L_1 method, and the Weighted L_1 method. (d) The singular values of W^*, the Nuclear method, and the Weighted Nuclear method.

with $\text{supp}(W^*) = 8 \times 80$ and $\text{rank}(W^*) = 10$. We compare the L_1 with Weighted L_1 methods (Algorithm 1), Nuclear with Weighted Nuclear methods (Algorithm 2). The gradient descent algorithm is employed with a learning rate of 0.2 and regularization coefficient $\gamma_N = 5$, and a number of iterations of $30,000$. The test error and weight/singular sparsity are used for evaluation. The followed sparsity represents the ratio of near zero weights/singulars. In practice, we set $\epsilon_w = 0.005, \epsilon_s = 0.05$.

$$\text{sparsity}_w(W) = \frac{\sum_i^d \sum_j^K I(|W_{ij}| < \epsilon_w)}{dK}, \ \text{sparsity}_s(W) = \frac{\sum_k^{\min\{d,K\}} I(|\sigma_k(W)| < \epsilon_s)}{\min\{d, K\}} \quad (19)$$

Results. Figure 1(a) 1(b) illustrate the test error and sparsity along iterations. For example, Figure 1(a) shows that the Weighted L_1 consistently has lower test error and higher sparsity than the L_1 method. These results indicate that the weighted method preserve more critical elements, which is consistent with our theoretical derivation. For further analysis, the distribution of weights and singular values are shown in Figure 1(c) 1(d), respectively. In Figure 1(c), the red Weighted L_1 has more parameters in the $[0, 0.005)$ interval than the blue L_1 method, thus stronger weight sparsity. In Figure 1(d), the singular values of the red Weighted Nuclear are closer to the target singular values than the blue Nuclear method when the ids lie between 10 and 19, indicating lower rank.

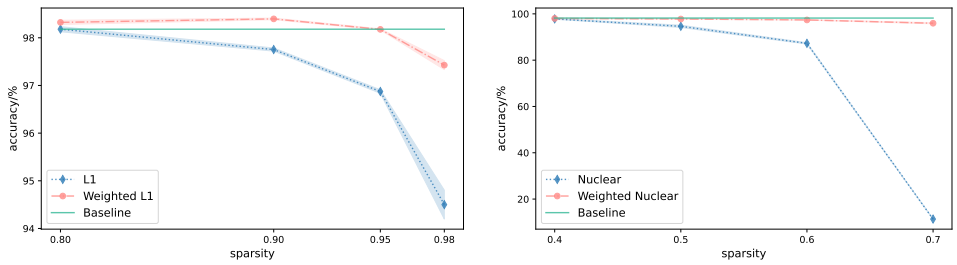
7.2 Results on Real Datasets

Datasets and Networks. Experiments are conducted on MNIST (LeCun et al., 1998), CIFAR-10 (Krizhevsky & Hinton, 2009) and Tiny-ImageNet (Le & Yang, 2015). MNIST is a dataset of handwritten digits with $60,000$ training images and $10,000$ test images. CIFAR-10 is a 10-class object recognition dataset with $50,000$ training images and $10,000$ test images. Tiny-ImageNet is a recognition dataset with two hundred classes, 100,000 training images, and 10,000 test images. The experimental networks include LeNet-300-100 (Krizhevsky & Hinton, 2009) for MNIST, VGG16

(Simonyan & Zisserman, 2015), ResNet20 (He et al., 2016), ResNet56 for CIFAR-10, and ResNet18 for Tiny-ImageNet.

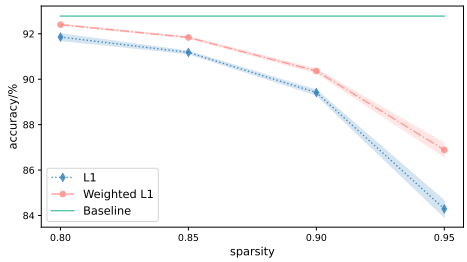
Settings. We compare the L_1 with the Weighted L_1 method, and the Nuclear with the Weighted Nuclear method. The weights of Weighted L_1 and Nuclear methods are initialized by the parameters of well-trained networks. Then the regularization is applied to compress networks. For a fair comparison, the hyper-parameters of non-weighted and weighted methods are kept the same during optimization. Each experiment is conducted three times across datasets and networks. The weight decay is set to 0 to eliminate its disturbance to the regularization. Details of other hyper-parameters such as learning rate and optimizer are listed in Appendix A.8.

Results. Table 1 shows the results between L_1 and Weighted L_1. Weighted L_1 generally has better performance than L_1 both in fully connected networks and convolutional neural networks. For example, Weighted L_1 has superior average accuracy to L_1 (86.89% vs. 84.29%) on ResNet-20 when the sparsity is 0.95, which indicates that the weighted method preserves more critical elements. Table 2 summarizes the results of the Nuclear with the Weighted Nuclear. We can see that the Weighted Nuclear significantly outperforms the Nuclear method, which suggests the weighted method is helpful to preserve the key singular values. Moreover, Figure 2 visualizes that the weighted method has higher accuracy than the non-weighted method at the same sparsity, verifying the effectiveness of the proposed methods.

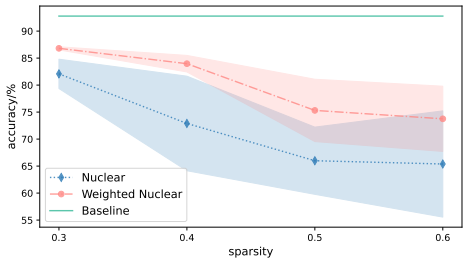


(a) MNIST, LeNet-300-100, L1 vs. Weighted L1

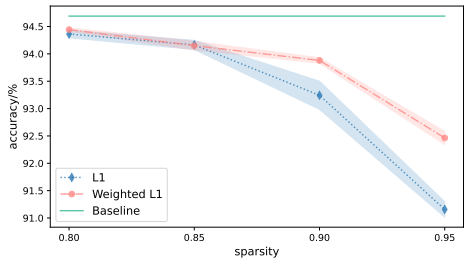
(b) MNIST, LeNet-300-100, Nuclear vs. Weighted Nuclear



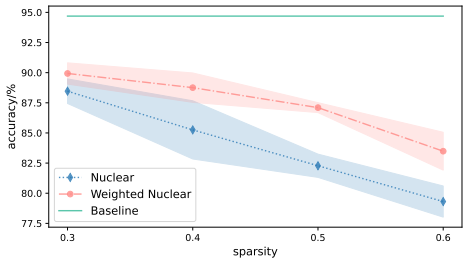
(c) CIFAR-10, ResNet20, L1 vs. Weighted L1



(d) CIFAR-10, ResNet20, Nuclear vs. Weighted Nuclear



(e) CIFAR-10, ResNet56, L1 vs. Weighted L1



(f) CIFAR-10, ResNet56, Nuclear vs. Weighted Nuclear

Figure 2: Comparison of non-weighted and weighted regularization methods on real datasets and networks.

Table 1: Comparison between L_1 and Weighted L_1 Regularization. The accuracy of Mean $\pm$ Std is reported by three independent experiments.

Dataset	Model	Baseline (%)	sparsity_w	L_1 (%)	Weighted L_1 (%)
MNIST	LeNet-300-100	98.18	0.9	97.75±0.03	**98.40±0.02**
			0.98	94.50±0.30	**97.43±0.09**
CIFAR-10	VGG16	93.14	0.8	93.21±0.07	**93.31±0.06**
			0.9	**93.30±0.01**	93.18±0.03
			0.99	**93.20±0.03**	93.19±0.18
	ResNet20	92.78	0.8	91.85±0.15	**92.40±0.03**
			0.9	89.41±0.11	**90.36±0.09**
			0.95	84.29±0.36	**86.89±0.30**
	ResNet56	94.69	0.8	94.36±0.07	**94.44±0.03**
			0.9	93.24±0.26	**93.88±0.05**
			0.95	91.15±0.14	**92.46±0.12**
Tiny-ImageNet	ResNet18	53.18	0.9	**51.21±1.24**	51.01±1.24
			0.95	47.74±3.56	**48.17±3.24**

Table 2: Comparison between Nuclear and Weighted Nuclear Regularization. The accuracy of Mean $\pm$ Std is reported by three independent experiments.

Dataset	Model	Baseline (%)	sparsity_s	Nuclear (%)	Weighted Nuclear (%)
MNIST	LeNet-300-100	98.18	0.5	94.62±0.49	**97.77±0.11**
			0.6	87.26±0.29	**97.38±0.09**
CIFAR-10	VGG16	93.14	0.5	91.76±0.65	**91.87±0.20**
			0.6	90.54±0.17	**91.12±0.33**
			0.7	87.18±0.77	**89.60±0.86**
	ResNet20	92.78	0.3	82.08±2.72	**86.81±0.29**
			0.4	72.90±8.76	**83.97±1.53**
			0.5	65.99±6.22	**75.31±5.77**
	ResNet56	94.69	0.3	88.46±1.02	**89.93±0.88**
			0.4	85.25±2.41	**88.76±1.22**
			0.5	82.28±0.97	**87.10±0.41**
Tiny-ImageNet	ResNet18	53.18	0.4	47.17±4.58	**50.77±1.59**
			0.5	44.29±6.19	**50.27±1.94**

8 CONCLUSION

In this paper, we propose two weighted regularization methods for sparse recovery and low-rank recovery respectively. For a fully connected single hidden layer neural network, we theoretically establish the error upper bound of the estimates obtained by the two methods and prove the sparse recovery consistency and low-rank recovery consistency respectively. The simulation experiment shows that the sparsity and rank of the weighted method are closer to the ground truth parameters and have better generalization, that is, the test error is smaller. Experiments on public datasets indicate that the weighted regularization method can achieve higher accuracy with the same element sparsity or singular value sparsity. However, the limitation of the proposed method is that the weight setting of the regularization term depends on the accuracy of the initial estimation $\widehat{W}_N$ to some sparse and low-rank minimum point. In the future work, more forms of weighted regularization methods, such as the non-convex regularization term, will be considered, and deeper neural networks will be theoretically analyzed.

REFERENCES

Jose M Alvarez and Mathieu Salzmann. Compression-aware training of deep networks. In *Advances in Neural Information Processing Systems*, 2017.

Demba Ba, Behtash Babadi, Patrick L. Purdon, and Emery N. Brown. Convergence and stability of iteratively re-weighted least squares algorithms. *IEEE Transactions on Signal Processing*, 2014.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 2019.

Kevin Bui, Fredrick Park, Shuai Zhang, Yingyong Qi, and Jack Xin. Structured sparsity of convolutional neural networks via nonconvex sparse group regularization. *Frontiers in applied mathematics and statistics*, 2021.

Emmanuel J. Candés, Michael B. Wakin, and Stephen P. Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier Analysis and Applications*, 2008.

Rick Chartrand and Wotao Yin. Iteratively reweighted algorithms for compressive sensing. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008.

Tianyi Chen, Bo Ji, Yixin Shi, Tianyu Ding, Biyi Fang, Sheng Yi, and Xiao Tu. Neural network compression via sparse optimization. *arXiv preprint arXiv:2011.04868*, 2020.

Ingrid Daubechies, Ronald DeVore, Massimo Fornasier, and Sinan Gunturk. Iteratively re-weighted least squares minimization: Proof of faster than linear rate for sparse recovery. In *Annual Conference on Information Sciences and Systems*, 2008.

Vu C Dinh and Lam S Ho. Consistent feature selection for analytic deep neural networks. In *Advances in Neural Information Processing Systems*, 2020.

Haoyu Fu, Yuejie Chi, and Yingbin Liang. Guaranteed recovery of one-hidden-layer neural networks via cross entropy. *IEEE Transactions on Signal Processing*, 2020.

Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng. Weighted nuclear norm minimization with application to image denoising. In *Conference on Computer Vision and Pattern Recognition*, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, 2016.

Yihui He, Xiangyu Zhang, and Jian Sun. Channel pruning for accelerating very deep neural networks. In *International Conference on Computer Vision*, 2017.

Yan Huang, Guisheng Liao, Yijian Xiang, Lei Zhang, Jie Li, and Arye Nehorai. Low-rank approximation via generalized reweighted iterative nuclear and frobenius norms. *IEEE Transactions on Image Processing*, 2020.

Yerlan Idelbayev and Miguel Á. Carreira-Perpiñán. Low-rank compression of neural nets: Learning the rank of each layer. In *Conference on Computer Vision and Pattern Recognition*, 2020.

Yerlan Idelbayev and Miguel Á. Carreira-Perpiñán. Exploring the effect of ℓ_0/ℓ_2 regularization in neural network pruning using the lc toolkit. In *International Conference on Acoustics, Speech and Signal Processing*, 2022.

Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman. Speeding up convolutional neural networks with low rank expansions. *arXiv preprint arXiv:1405.3866*, 2014.

Hyeji Kim, Muhammad Umar Karim Khan, and Chong-Min Kyung. Efficient neural network compression. In *Conference on Computer Vision and Pattern Recognition*, 2019.

A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 2015.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

Jiashi Li, Qi Qi, Jingyu Wang, Ce Ge, Yujian Li, Zhangzhang Yue, and Haifeng Sun. Oicsr: Out-in-channel sparsity regularization for compact deep neural networks. In *Conference on Computer Vision and Pattern Recognition*, 2019.

Yawei Li, Shuhang Gu, Christoph Mayer, Luc Van Gool, and Radu Timofte. Group sparsity: The hinge between filter pruning and decomposition for network compression. In *Conference on Computer Vision and Pattern Recognition*, 2020.

Junjie Liu, Zhe Xu, Runbin Shi, Ray CC Cheung, and Hayden KH So. Dynamic sparse training: Find efficient sparse network from scratch with trainable masked layers. In *International Conference on Learning Representations*, 2020.

Christos Louizos, Max Welling, and Diederik P Kingma. Learning sparse neural networks through l_0 regularization. *arXiv preprint arXiv:1712.01312*, 2017.

Rongrong Ma, Jianyu Miao, Lingfeng Niu, and Peng Zhang. Transformed ℓ_1 regularization for learning sparse deep neural networks. *Neural Networks*, 2019.

Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 2018.

Kakeru Mitsuno and Takio Kurita. Filter pruning using hierarchical group sparse regularization for deep convolutional neural networks. In *International conference on pattern recognition*, 2021.

Samet Oymak. Learning compact neural networks with regularization. In *International Conference on Machine Learning*, 2018.

Mohammad Khalid Pandit, Roohie Naaz, and Mohammad Ahsan Chishti. Learning sparse neural networks using non-convex regularization. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2021.

Dimitris Papadimitriou and Swayambhoo Jain. Data-driven low-rank neural network compression. In *International Conference on Image Processing*, 2021.

Simone Scardapane, Danilo Comminiello, Amir Hussain, and Aurelio Uncini. Group sparse regularization for deep neural networks. *Neurocomputing*, 2017.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

Suraj Srinivas, Akshayvarun Subramanya, and R. Venkatesh Babu. Training sparse neural networks. In *Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

Sridhar Swaminathan, Deepak Garg, Rajkumar Kannan, and Frederic Andres. Sparse low rank factorization for deep neural network compression. *Neurocomputing*, 2020.

Cheng Tai, Tong Xiao, Yi Zhang, Xiaogang Wang, and Weinan E. Convolutional neural networks with low-rank regularization. *arXiv preprint arXiv:1511.06067*, 2015.

Anda Tang, Pei Quan, Lingfeng Niu, and Yong Shi. A survey for sparse regularization based compression methods. *Annals of Data Science*, 2022.

Enzo Tartaglione, Andrea Bragagnolo, Francesco Odierna, Attilio Fiandrotti, and Marco Grangetto. Serene: Sensitivity-based regularization of neurons for structured sparsity in neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Wei Wen, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Learning structured sparsity in deep neural networks. In *Advances in Neural Information Processing Systems*, 2016.

David Wipf and Srikantan Nagarajan. Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions. *IEEE Journal of Selected Topics in Signal Processing*, 2010.

Jun Xu, Lei Zhang, David Zhang, and Xiangchu Feng. Multi-channel weighted nuclear norm minimization for real color image denoising. In *International Conference on Computer Vision*, 2017.

Yuhui Xu, Yuxi Li, Shuai Zhang, Wei Wen, Botao Wang, Wenrui Dai, Yingyong Qi, Yiran Chen, Weiyao Lin, and Hongkai Xiong. Trained rank pruning for efficient deep neural networks. In *Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS Edition*, 2019.

Noam Yair and Tomer Michaeli. Multi-scale weighted nuclear norm image restoration. In *Conference on Computer Vision and Pattern Recognition*, 2018.

Xiyu Yu, Tongliang Liu, Xinchao Wang, and Dacheng Tao. On compressing deep models by low rank and sparse decomposition. In *Conference on Computer Vision and Pattern Recognition*, 2017.

Kai Zhong, Zhao Song, Prateek Jain, Peter L. Bartlett, and Inderjit S. Dhillon. Recovery guarantees for one-hidden-layer neural networks. In *International Conference on Machine Learning*, 2017.

A APPENDIX

A.1 PRELIMINARY

For $\forall W \in \mathbb{R}^{d\times K}$, denote the k-th column of W as w_k. Given that $y = \sum_{k=1}^{K} \phi\left(w_k^{*\top} x\right)$ is only related to x, $\ell\left(W; x, y\right)$ is often abbreviated as $\ell\left(W; x\right)$ for the convenience of statement in subsequent theoretical analysis. For subsequent theoretical analysis, we first calculate the gradient and the Hessian of $f_N(W)$ and $f(W)$. For each $j \in \{1, \cdots, K\}$, the partial gradient of $f(W)$ with respect to w_j can be represented as

$$\frac{\partial f(W)}{\partial w_j} = \mathbb{E}\left[\left(\sum_{k=1}^{K} \phi\left(w_k^\top x\right) - y\right) \phi'\left(w_j^\top x\right) x\right] \tag{20}$$

For each $j, l \in \{1, \cdots, K\}$, the second partial derivative of $f(W)$ for the (j, l)-th block is,

$$\frac{\partial^2 f(W)}{\partial w_j \partial w_l} = \mathbb{E}\left[\xi_{j,l}(W; x) \cdot x x^\top\right] \tag{21}$$

where

$$\xi_{j,l}(W; x) = \begin{cases} \phi'\left(w_j^\top x\right) \phi'\left(w_l^\top x\right) & \text{for } j \neq l, \\ \left(\sum_{k=1}^{K} \phi\left(w_k^\top x\right) - y\right) \phi''\left(w_j^\top x\right) + \left(\phi'\left(w_j^\top x\right)\right)^2 & \text{for } j = l. \end{cases} \tag{22}$$

Accordingly, the gradient of empirical risk is $\nabla f_N(W) = [\frac{\partial f_N(W)}{\partial w_1}, \cdots, \frac{\partial f_N(W)}{\partial w_K}]$, where for $\forall j \in [K]$,

$$\frac{\partial f_N(W)}{\partial w_j} = \frac{1}{N} \sum_{i=1}^{N} \left[\left(\sum_{k=1}^{K} \phi\left(w_k^\top x_i\right) - y_i\right) \phi'\left(w_j^\top x_i\right) x_i\right]. \tag{23}$$

The Hessian matrix of empirical risk is $\nabla^2 f_N(W) = [\frac{\partial f_N(W)}{\partial w_j \partial w_l}]_{j\in[K], l\in[K]}$, where

$$\frac{\partial^2 f_N(W)}{\partial w_j \partial w_l} = \frac{1}{N} \sum_{i=1}^{N} \left[\xi_{j,l}(W; x_i) \cdot x_i x_i^\top\right]. \tag{24}$$

Lemma A.1 *$W \in \mathcal{H}^*$ if and only if $h(W; x) = h(W^*; x)$, $\forall x \in \mathbb{R}^d$, i.e., $\sum_{k=1}^{K} \phi(w_i^T x) = \sum_{k=1}^{K} \phi(w_i^{*T} x)$, $\forall x \in \mathbb{R}^d$.*

By Lemma A.1, we can denote any point in $\mathcal{H}^*$ as W^* and it satisfies $y = \sum_{k=1}^{K} \phi\left(w_k^{*\top} x\right)$.

Lemma A.2 *(The intersection of events with high probability is still a high probability event). $\forall \delta_A > 0$, there exists $N_0(\delta_A) > 0$, for $\forall N \geq N_0(\delta_A)$, we have $\mathbb{P}(A_N) \geq 1 - \delta_A$.*

$\forall \delta_B > 0$, there exists $N_0(\delta_B) > 0$, for $\forall N \geq N_0(\delta_B)$, we have $\mathbb{P}(B) \geq 1 - \delta_B$.

Then for $\forall N \geq \max\{N_0(\delta_A), N_0(\delta_B)\}$, we have $\mathbb{P}(AB) \geq 1 - (\delta_A + \delta_B)$.

Proof *Denote the complement of set B as B^C. Noting that $P(B^C) \leq \delta_B$, we have $\mathbb{P}(AB) = P(A) - P(AB^C) \geq 1 - \delta_A - P(AB^C) \geq 1 - \delta_A - P(B^C) \geq 1 - \delta_A - \delta_B$ where the second equality comes from $P(AB^C) \leq P(B^C) \leq \delta_B$.* ■

A.2 THE PROOF OF THEOREM 6.1

Before the proof of Theorem 6.1, two key lemmas are shown. The first lemma can be understood as an extension of Taylor's expansion when the Hessian matrix is singular.

Lemma A.3 *(Lemma 3.2 in Dinh & Ho (2020)). There exist C_1, $\nu > 0$ and such that $f(W) - f(W_{\mathcal{H}^*}) \geq C_1 \cdot dist\left(W, \mathcal{H}^*\right)^\nu$ for all $W_{\mathcal{H}^*} \in \mathcal{H}^*$ and $W \in \mathcal{B}\left(W_{\mathcal{H}^*}, R\right) \triangleq \left\{W \in \mathbb{R}^{d\times K} : \|W - W_{\mathcal{H}^*}\|_F \leq R\right\}$.*

The second one provides the generalization bound.

Lemma A.4 *(Lemma 3.3 in Dinh & Ho (2020)). For any $\delta_1 > 0$, there exist $C_{\delta_1} > 0$ and $N_{\delta_1} > 0$ such that $\forall N \geq N_{\delta_1}$ the generalization bound is*

$$|f_N(W) - f(W)| = C_{\delta_1} \cdot \frac{\log N}{\sqrt{N}},\ \forall W \in \mathcal{B}\left(W_{\mathcal{H}^*}, R\right) \tag{25}$$

with probability at least $1 - \delta_1$.

Based on Lemma A.3 and A.4, we prove Lemma 6.1.

Proof *Note* $\operatorname{dist}\left(\widehat{W}_N, \mathcal{H}^*\right) \triangleq \min_{W \in \mathcal{H}^*} \|W - \widehat{W}_N\|_F$. *Define* $W_{\mathcal{H}^*} \triangleq \operatorname*{argmin}_{W \in \mathcal{H}^*} \|W - \widehat{W}_N\|_F$. *Then* $\operatorname{dist}\left(\widehat{W}_N, \mathcal{H}^*\right) = \|W_{\mathcal{H}^*} - \widehat{W}_N\|_F$. *By Lemma A.3, we can obtain*

$$\begin{aligned} C_1 dist\left(\widehat{W}_N, \mathcal{H}^*\right)^\nu &= C_1 \|\widehat{W}_N - W_{\mathcal{H}^*}\|_F^\nu \\ &\leq f(\widehat{W}_N) - f(W_{\mathcal{H}^*}) \\ &= [f(\widehat{W}_N) - f_N(\widehat{W}_N)] + [f_N(\widehat{W}_N) - f_N(W_{\mathcal{H}^*})] + [f_N(W_{\mathcal{H}^*}) - f(W_{\mathcal{H}^*})] \end{aligned} \tag{26}$$

For the second term to the right of the inequality, the optimality of $\widehat{W}_N$ follows

$$f_N(\widehat{W}_N) - f_N(W_{\mathcal{H}^*}) \leq 0. \tag{27}$$

Combined with absolute value inequality, further directly we have

$$C_1 dist\left(\widehat{W}_N, \mathcal{H}^*\right)^\nu \leq |f(\widehat{W}_N) - f_N(\widehat{W}_N)| + |f_N(W_{\mathcal{H}^*}) - f(W_{\mathcal{H}^*})| \tag{28}$$

Then we just have to consider generalization bounds which is the bound of $|f_N(W) - f(W)|$. Combined with Lemma A.4, there exists $C_2 > 0$ such that

$$C_1 \|\widehat{W}_N - W_{\mathcal{H}^*}\|_F^\nu \leq |f(\widehat{W}_N) - f_N(\widehat{W}_N)| + |f_N(W_{\mathcal{H}^*}) - f(W_{\mathcal{H}^*})| \leq 2C_{\delta_1} \cdot \frac{\log N}{\sqrt{N}}. \tag{29}$$

Further, $\|\widehat{W}_N - W_{\mathcal{H}^*}\|_F \leq C_{\delta_1} \cdot \left(\frac{\log N}{\sqrt{N}}\right)^{1/\nu} = O\left(\left(\frac{\log N}{\sqrt{N}}\right)^{1/\nu}\right)$ can be drawn directly. ■

Remark A.1 *Since point $W_{\mathcal{H}^*}$ in local region $\mathcal{B}(W_{\mathcal{H}^*}, R)$ of Algorithm 1 and 2 is unknown, consider replacing $W_{\mathcal{H}^*}$ with the estimated value $\widehat{W}_N$. For any point $W \in \mathcal{B}(W_{\widehat{W}_N}, R)$, $\|W - \widehat{W}_N\|_F \leq R$. Note that*

$$\begin{aligned} \|W - W_{\mathcal{H}^*}\|_F &\leq \|(W - \widehat{W}_N) + (\widehat{W}_N - W_{\mathcal{H}^*})\|_F \leq \|W - \widehat{W}_N\|_F + \|\widehat{W}_N - W_{\mathcal{H}^*}\|_F \\ &\leq \|W - \widehat{W}_N\|_F + C_{\delta_1} \cdot \left(\frac{\log N}{\sqrt{N}}\right)^{1/\nu} \leq R. \end{aligned} \tag{30}$$

Noting that $\|W - \widehat{W}_N\|_F \leq R - C_{\delta_1} \cdot \left(\frac{\log N}{\sqrt{N}}\right)^{1/\nu} \triangleq \widehat{R}$, we know that $\mathcal{B}(W_{\widehat{W}_N}, \widehat{R})$ is a subset of $\mathcal{B}(W_{\mathcal{H}^*}, R)$. By $\lambda_N > \max\left\{C_{\delta_1} \cdot \left(\frac{\log N}{\sqrt{N}}\right)^{1/\nu}, C_{\delta_2}\sqrt{\frac{\log N}{N}}\right\}$ in Assumption 6.3, we have $\widehat{R} > 0$.

A.3 THE PROOF OF LEMMA 6.1

Definition A.1 *(Sub-exponential norm). For random variable $X \in \mathbb{R}^d$, the sub-exponential norm of X is defined as*

$$\|X\|_{\psi_1} \triangleq \sup_{t \geq 1} \frac{1}{t} \left[\mathbb{E}|X|^t\right]^{1/t}. \tag{31}$$

Definition A.2 *(Sub-gaussian norm). For random variable $X \in \mathbb{R}^d$, the sub-gaussian norm of X is defined as*

$$\|X\|_{\psi_2} \triangleq \sup_{t \geq 1} \frac{1}{\sqrt{t}} \left[\mathbb{E}|X|^t\right]^{1/t}. \tag{32}$$

Lemma A.5 *(Bernstein inequality for sub-exponential random variables). Let $X_1, \cdots, X_n$ be independent sub-exponential random variables with $\|X_i\|_{\psi_1} \leq b$, and define $S_n = \sum_{i=1}^n (X_i - \mathbb{E}X_i)$. Then there exists a universal constant c such that, for all $t > 0$,*

$$\mathbb{P}(S_n \geq t) \leq \left\{-c \min\left(\frac{t^2}{nb^2}, \frac{t}{b}\right)\right\}. \tag{33}$$

Lemma A.6 *(Lemma 4 in Mei et al. (2018)). Let $M \in \mathbb{R}^{d \times d}$ be a symmetric $d \times d$ matrix and V_ϵ be an ϵ-cover of unit-Euclidean-norm ball $\mathcal{B}(0, 1)$, then*

$$\|M\| \leq \frac{1}{1 - 2\epsilon} \sup_{v \in V_\epsilon} |\langle v, Mv \rangle|.$$

Firstly, we prove that $\|G_i\|_{\psi_1}$ is upper bounded.

Lemma A.7 *($\|G_i\|_{\psi_1}$ is upper bounded). For $G_i = \left\langle u, \left(\nabla^2 \ell\left(W; x_i\right) - \mathbb{E}\left[\nabla^2 \ell(W; x)\right]\right) u \right\rangle$ where $u \in \mathcal{B}(\mathbf{0}, 1) = \left\{W \in \mathbb{R}^{d \times K} : \|W\|_{\mathrm{F}} = 1\right\}$. There exists some constant C such that*

$$\|G_i\|_{\psi_1} \leq C :\equiv \tau^2. \tag{34}$$

Proof *Note $\|G_i\|_{\psi_1} \leq \|\langle u, \nabla \ell(W; x) u \rangle\|_{\psi_1} + \|\nabla^2 f(W; x)\|$. For the upper bound of $\|\nabla^2 f(W; x)\|$ is known, now let's focus on the first term with $u = \left[u_1^\top, \cdots, u_K^\top\right]^\top \in \mathbb{R}^{dK}$.*

$$\begin{aligned}
\|\langle u, \nabla \ell(W; x) u \rangle\|_{\psi_1} &\leq \sum_{j=1}^K \sum_{l=1}^K \left\|\xi_{jl} \cdot u_j^\top x x^\top u_l\right\|_{\psi_1} \\
&\leq \sum_{j=1}^K \sum_{l=1}^K \sup_{t \geq 1} t^{-1} \left(\mathbb{E}\left|\xi_{jl} \cdot u_j^\top x x^\top u_l\right|^t\right)^{1/t}
\end{aligned} \tag{35}$$

where

$$\xi_{j,l}(W) = \begin{cases} \phi'\left(w_j^\top x\right) \phi'\left(w_l^\top x\right) & \text{for } j \neq l, \\ \left(\sum_{k=1}^K \phi\left(w_k^\top x\right) - y\right) \phi''\left(w_j^\top x\right) + \left(\phi'\left(w_j^\top x\right)\right)^2 & \text{for } j = l. \end{cases} \tag{36}$$

By the Holder inequality, we can get that

$$\left(\mathbb{E}\left|\xi_{jl} \cdot u_j^\top x x^\top u_l\right|^t\right)^{1/t} = \left(\mathbb{E}\left|\xi_{jl}\right|^t \cdot \left|u_j^\top x x^\top u_l\right|^t\right)^{1/t} \leq \left(\sqrt{\mathbb{E}\left|\xi_{jl}\right|^{2t}} \cdot \sqrt{\mathbb{E}\left|u_j^\top x x^\top u_l\right|^{2t}}\right)^{1/t} \tag{37}$$

For $\forall j, l \in [K]$, we have

$$\begin{aligned}
|\xi_{jl}| &\leq L_2 \left|\left(\sum_{k=1}^K \phi\left(w_k^\top x\right) - \sum_{k=1}^K \phi\left(w_k^{*\top} x\right)\right)\right| + L_1^2 \\
&\leq L_1 L_2 \sum_{k=1}^K \|w_k - w_k^*\| \cdot \|x\| + L_1^2 \\
&= L_1 L_2 \|W - W^*\|_F \cdot \|x\| + L_1^2.
\end{aligned} \tag{38}$$

Noting that the input data $x \sim \mathcal{N}(0, I)$ and the moments of normal random variables are bounded, we can obtain

$$\begin{aligned}
t^{-1} \left(\mathbb{E}\left|\xi_{jl} \cdot u_j^\top x x^\top u_l\right|^t\right)^{1/t} &\leq C t^{-1} \left(\mathbb{E}\left|u_j^\top x x^\top u_l\right|^{2t}\right)^{1/2t} \\
&\leq C t^{-1} \left(\sqrt{\mathbb{E}\left|u_j^\top x\right|^{4t}} \cdot \sqrt{\mathbb{E}\left|x^\top u_l\right|^{4t}}\right)^{1/2t}.
\end{aligned} \tag{39}$$

From the equation (90) and (91) in Fu et al. (2020), we can directly get $\|G_i\|_{\psi_1} \leq C \triangleq \tau^2$. ■

Now we start the proof of Lemma 6.1 which is listed here again.

Lemma A.8 *(Lemma 6.1: The empirical Hessian is close to the expected Hessian). Assume Assumption 6.1 and 6.2 hold. The empirical Hessian converges uniformly to the expected Hessian. Namely, for $\forall \delta_2 > 0$, there exists $C_{\delta_2} > 0$, if $N \geq C_{\delta_2} \cdot dK \log dK \triangleq N_0(\delta_2)$, we have*

$$\mathbb{P}\left(\sup_{W \in \mathcal{B}(W_{\mathcal{H}^*}, R)} \left\|\nabla^2 f_N(W) - \nabla^2 f(W)\right\| \leq C_{\delta_2}\sqrt{\frac{\log N}{N}}\right) \geq 1 - \delta_2. \tag{40}$$

Proof *Inspired by Fu et al. (2020), we just need to verify the conditions in Mei et al. (2018). Similar to the analysis in (Fu et al., 2020; Mei et al., 2018), we also apply the covering number theory. Let N_ϵ be the $\epsilon-$covering number of the Euclidean ball $\mathcal{B}(W^*, R)$. From Lemma 5.2 in Vershynin (2010), it is known that $\log N_\epsilon \leq dK \log(3R/\epsilon)$. Let $\mathcal{W}_\epsilon = \{W_1, \cdots, W_{N_\epsilon}\}$ be the $\epsilon-$cover set with N_ϵ elements. For any $W \in \mathcal{B}(W^*, R)$, let $j(W) = \underset{j \in [N_\epsilon]}{\operatorname{argmin}} \|W - W_{j(W)}\|_F \leq \epsilon$ for all $W \in \mathcal{B}(W^*, R)$. For any $W \in \mathcal{B}(W^*, R)$, we have*

$$\begin{aligned}
\left\|\nabla^2 f_N(W) - \nabla^2 f(W)\right\| \leq & \frac{1}{N}\left\|\sum_{i=1}^N \left[\nabla^2 \ell\left(W; x_i\right) - \nabla^2 \ell\left(W_{j(W)}; x_i\right)\right]\right\| \\
& + \left\|\frac{1}{N}\sum_{i=1}^N \nabla^2 \ell\left(W_{j(W)}; x_i\right) - \mathbb{E}\left[\nabla^2 \ell\left(W_{j(W)}; x\right)\right]\right\| \\
& + \left\|\mathbb{E}\left[\nabla^2 \ell\left(W_{j(W)}; x\right)\right] - \mathbb{E}\left[\nabla^2 \ell(W; x)\right]\right\|.
\end{aligned} \tag{41}$$

Hence, we have

$$\mathbb{P}\left(\sup_{W \in \mathcal{B}(W^*, R)} \left\|\nabla^2 f_N(W) - \nabla^2 f(W)\right\| \geq t\right) \leq \mathbb{P}\left(A_t\right) + \mathbb{P}\left(B_t\right) + \mathbb{P}\left(C_t\right),$$

where the events A_t, B_t and C_t are defined as

$$A_t = \left\{\sup_{W \in \mathcal{B}(W^*, R)} \frac{1}{N}\left\|\sum_{i=1}^N \left[\nabla^2 \ell\left(W; x_i\right) - \nabla^2 \ell\left(W_{j(W)}; x_i\right)\right]\right\| \geq \frac{t}{3}\right\}, \tag{42}$$

$$B_t = \left\{\sup_{W \in \mathcal{W}_\epsilon} \left\|\frac{1}{N}\sum_{i=1}^N \nabla^2 \ell\left(W; x_i\right) - \mathbb{E}\left[\nabla^2 \ell(W; x)\right]\right\| \geq \frac{t}{3}\right\}, \tag{43}$$

$$C_t = \left\{\sup_{W \in \mathcal{B}(W^*, R)} \left\|\mathbb{E}\left[\nabla^2 \ell\left(W_{j(W)}; x\right)\right] - \mathbb{E}\left[\nabla^2 \ell(W; x)\right]\right\| \geq \frac{t}{3}\right\}. \tag{44}$$

Above all, we bound the terms $\mathbb{P}\left(A_t\right), \mathbb{P}\left(B_t\right)$, and $\mathbb{P}\left(C_t\right)$, separately.

1) Upper bound on $\mathbb{P}\left(B_t\right)$. Inspired by the proof of Lemma 3 in Fu et al. (2020), let $V_{1/4}$ be a $\left(\frac{1}{4}\right)$-cover of the ball $\mathcal{B}(0,1) = \left\{W \in \mathbb{R}^{d \times K} : \|W\|_{\mathrm{F}} = 1\right\}$, where $\log\left|V_{1/4}\right| \leq dK \log 12$. Following from Lemma A.6, we have

$$\left\|\frac{1}{N}\sum_{i=1}^N \nabla^2 \ell\left(W; x_i\right) - \mathbb{E}\left[\nabla^2 \ell(W; x)\right]\right\| \leq \sup_{v \in V_{1/4}} \left|\left\langle v, \left(\frac{1}{N}\sum_{i=1}^N \nabla^2 \ell\left(W; x_i\right) - \mathbb{E}\left[\nabla^2 \ell(W; x)\right]\right) v\right\rangle\right|.$$

Taking the union bound over $\mathcal{W}_\epsilon$ and $V_{1/4}$ yields

$$\begin{aligned}
\mathbb{P}(B_t) & \leq \mathbb{P}\left(\sup_{W \in \mathcal{W}_\epsilon, v \in V_{1/4}} \left|\frac{1}{N}\sum_{i=1}^N G_i\right| \geq \frac{t}{6}\right) \\
& \leq \exp\left(dK\left(\log\frac{3r}{\epsilon} + \log 12\right)\right) \sup_{W \in \mathcal{W}_\epsilon, v \in V_{1/4}} \mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^N G_i\right| \geq \frac{t}{6}\right).
\end{aligned} \tag{45}$$

where $G_i = \left\langle v, \left(\nabla^2 \ell\left(W; x_i\right) - \mathbb{E}\left[\nabla^2 \ell(W; x)\right]\right) v\right\rangle$ and $\mathbb{E}\left[G_i\right] = 0$. Let $a = \left[a_1^\top, \cdots, a_K^\top\right] \in \mathbb{R}^{dK}$.

Then we will show that $\|G_i\|_{\psi_1}$ is upper bounded, i.e., there exists some constant C such that

$$\|G_i\|_{\psi_1} \leq C :\equiv \tau^2.$$

Applying the Bernstein inequality for sub-exponential random variables to (45), by Theorem 9 in Mei et al. (2018), we have that for fixed $W \in \mathcal{W}_\epsilon, v \in V_{\frac{1}{4}}$,

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\left\langle v, \left(\nabla^2\ell\left(W;x_i\right) - \mathbb{E}\left[\nabla^2\ell(W;x)\right]\right)v\right\rangle\right| \geq \frac{t}{6}\right) \leq 2\exp\left(-C\cdot N\cdot\min\left(\frac{t^2}{\tau^4}, \frac{t}{\tau^2}\right)\right), \tag{46}$$

for some universal constant C. Combining (45) and (46), $\mathbb{P}(B_t)$ is upper bounded by

$$\mathbb{P}(B_t) \leq 2\exp\left(-C\cdot N\cdot\min\left(\frac{t^2}{\tau^4}, \frac{t}{\tau^2}\right) + dK\log\frac{3r}{\epsilon} + dK\log 12\right). \tag{47}$$

Above all, if

$$t > C\cdot\max\left\{\sqrt{\frac{\tau^4\left(dK\log\frac{36r}{\epsilon} + \log\frac{4}{\delta}\right)}{N}}, \frac{\tau^2\left(dK\log\frac{36r}{\epsilon} + \log\frac{4}{\delta}\right)}{N}\right\} \tag{48}$$

for some large enough constant C, we have $\mathbb{P}(B_t) \leq \frac{\delta}{2}$.

2) Upper bound on $\mathbb{P}(A_t)$ and $\mathbb{P}(C_t)$.

These two events will be bounded in a similar way. By the third derivative of the activation function is bounded in assumption 6.1, for $\forall W \neq W' \in \mathcal{B}\left(W^*, R\right)$, we have that

$$\left\|\nabla^2\ell(W;z) - \nabla^2\ell(W';z)\right\| \leq \sum_{j=1}^{K}\sum_{l=1}^{K}\left|\xi_{j,l}(W) - \xi_{j,l}(W')\right|\cdot\|xx^\top\| \tag{49}$$

$$\left|\xi_{j,l}(W) - \xi_{j,l}(W')\right| \leq \left(\max_k\left|T_{i,j,k}\right|\right)\cdot\|x\|\cdot\sqrt{K}\cdot\|W - W'\| \tag{50}$$

For $\forall j, l, k \in [K]$, we have

$$\begin{aligned}
\left|T_{j,l,k}\right| &\leq 2L_1 + L_1L_2 + L_3\left|\sum_{k=1}^{K}\phi\left(w_k^\top x\right) - \sum_{k=1}^{K}\phi\left(w_k^{*\top}x\right)\right| \\
&\leq 2L_1 + L_1L_2 + L_1L_3\sum_{k=1}^{K}\left|\left(w_k - w_k^*\right)^\top x\right| \\
&\leq 2L_1 + L_1L_2 + L_1L_3K\left\|W - W^*\right\|\cdot\|x\|.
\end{aligned} \tag{51}$$

If $\|W - W^*\| \leq \min\left\{\sqrt{\frac{(2L_1+L_1L_2)^2}{2L_1^2L_3^2dK}}, \frac{2+L_2}{KL_3}\right\}$, we have $\max_{j,l,k}\left|T_{j,l,k}\right| \leq (2L_1 + L_1L_2)(1 + \|x\|)$.

$$\begin{aligned}
\mathbb{E}\left[\sup_{W\neq W'\in\mathcal{B}(W^*,R)}\frac{\left\|\nabla^2\ell(W;x) - \nabla^2\ell\left(W';x\right)\right\|}{\left\|W - W'\right\|_{\mathrm{F}}}\right] &\leq \sqrt{K}\cdot K^2\cdot\mathbb{E}\left[\left(\max_{j,l,k}\left|T_{j,l,k}\right|\right)\cdot\|x\|\cdot\|xx^\top\|\right] \\
&\leq C\cdot dK^{3/2}.
\end{aligned} \tag{52}$$

For the event C_t which is a deterministic event, we have

$$\begin{aligned}
&\sup_{W\in\mathcal{B}(W^*,R)}\left\|\mathbb{E}\left[\nabla^2\ell\left(W_{j(W)};x\right)\right] - \mathbb{E}\left[\nabla^2\ell(W;x)\right]\right\| \\
&\leq \sup_{W\in\mathcal{B}(W^*,R)}\frac{\left\|\mathbb{E}\left[\nabla^2\ell\left(W_{j(W)};x\right)\right] - \mathbb{E}\left[\nabla^2\ell(W;x)\right]\right\|}{\|W - W_{j(W)}\|}\cdot\sup_{W\in\mathcal{B}(W^*,R)}\|W - W_{j(W)}\| \\
&\leq C\cdot dK^{3/2}\cdot\epsilon
\end{aligned} \tag{53}$$

by letting $W' = W_{j(W)}$. Thus if $t > C \cdot dK^{3/2} \cdot \epsilon$, C_t holds.

Similarly, we can bound the event A_t as below.

$$\begin{aligned}
&\mathbb{P}\left(\sup_{W\in\mathcal{B}(W^*,R)} \frac{1}{N}\left\|\sum_{i=1}^{n}\left[\nabla^2\ell\left(W;x_i\right)-\nabla^2\ell\left(W_{j(W)};x_i\right)\right]\right\| \geq \frac{t}{3}\right)\\
&\leq \frac{3}{t}\mathbb{E}\left[\sup_{W\in\mathcal{B}(W^*,R)}\left\|\frac{1}{N}\sum_{i=1}^{N}\left[\nabla^2\ell\left(W;x_i\right)-\nabla^2\ell\left(W_{j(W)};x_i\right)\right]\right\|\right]\\
&\leq \frac{3}{t}\mathbb{E}\left[\sup_{W\in\mathcal{B}(W^*,R)}\left\|\nabla^2\ell\left(W;x_i\right)-\nabla^2\ell\left(W_{j(W)};x_i\right)\right\|\right]\\
&\leq \frac{3}{t}\mathbb{E}\left[\sup_{W\in\mathcal{B}(W^*,R)}\frac{\left\|\nabla^2\ell\left(W;x_i\right)-\nabla^2\ell\left(W_{j(W)};x_i\right)\right\|}{\left\|W-W_{j(W)}\right\|_{\mathrm{F}}}\right]\cdot\sup_{W\in\mathcal{B}(W^*,R)}\left\|W-W_{j(W)}\right\|_{\mathrm{F}}\\
&\leq \frac{C\cdot dK^{3/2}\cdot\epsilon}{t}
\end{aligned} \tag{54}$$

where the first inequality follows from the Markov inequality. Thus, taking $t \geq \frac{C\cdot dK^{3/2}\cdot\epsilon}{\delta}$ ensures that $\mathbb{P}\left(A_t\right) \leq \frac{\delta}{2}$.

3) Final step. By choosing $\epsilon = \frac{\delta\tau^2}{dK^{3/2}\cdot NdK}$, we have

$$t > \tau^2\cdot\max\left\{\frac{1}{NdK},\, C\sqrt{\frac{\tau^4\left(dK\log\frac{36r}{\epsilon}+\log\frac{4}{\delta}\right)}{N}},\, C\frac{\tau^2\left(dK\log\frac{36r}{\epsilon}+\log\frac{4}{\delta}\right)}{N}\right\}. \tag{55}$$

There exists C_δ as log as $N \geq C_\delta dK\log dK$, we have

Thus there exists C_δ, for $\forall N \geq C_\delta\cdot dK\log dK$, we can obtain that

$$\mathbb{P}\left(\sup_{W\in\mathcal{B}(W^*,R)}\left\|\nabla^2 f_N(W)-\nabla^2 f(W)\right\| \geq \tau^2\sqrt{\frac{C_\delta dK\log N}{N}}\right)\leq\delta. \tag{56}$$

■

A.4 THE PROOF OF LEMMA 6.2

Firstly, we prove Hessian smoothness of expected loss to extend the convexity from the point $\mathcal{H}^*$ to the region $\mathcal{B}(W_{\mathcal{H}^*}, R_0)$.

A.4.1 HESSIAN SMOOTHNESS OF EXPECTED LOSS: THE PROOF OF LEMMA A.9

Lemma A.9 *(Hessian smoothness of expected loss). Assume $W \in \mathcal{B}(W_{\mathcal{H}^*}, R_0)$ where $R_0 = \sqrt{\frac{(2L_1+L_1L_2)^2}{2L_1^2L_3^2dK}}$. Then*

$$\left\|\nabla^2 f(W)-\nabla^2 f(W_{\mathcal{H}^*})\right\| \leq 4\left(2L_1+L_1L_2\right)^2K^{\frac{7}{2}}\cdot\left\|W-W_{\mathcal{H}^*}\right\|_F. \tag{57}$$

Proof *Denote $W_{\mathcal{H}^*}$ as W^* for the sake of statement. Let $\Delta = \nabla^2\mathbb{E}\left[\ell(W;x)\right]-\nabla^2\mathbb{E}\left[\ell\left(W';x\right)\right]$. Denote the (j,l)-th block of Δ as $\Delta_{j,l}\in\mathbb{R}^{d\times d}$ for $\forall(j,l)\in[K]\times[K]$. Let $a = \left[a_1^\top,\cdots,a_K^\top\right]^\top \in \mathbb{R}^{dK}$. By the definition of spectral norm,*

$$\left\|\nabla^2 f(W)-\nabla^2 f(W^*)\right\| = \max_{\|a\|=1} a^\top\left(\nabla^2 f(W)-\nabla^2 f(W^*)\right)a = \max_{\|a\|=1}\sum_{j=1}^{K}\sum_{l=1}^{K}a_j^\top\Delta_{j,l}a_l. \tag{58}$$

Denoting $\xi_{j,l}(W) \triangleq \frac{\partial^2\mathbb{E}[\ell(W;x)]}{\partial w_j\partial w_l}$ and $\xi_{j,l}(W^*) \triangleq \frac{\partial^2\mathbb{E}[\ell(W^*;x)]}{\partial w_j^*\partial w_l^*}$, we have

$$\Delta_{j,l} = \frac{\partial^2\mathbb{E}\left[\ell(W;x)\right]}{\partial w_j\partial w_l}-\frac{\partial^2\mathbb{E}\left[\ell(W^*;x)\right]}{\partial w_j^*\partial w_l^*} = \mathbb{E}\left[\left(\xi_{j,l}(W)-\xi_{j,l}(W^*)\right)\cdot xx^\top\right] \tag{59}$$

where

$$\xi_{j,l}(W) = \begin{cases} \phi'\left(w_j^\top x\right)\phi'\left(w_l^\top x\right) & \text{for } j \neq l, \\ \left(\sum_{k=1}^K \phi\left(w_k^\top x\right) - y\right)\phi''\left(w_j^\top x\right) + \left(\phi'\left(w_j^\top x\right)\right)^2 & \text{for } j = l. \end{cases} \tag{60}$$

By the mean value theorem, we have

$$\xi_{j,l}\left(W\right) = \xi_{j,l}\left(W^*\right) + \sum_{k=1}^K \left\langle \frac{\partial \xi_{j,l}\left(\widetilde{W}\right)}{\partial \widetilde{w}_k}, w_k - w_k^* \right\rangle \tag{61}$$

where $\widetilde{W} = \eta W + (1-\eta)W^*$. *Then we can obtain*

$$\Delta_{j,l} = \mathbb{E}\left[\left(\sum_{k=1}^K \left\langle \frac{\partial \xi_{j,l}\left(\widetilde{W}\right)}{\partial \widetilde{w}_k}, w_k - w_k^* \right\rangle\right) xx^\top\right]. \tag{62}$$

Denote $\frac{\partial \xi_{j,l}(W)}{\partial w_k} \triangleq T_{j,l,k} \cdot x$, *where* $T_{j,l,k} \in \mathbb{R}$ *is a scalar. From (60) we can further calculate* $\frac{\partial \xi_{j,l}(W)}{\partial w_k}$ *as follows.*

$$\frac{\partial \xi_{j,l}\left(W\right)}{\partial w_k} \triangleq T_{j,l,k} \cdot x = \begin{cases} \phi''\left(w_j^\top x\right)\phi'\left(w_l^\top x\right)x & \text{for } k = j, \\ \phi'\left(w_j^\top x\right)\phi''\left(w_l^\top x\right)x & \text{for } k = l, \\ 0 & \text{for } k \neq j \text{ and } k \neq l. \end{cases} \tag{63}$$

and

$$\begin{aligned} &\frac{\partial \xi_{j,j}\left(W\right)}{\partial w_k} \triangleq T_{j,j,k} \cdot x \\ &= \begin{cases} \left[2\phi'\left(w_j^\top x\right) + \phi'\left(w_j^\top x\right)\cdot\phi''\left(w_j^\top x\right) + \left(\sum_{k=1}^K \phi\left(w_k^\top x\right) - y\right)\cdot\phi'''\left(w_j^\top x\right)\right]x & \text{for } k = j, \\ \phi'\left(w_k^\top x\right)\cdot\phi''\left(w_j^\top x\right)x & \text{for } k \neq j. \end{cases} \end{aligned} \tag{64}$$

Since the first derivative, second derivative and third derivative of the activation function are bounded, i.e, $|\phi'(\cdot)| \leq L_1$, $|\phi''(\cdot)| \leq L_2$ *and* $|\phi'''(\cdot)| \leq L_3$. *It's relatively easy to get*

$$|T_{j,l,k}| \leq L_1 L_2 \tag{65}$$

and

$$|T_{j,j,k}| \leq 2L_1 + L_1 L_2 + L_3 \left|\sum_{k=1}^K \phi\left(w_k^\top x\right) - \sum_{k=1}^K \phi\left(w_k^{*\top} x\right)\right|. \tag{66}$$

Then for $\forall j, l, k \in [K]$,

$$|T_{j,l,k}| \leq 2L_1 + L_1 L_2 + L_3 \left|\sum_{k=1}^K \phi\left(w_k^\top x\right) - \sum_{k=1}^K \phi\left(w_k^{*\top} x\right)\right|. \tag{67}$$

To sum up, by the equation (33-34) in Fu et al. (2020), we can obtain that

$$\left\|\nabla^2 f(W) - \nabla^2 f(W^*)\right\| \leq \max_{\|a\|=1} \sum_{j=1}^K \sum_{l=1}^K \sqrt{\sum_{k=1}^K \mathbb{E}\left[T_{j,l,k}^2\right]} \cdot \sqrt{\sum_{k=1}^K \|w_k - w_k^*\|^2 \cdot \|a_j\|^2 \cdot \|a_l\|^2}. \tag{68}$$

Now we focus on the upper bound of $\mathbb{E}\left[T_{j,l,k}^2\right]$. *Noticing that*

$$\begin{aligned} \left|\sum_{k=1}^K \phi\left(w_k^\top x\right) - \sum_{k=1}^K \phi\left(w_k^{*\top} x\right)\right| &\leq \sum_{k=1}^K \left(\left|\phi\left(w_k^\top x\right) - \phi\left(w_k^{*\top} x\right)\right|\right) \\ &\leq L_1 \sum_{k=1}^K \left|\left(w_k - w_k^*\right)^\top x\right|, \end{aligned} \tag{69}$$

we can obtain

$$\begin{aligned}\mathbb{E}\left[T_{j,l,k}^2\right] &\leq 2\left(2L_1 + L_1L_2\right)^2 + 4L_1^2L_3^2K\|W - W^*\|_F^2 \cdot \mathbb{E}[x^\top x] \\ &= (2(2L_1 + L_1L_2)^2 + 4L_1^2L_3^2dK\|W - W^*\|_F^2 \end{aligned} \tag{70}$$

by $\|w_k - w_k^*\| \leq \|W - W^*\|_F$ *and* $\mathbb{E}[x^\top x] = d$. *Choosing* $\|W - W^*\|_F \leq \sqrt{\frac{(2L_1+L_1L_2)^2}{2L_1^2L_3^2dK}}$, *we have* $\mathbb{E}\left[T_{j,l,k}^2\right] \leq 4\left(2L_1 + L_1L_2\right)^2$. *Further we can obtain*

$$\begin{aligned}\left\|\nabla^2 f(W) - \nabla^2 f(W^*)\right\| &\leq 4\left(2L_1 + L_1L_2\right)^2 K^{\frac{5}{2}} \cdot \|W - W^*\|_F \cdot \max_{\|a\|=1} \sum_{j=1}^{K}\sum_{l=1}^{K} \|a_j\|\|a_l\| \\ &\leq 4\left(2L_1 + L_1L_2\right)^2 K^{\frac{7}{2}} \cdot \|W - W^*\|_F . \end{aligned} \tag{71}$$

■

A.4.2 Local uniform strong convexity and smoothness of expected loss: The proof of Lemma A.10

Lemma A.10 *(Local strong convexity and smoothness of expected loss). For the population loss* $f(\cdot)$ *and* $\forall W \in \mathcal{B}\left(W_{\mathcal{H}^*}, \sqrt{\frac{(2L_1+L_1L_2)^2}{2L_1^2L_3^2dK}}\right)$, *we have*

$$l_f \cdot I \preceq \nabla^2 f(W) \preceq L_f \cdot I \tag{72}$$

where $l_f = -4\left(2L_1 + L_1L_2\right)^2 K^{\frac{7}{2}} \cdot \|W - W_{\mathcal{H}^*}\|_F$ *and* $L_f = 4\left(2L_1 + L_1L_2\right)^2 K^{\frac{7}{2}} \cdot \|W - W_{\mathcal{H}^*}\|_F + K\sigma_1^2$.

Proof *By Lemma A.9 in the appendix, if* $\|W - W_{\mathcal{H}^*}\|_F \leq \sqrt{\frac{(2L_1+L_1L_2)^2}{2L_1^2L_3^2dK}}$, *we have the second-order smoothness near the ground truth*

$$\left\|\nabla^2 f(W) - \nabla^2 f(W_{\mathcal{H}^*})\right\| \leq 4\left(2L_1 + L_1L_2\right)^2 K^{\frac{7}{2}} \cdot \|W - W_{\mathcal{H}^*}\|_F . \tag{73}$$

From Lemma D.6 and D.7 in Zhong et al. (2017), we get the upper and lower bounds on $\nabla^2 f(W_{\mathcal{H}^*})$

$$\rho\left(\sigma_{\min}(W_{\mathcal{H}^*})\right) / \left(\kappa^2\lambda\right) \cdot I \preceq \nabla^2 f(W_{\mathcal{H}^*}) \preceq K\left(\sigma_{\max}(W_{\mathcal{H}^*})\right)^2 \cdot I \tag{74}$$

where $\kappa = \frac{\sigma_{\max}(W_{\mathcal{H}^*})}{\sigma_{\min}(W_{\mathcal{H}^*})}$, $\lambda = \frac{\prod_{k=1}^{K}\sigma_k(W_{\mathcal{H}^*})}{(\sigma_{\min}(W_{\mathcal{H}^*}))^K}$.

By the equation (73), if $\|W - W_{\mathcal{H}^*}\|_F \leq \sqrt{\frac{(2L_1+L_1L_2)^2}{2L_1^2L_3^2dK}}$, *we have*

$$\begin{aligned}\sigma_{\min}\left(\nabla^2 f(W)\right) &\geq \sigma_{\min}\left(\nabla^2 f(W_{\mathcal{H}^*})\right) - \|\nabla^2 f(W) - \nabla^2 f(W_{\mathcal{H}^*})\| \\ &\geq \rho\left(\sigma_K\right) / \left(\kappa^2\lambda\right) - 4\left(2L_1 + L_1L_2\right)^2 K^{\frac{7}{2}} \cdot \|W - W_{\mathcal{H}^*}\|_F \\ &= -4\left(2L_1 + L_1L_2\right)^2 K^{\frac{7}{2}} \cdot \|W - W_{\mathcal{H}^*}\|_F . \end{aligned} \tag{75}$$

By the triangle inequality, the spectral norm of $\nabla^2 f(W)$, *has the following upper bound*

$$\begin{aligned}\|\nabla^2 f(W)\| &\leq \|\nabla^2 f(W) - \nabla^2 f(W_{\mathcal{H}^*})\| + \|\nabla^2 f(W_{\mathcal{H}^*})\| \\ &\leq 4\left(2L_1 + L_1L_2\right)^2 K^{\frac{7}{2}} \cdot \|W - W_{\mathcal{H}^*}\|_F + K\sigma_1^2. \end{aligned} \tag{76}$$

■

A.4.3 Local uniform strong convexity and smoothness of empirical loss: The proof of Lemma 6.2

The square term of our method can strengthen the local convexity to the local strong convexity in Lemma 6.2, which is convenient to analyze and estimate the recovery degree relative to $W_{\mathcal{H}^*}$.

Proof *For the smallest singular value of $\nabla^2 f_N(W)$, we have*

$$\sigma_{\min}\left(\nabla^2 f_N(W)\right) \geq \sigma_{\min}\left(\nabla^2 f(W)\right) - \|\nabla^2 f_N(W) - \nabla^2 f(W)\|. \tag{77}$$

Noting the definition $F_N(W) \triangleq f_N(W) + \lambda_N \|W - \widehat{W}_N\|_F^2$, we have

$$\nabla^2 F_N(W) \succeq \left(\sigma_{\min}\left(\nabla^2 f(W)\right) - \|\nabla^2 f_N(W) - \nabla^2 f(W)\| + 2\lambda_N\right) \cdot I. \tag{78}$$

Noting that

$$\|\nabla^2 f_N(W)\| \leq \|\nabla^2 f_N(W) - \nabla^2 f(W)\| + \|\nabla^2 f(W)\|, \tag{79}$$

we can obtain

$$\|\nabla^2 F_N(W)\| = \|\nabla^2 f_N(W) + 2\lambda_N \cdot I\| \leq \|\nabla^2 f_N(W)\| + 2\lambda_N. \tag{80}$$

We first analyze the lower bound of $\nabla^2 f_N(W)$. Combining (77), by Lemma 6.1, we have

$$\begin{aligned}
&\sigma_{\min}\left(\nabla^2 f(W)\right) - \|\nabla^2 f_N(W) - \nabla^2 f(W)\| \geq l_f - C_{\delta_2}\sqrt{\frac{\log N}{N}} \\
&\geq -4\left(2L_1 + L_1 L_2\right)^2 K^{\frac{7}{2}} \cdot \|W - W_{\mathcal{H}^*}\|_F - C_{\delta_2}\sqrt{\frac{\log N}{N}}
\end{aligned} \tag{81}$$

with probability at least $1 - \delta_2$. Thus if we choose $\lambda_N > C_{\delta_2}\sqrt{\frac{\log N}{N}}$, there exists $l_N \triangleq \lambda_N - C_{\delta_2}\sqrt{\frac{\log N}{N}} > 0$ such that

$$\nabla^2 F_N(W) \succeq l_N \cdot I \tag{82}$$

with probability at least $1 - \delta_2$.

Then we analyze the upper bound of $\nabla^2 f_N(W)$. Combining (80), by Lemma 6.1, we have

$$\|\nabla^2 f_N(W) - \nabla^2 f(W)\| + \|\nabla^2 f(W)\| \leq C_{\delta_2}\sqrt{\frac{\log N}{N}} + L_f \tag{83}$$

with probability at least $1 - \delta_2$. By denoting $L_N \triangleq C_{\delta_2}\sqrt{\frac{\log N}{N}} + L_f + 2\lambda_N$,

$$\|\nabla^2 F_N(W)\| \leq \|\nabla^2 f_N(W)\| + 2\lambda_N \leq L_N. \tag{84}$$

with probability at least $1 - \delta_2$.

Above all, $l_N \cdot I \preceq \nabla^2 F_N(W) \preceq L_N \cdot I$ with probability at least $1 - \delta_2$. ■

A.5 THE PROOF OF LEMMA A.12

Firstly prove a key lemma.

Lemma A.11 *For $u \in \mathcal{B}(0,1) = \left\{W \in \mathbb{R}^{d\times K} : \|W\|_{\mathrm{F}} = 1\right\}$, $\|\langle \nabla \ell(W), u\rangle\|_{\psi_2}$ is upper bounded, i.e., there exists some constant C such that*

$$\|\langle \nabla \ell(W), u\rangle\|_{\psi_2} \leq C. \tag{85}$$

Proof *Note $\nabla \ell(W; x) = [\frac{\partial \ell(W)}{\partial W_1}, \cdots, \frac{\partial \ell(W)}{\partial W_K}]$ where $\frac{\partial \ell(W)}{\partial W_j} = \left(\sum_{k=1}^K \phi\left(w_k^\top x\right) - y\right)\phi'\left(w_j^\top x\right)x$, $\forall j \in [K]$. Hence*

$$\langle \nabla \ell(W; x), u\rangle = \left(\sum_{k=1}^K \phi\left(w_k^\top x\right) - y\right) \cdot \sum_{j=1}^K \phi'\left(w_j^\top x\right)\left(u_j^\top x\right) \tag{86}$$

By the definition of sub-gaussian norm and denoting $\zeta_j = \left(\sum_{k=1}^K \phi\left(w_k^\top x\right) - y\right) \cdot \phi'\left(w_j^\top x\right)$, we have

$$\begin{aligned}
\|\langle \nabla \ell(W), u\rangle\|_{\psi_2} &\leq \sum_{j=1}^K \left\|\zeta_j \cdot \left(u_j^\top x\right)\right\|_{\psi_2} = \sum_{j=1}^K \sup_{t\geq 1} \frac{1}{\sqrt{t}}\left(\mathbb{E}\left|\zeta_j \cdot \left(u_j^\top x\right)\right|^t\right)^{1/t} \\
&\leq \sum_{j=1}^K \sup_{t\geq 1} \frac{1}{\sqrt{t}}\left(\sqrt{\mathbb{E}\left|\zeta_j\right|^{2t}} \cdot \sqrt{\mathbb{E}\left|\left(u_j^\top x\right)\right|^{2t}}\right)^{1/t}.
\end{aligned} \tag{87}$$

and

$$|\zeta_j| = \left| \left(\sum_{k=1}^{K} \phi\left(w_k^\top x\right) - \sum_{k=1}^{K} \phi\left(w_k^{*\top} x\right) \right) \cdot \phi'\left(w_j^\top x\right) \right| \leq L_1^2 \cdot K \|W - W^*\|_F \cdot \|x\| \tag{88}$$

combining (87) and (88), we have $\|\langle \nabla \ell(W), u \rangle\|_{\psi_2} \leq C \cdot K$.

Lemma A.12 *(The empirical gradient is close to the expected gradient). The empirical gradient converges uniformly to the expected gradient. Namely, for* $\forall \delta_3 > 0$*, if* $N \geq C_{\delta_3} \cdot dK \log dK \triangleq N_0(\delta_3)$*, we have*

$$\mathbb{P}\left(\sup_{W \in \mathcal{B}(W_{\mathcal{H}^*}, R)} \|\nabla f_N(W) - \nabla f(W)\|_F \leq C_{\delta_3} \sqrt{\frac{dK \log N}{N}} \right) \geq 1 - \delta_3. \tag{89}$$

Now we start the proof of A.12.

Proof *Combining Lemma A.11, similar to the proof of Lemma A.9, we can get the following concentration inequality about the gradient of loss function.* ■

A.6 THE PROOF OF THEOREM 6.2

The proof is based on the above strong local convexity, we use the theory of convex optimization to analyze the error upper bounds of the estimates obtained by the two weighted regularization methods.

Proof ***Step 1.*** *The error bound of the estimates of sparse recovery optimization problem.*

Denote $W_{N,1} = W_{\mathcal{H}^*} + M_{N,1}$*, then we just have to focus on the bounds of* $M_{N,1}$*. By the optimality of* $W_{N,1}$*, we have*

$$\begin{aligned}
0 &\geq J_N(W_{\mathcal{H}^*} + M_{N,1}) - J_N(W_{\mathcal{H}^*}) \\
&= f_N(W_{\mathcal{H}^*} + M_{N,1}) + \lambda_N \|W_{\mathcal{H}^*} + M_{N,1} - \widehat{W}_N\|_F^2 - \left(f_N(W_{\mathcal{H}^*}) + \lambda_N \|W_{\mathcal{H}^*} - \widehat{W}_N\|_F^2 \right) \\
&\quad + \gamma_N \|W_{\mathcal{H}^*} + M_N\|_{\widehat{W}_N^w, 1} - \gamma_N \|W_{\mathcal{H}^*}\|_{\widehat{W}_N^w, 1}.
\end{aligned} \tag{90}$$

By the definition of $\|\cdot\|_{\widehat{W}_N^w, 1}$ *and absolute value inequality, we can get*

$$\begin{aligned}
&\|W_{\mathcal{H}^*} + M_{N,1}\|_{\widehat{W}_N^w, 1} - \|W_{\mathcal{H}^*}\|_{\widehat{W}_N^w, 1} = \sum_{s=1}^{d} \sum_{t=1}^{K} \frac{|W_{\mathcal{H}^*}(s,t) + M_{N,1}(s,t)| - |W_{\mathcal{H}^*}(s,t)|}{\left|\widehat{W}_N^w(s,t)\right|} \\
&= \sum_{(s,t) \in \mathcal{A}_{\mathcal{H}^*}} \frac{|M_{N,1}(s,t)|}{\left|\widehat{W}_N(s,t)\right|} + \sum_{(s,t) \in \mathcal{A}_{\mathcal{H}^*}^C} \frac{|W_{\mathcal{H}^*}(s,t) + M_{N,1}(s,t)| - |W_{\mathcal{H}^*}(s,t)|}{\left|\widehat{W}_N^w(s,t)\right|} \\
&\geq \sum_{(s,t) \in \mathcal{A}_{\mathcal{H}^*}^C} \frac{|W_{\mathcal{H}^*}(s,t) + M_{N,1}(s,t)| - |W_{\mathcal{H}^*}(s,t)|}{\left|\widehat{W}_N^w(s,t)\right|} \\
&\geq \sum_{(s,t) \in \mathcal{A}_{\mathcal{H}^*}^C} -\frac{|M_{N,1}(s,t)|}{\left|\widehat{W}_N^w(s,t)\right|} \\
&\geq -C_3 \|M_N\|_F .
\end{aligned} \tag{91}$$

where the last inequality follows from the equivalence of norms.

Denote $F_N(W) \triangleq f_N(W) + \lambda_N \|W - \widehat{W}_N\|_F^2$*. By substituting (91) into (90) and the mean value theorem, there exists* $W' = \xi W_{\mathcal{H}^*} + (1-\xi) W_N \in \mathcal{B}(W_{\mathcal{H}^*}, R)$ *for* $\xi \in [0,1]$ *such that*

$$\begin{aligned}
0 &\geq F_N(W_{\mathcal{H}^*} + M_N) - F_N(W_{\mathcal{H}^*}) - \gamma_N \cdot C_3 \|M_{N,1}\|_F \\
&= \langle \nabla F_N(W_{\mathcal{H}^*}), M_{N,1} \rangle + \frac{1}{2} \mathrm{vec}(M_{N,1})^T \nabla^2 F_N(W') \mathrm{vec}(M_{N,1}) - C_3 \cdot \gamma_N \|M_{N,1}\|_F \\
&= \langle \nabla f_N(W_{\mathcal{H}^*}), M_{N,1} \rangle + 2\lambda_N \langle W_{\mathcal{H}^*} - \widehat{W}_N, M_{N,1} \rangle + \frac{1}{2} \mathrm{vec}(M_{N,1})^T \nabla^2 F_N(W') \mathrm{vec}(M_{N,1}) \\
&\quad - C_3 \cdot \gamma_N \|M_{N,1}\|_F .
\end{aligned} \tag{92}$$

Note $W_{\mathcal{H}^*}$ is the global optimal point, then $\nabla f(W_{\mathcal{H}^*}) = 0$. By Lemma A.12, we can obtain for $N \geq N_0(\delta_3)$

$$\|\nabla f_N(W_{\mathcal{H}^*})\|_F \leq C_{\delta_3}\sqrt{\frac{\log N}{N}} \tag{93}$$

with probability at least $1-\delta_3$. From Lemma 6.2, by choosing $\lambda_N > C_{\delta_2}\sqrt{\frac{\log N}{N}}$, the Hessian matrix of $F_N(\cdot)$ at point $W' \in \mathcal{B}(W_{\mathcal{H}^*}, R)$ is positive definite, i.e, $\nabla^2 F_N(W') \succeq l_N \cdot I$ with probability at least $1-\delta_2$.

By Lemma A.2, for $N \geq \max\{N_0(\delta_2), N_0(\delta_3)\}$, we have

$$\begin{aligned}
0 &\geq J_N(W_{\mathcal{H}^*} + M_{N,1}) - J_N(W_{\mathcal{H}^*}) \\
&= \frac{l_N}{2}\|M_{N,1}\|_F^2 - \left(\|\nabla f_N(W_{\mathcal{H}^*})\|_F + 2\lambda_N\|W_{\mathcal{H}^*} - \widehat{W}_N\|_F\right)\|M_{N,1}\|_F - C_3\cdot\gamma_N\|M_{N,1}\|_F \\
&\geq \frac{l_N}{2}\|M_{N,1}\|_F\left(\|M_{N,1}\|_F - \frac{2}{l_N}\|\nabla f_N(W_{\mathcal{H}^*})\|_F - \frac{4\lambda_N}{l_N}\|W_{\mathcal{H}^*} - \widehat{W}_N\|_F - \frac{2C_3}{l_N}\gamma_N\right)
\end{aligned} \tag{94}$$

with probability at least $1-(\delta_2+\delta_3)$. By Lemma A.2 and Theorem 6.1, for $N \geq \max\{N_0(\delta_1), N_0(\delta_2), N_0(\delta_3)\} \triangleq N_0(\delta)$, further we can obtain

$$\begin{aligned}
\|W_{N,1} - W_{\mathcal{H}^*}\|_F &\leq \frac{2}{l_N}\|\nabla f_N(W_{\mathcal{H}^*})\|_F + \frac{4\lambda_N}{l_N}\|W_{\mathcal{H}^*} - \widehat{W}_N\|_F + \frac{2C_3}{l_N}\gamma_N \\
&\leq C_{\delta_2}\frac{2}{l_N}\sqrt{\frac{\log N}{N}} + C_{\delta_1}\frac{4\lambda_N}{l_N}\left(\frac{\log N}{\sqrt{N}}\right)^{1/\nu} + \frac{2C_3}{l_N}\gamma_N
\end{aligned} \tag{95}$$

with probability at least $1-(\delta_1+\delta_2+\delta_3) \triangleq \delta$.

Step 2. *The error bound of the estimates of low-rank recovery optimization problem.*

Denote $W_{N,*} = W_{\mathcal{H}^*} + M_{N,*}$, then we just have to focus on the bounds of $M_{N,*}$. By the optimality of $W_{N,*}$, we have

$$\begin{aligned}
0 &\geq J_N(W_{\mathcal{H}^*} + M_{N,*}) - J_N(W_{\mathcal{H}^*}) \\
&= f_N(W_{\mathcal{H}^*} + M_{N,*}) + \lambda_N\|W_{\mathcal{H}^*} + M_{N,*} - \widehat{W}_N\|_F^2 - \left(f_N(W_{\mathcal{H}^*}) + \lambda_N\|W_{\mathcal{H}^*} - \widehat{W}_N\|_F^2\right) \\
&\quad + \gamma_N\|W_{\mathcal{H}^*} + M_{N,*}\|_{\widehat{W}_N^w,*} - \gamma_N\|W_{\mathcal{H}^*}\|_{\widehat{W}_N^w,*}.
\end{aligned} \tag{96}$$

For the nuclear norm regularization part, we can obtain

$$\begin{aligned}
\|W_{N,*}\|_{\widehat{W}_N^w,*} - \|W_{\mathcal{H}^*}\|_{\widehat{W}_N^w,*} &= \sum_{l=1}^{K}\frac{\sigma_l(W_{N,*}) - \sigma_l(W_{\mathcal{H}^*})}{\sigma_l(\widehat{W}_N^w)} \\
= \sum_{l\in\mathcal{B}_{\mathcal{H}^*}}\frac{\sigma_l(W_{N,*})}{\sigma_l(\widehat{W}_N^w)} + \sum_{l\in\mathcal{B}_{\mathcal{H}^*}^C}\frac{\sigma_l(W_N) - \sigma_l(W_{\mathcal{H}^*})}{\sigma_l(\widehat{W}_N^w)} &\geq -C_4\|M_N\|_F.
\end{aligned} \tag{97}$$

Similar to Step 1, for $N \geq N_0(\delta)$. we can also get the error bound of $W_{N,*}$, i.e.,

$$\begin{aligned}
\|W_{N,*} - W_{\mathcal{H}^*}\|_F &\leq \frac{2}{l_N}\|\nabla f_N(W_{\mathcal{H}^*})\|_F + \frac{4\lambda_N}{l_N}\|W_{\mathcal{H}^*} - \widehat{W}_N\|_F + \frac{2C_4}{l_N}\gamma_N \\
&\leq C_{\delta_2}\frac{2}{l_N}\sqrt{\frac{\log N}{N}} + C_{\delta_1}\frac{4\lambda_N}{l_N}\left(\frac{\log N}{\sqrt{N}}\right)^{1/\nu} + \frac{2C_4}{l_N}\gamma_N
\end{aligned} \tag{98}$$

with probability at least $1-\delta$. ■

A.7 THE PROOF OF THEOREM 6.3

The proof is based on the optimality of $W_{N,1}$, $W_{N,*}$ and the selection of regularization coefficient γ_N in Assumption 6.3.

Proof ***Step 1.*** *Firstly, we prove the sparse consistency of the estimates* $W_{N,1}$.

Denote $W_{N,1} = W_{\mathcal{H}^*} + M_{N,1}$ *and* $\overline{W}_{N,1} = W_{\mathcal{H}^*} + \overline{M}_{N,1}$, *where*

$$\overline{M}_{N,1}(s,t) = M_{N,1}(s,t), \textit{if } (s,t) \in \mathcal{A}^C_{\mathcal{H}^*} \tag{99}$$

$$\overline{M}_{N,1}(s,t) = 0, \qquad \textit{if } (s,t) \in \mathcal{A}_{\mathcal{H}^*}. \tag{100}$$

Suppose that there exists some $(s_0,t_0) \in \mathcal{A}_{\mathcal{H}^*}$ *such that* $M_N(s_0,t_0) \neq 0$.

By the optimality of $W_{N,1}$, *we can obtain*

$$\begin{aligned} 0 &\geq J_N(W_{\mathcal{H}^*} + M_{N,1}) - J_N(W_{\mathcal{H}^*} + \overline{M}_{N,1}) \\ &= F_N(W_{\mathcal{H}^*} + M_{N,1}) - F_N(W_{\mathcal{H}^*} + \overline{M}_{N,1}) \\ &\quad + \gamma_N \|W_{\mathcal{H}^*} + M_{N,1}\|_{\widehat{W}^w_N,1} - \gamma_N \|W_{\mathcal{H}^*} + \overline{M}_{N,1}\|_{\widehat{W}^w_N,1}. \end{aligned} \tag{101}$$

Note that

$$\begin{aligned} &\|W_{\mathcal{H}^*} + M_{N,1}\|_{\widehat{W}^w_N,1} - \|W_{\mathcal{H}^*} + \overline{M}_{N,1}\|_{\widehat{W}^w_N,1} \\ &= \sum_{s=1}^{d}\sum_{t=1}^{K} \frac{|W_{\mathcal{H}^*}(s,t) + M_{N,1}(s,t)| - \left|W_{\mathcal{H}^*}(s,t) + \overline{M}_{N,1}(s,t)\right|}{\left|\widehat{W}^w_N(s,t)\right|} \\ &= \sum_{(s,t)\in\mathcal{A}_{\mathcal{H}^*}} \frac{|M_{N,1}(s,t)|}{\left|\widehat{W}^w_N(s,t)\right|} + \sum_{(s,t)\in\mathcal{A}^C_{\mathcal{H}^*}} \frac{|W_{\mathcal{H}^*}(s,t) + M_{N,1}(s,t)| - \left|W_{\mathcal{H}^*}(s,t) + \overline{M}_{N,1}\right|}{\left|\widehat{W}^w_N(s,t)\right|} \\ &= \sum_{(s,t)\in\mathcal{A}_{\mathcal{H}^*}} \frac{|M_{N,1}(s,t)|}{\left|\widehat{W}^w_N(s,t)\right|} = \frac{|M_{N,1}(s_0,t_0)|}{\left|\widehat{W}^w_N(s_0,t_0)\right|} \end{aligned} \tag{102}$$

where the second equation comes from $W_{\mathcal{H}^*}(s,t) = 0,\ \forall (s,t) \in \mathcal{A}_{\mathcal{H}^*}$ *and the third equation comes from* $\overline{M}_{N,1}(s,t) = M_{N,1}(s,t),\ \forall (s,t) \in \mathcal{A}^C_{\mathcal{H}^*}$. *Further we can get*

$$\gamma_N \frac{1}{|\widehat{W}^w_N(s_0,t_0)|} |M_{N,1}(s_0,t_0)| \leq F_N(W_{\mathcal{H}^*} + \overline{M}_{N,1}) - F_N(W_{\mathcal{H}^*} + M_{N,1}) \tag{103}$$

By $l_N \cdot I \preceq \nabla^2 F_N(W) \preceq L_N \cdot I$, *we can obtain*

$$F_N(\overline{W}_{N,1}) - F_N(W_{N,1}) = \langle \nabla F_N(\widetilde{W}_{N,1}), \overline{W}_{N,1} - W_{N,1} \rangle \leq \|\nabla F_N(\widetilde{W}_{N,1})\|_F \cdot \|\overline{W}_{N,1} - W_{N,1}\|_F \tag{104}$$

and

$$\mathrm{vec}(\nabla F_N(\widetilde{W}_{N,1})) - \mathrm{vec}(\nabla F_N(W_{\mathcal{H}^*})) = \nabla^2 F_N(\widetilde{W}'_N) \cdot \mathrm{vec}(\widetilde{W}_{N,1} - W_{\mathcal{H}^*}) \tag{105}$$

where $\widetilde{W}_{N,1} = \xi W_{N,1} + (1-\xi)\overline{W}_{N,1},\ \xi \in [0,1]$ *and* $\widetilde{W}'_{N,1} = \xi' W_{\mathcal{H}^*} + (1-\xi')\widetilde{W}_{N,1},\ \xi' \in [0,1]$. *Obviously, we have* $\widetilde{W}_{N,1},\ \widetilde{W}'_{N,1} \in \mathcal{B}(W_{\mathcal{H}^*}, R)$. *Thus*

$$\|\nabla F_N(\widetilde{W}_{N,1})\|_F \leq \|\nabla F_N(W_{\mathcal{H}^*})\|_F + L_N \|\widetilde{W}_{N,1} - W_{\mathcal{H}^*}\|_F \tag{106}$$

Noting that

$$\nabla F_N(W_{\mathcal{H}^*}) = \nabla f_N(W_{\mathcal{H}^*}) + 2\lambda_N \langle W_{\mathcal{H}^*} - \widehat{W}_N \rangle, \tag{107}$$

we have

$$\|\nabla F_N(W_{\mathcal{H}^*})\|_F \leq \|\nabla f_N(W_{\mathcal{H}^*})\|_F + 2\lambda_N \|W_{\mathcal{H}^*} - \widehat{W}_N\|_F, \tag{108}$$

further

$$\begin{aligned} &F_N(\overline{W}_{N,1}) - F_N(W_{N,1}) \\ &\leq \left(\|\nabla f_N(W_{\mathcal{H}^*})\|_F + 2\lambda_N \|W_{\mathcal{H}^*} - \widehat{W}_N\|_F + L_N \|\widetilde{W}_{N,1} - W_{\mathcal{H}^*}\|_F\right) \cdot \|\overline{W}_{N,1} - W_{N,1}\|_F \\ &\triangleq H_N \cdot \|\overline{W}_{N,1} - W_{N,1}\|_F \end{aligned} \tag{109}$$

where $H_N \triangleq \|\nabla f_N(W_{\mathcal{H}^*})\|_F + 2\lambda_N\|W_{\mathcal{H}^*} - \widehat{W}_N\|_F + L_N\|\widetilde{W}_{N,1} - W_{\mathcal{H}^*}\|_F$. *Further we have*

$$\gamma_N \frac{1}{|\widehat{W}_N(s_0,t_0)|}\,|M_{N,1}(s_0,t_0)| \le F_N(W_{\mathcal{H}^*} + \overline{M}_{N,1}) - F_N(W_{\mathcal{H}^*} + M_{N,1}) \le H_N\|\overline{M}_{N,1} - M_{N,1}\| = H_N\,|M_{N,1}(s_0,t_0)|\,. \tag{110}$$

On one hand, since $M_{N,1}(s_0,t_0) \neq 0$, *we deduce that*

$$\gamma_N \frac{1}{|\widehat{W}^w_N(s_0,t_0)|} \le H_N \tag{111}$$

On the other hand, by Theorem 6.1, we have $|\widehat{W}^w_N(s_0,t_0)| \le C_{\delta_1}\left(\frac{\log N}{\sqrt{N}}\right)^{1/\nu}$. *By the error bound of* $W_{N,1}$, *the definition of* $\overline{W}_{N,1}$ *and Lemma A.12, for* $N \ge N_0(\delta)$, *we can get*

$$\begin{aligned} H_N &= \|\nabla f_N(W_{\mathcal{H}^*})\|_F + 2\lambda_N\|W_{\mathcal{H}^*} - \widehat{W}_N\|_F + L_N\|\widetilde{W}_{N,1} - W_{\mathcal{H}^*}\|_F \\ &\le C_{\delta_3}\sqrt{\frac{\log N}{N}} + 2\lambda_N C_{\delta_1}\left(\frac{\log N}{\sqrt{N}}\right)^{1/\nu} \\ &\quad + L_N\left(C_{\delta_2}\frac{2}{l_N}\sqrt{\frac{\log N}{N}} + C_{\delta_1}\frac{4\lambda_N}{l_N}\left(\frac{\log N}{\sqrt{N}}\right)^{1/\nu} + \frac{2C_3}{l_N}\gamma_N\right) \to 0 \text{ as } N \to \infty \end{aligned} \tag{112}$$

with probability at least $1-\delta$. *By the choice of* γ_N, *i.e.,* $\frac{\gamma_N}{\left(\frac{\log N}{\sqrt{N}}\right)^{1/\nu}} \to \infty$ *as* $N \to \infty$, *there exists* $N_1(\delta)$ *for* $\forall N \ge N_1(\delta)$ *we can obtain*

$$\gamma_N \frac{1}{|\widehat{W}^w_N(s_0,t_0)|} \ge \frac{1}{C_{\delta_1}} \frac{\gamma_N}{\left(\frac{\log N}{\sqrt{N}}\right)^{1/\nu}} > H_N \tag{113}$$

with probability at least $1-\delta$. *It contradicts equation (111). Now we complete the proof of the sparse consistency.*

Step 2. *Nextly, we prove the low-rank consistency of the estimates* $W_{N,*}$.

Denote $W_{N,*} = U_{N,*}\Sigma_{N,*}V^T_{N,*}$ *and* $\overline{W}_{N,*} = U_{N,*}\overline{\Sigma}_{N,*}V^T_{N,*}$, *where* $\Sigma_{N,*}(l) = \sigma_l(W_{N,*})$ *and* $\overline{\Sigma}_{N,*}(l) = \sigma_l(\overline{W}_{N,*})$.

$$\overline{\Sigma}_{N,*}(l) = \Sigma_{N,*}(l), \text{if } l \in \mathcal{B}^C_{\mathcal{H}^*} \tag{114}$$

$$\overline{\Sigma}_{N,*}(l) = 0, \qquad \text{if } l \in \mathcal{B}_{\mathcal{H}^*}. \tag{115}$$

Suppose that there exists some $l_0 \in \mathcal{B}_{\mathcal{H}^*}$ *such that* $\sigma_{l_0}(\widehat{W}_N) \neq 0$.

Still using the optimality of $W_{N,*}$, *we can obtain*

$$\begin{aligned} 0 &\ge J_N(W_{N,*}) - J_N(\overline{W}_{N,*}) \\ &= F_N(W_{N,*}) + \gamma_N\|W_{N,*}\|_{\widehat{W}^w_N,*} - F_N(\overline{W}_{N,*}) - \gamma_N\|\overline{W}_{N,*}\|_{\widehat{W}^w_N,*} \end{aligned} \tag{116}$$

Note that

$$\begin{aligned} \|W_{N,*}\|_{\widehat{W}^w_N,*} - \|\overline{W}_{N,*}\|_{\widehat{W}^w_N,*} &= \sum_{l=1}^{K}\frac{\sigma_l(W_{N,*}) - \sigma_l(\overline{W}_{N,*})}{\sigma_l(\widehat{W}^w_N)} \\ = \sum_{l\in\mathcal{B}_{\mathcal{H}^*}}\frac{\sigma_l(W_{N,*})}{\sigma_l(\widehat{W}^w_N)} + \sum_{l\in\mathcal{B}^C_{\mathcal{H}^*}}\frac{\sigma_l(W_{N,*}) - \sigma_l(\overline{W}_{N,*})}{\sigma_l(\widehat{W}^w_N)} &= \frac{\sigma_{l_0}(W_{N,*})}{\sigma_{l_0}(\widehat{W}^w_N)}. \end{aligned} \tag{117}$$

Combining (116), further we can get

$$\gamma_N \frac{1}{\sigma_{l_0}(\widehat{W}^w_N)}\sigma_{l_0}(W_{N,*}) \le F_N(\overline{W}_{N,*}) - F_N(W_{N,*}) \tag{118}$$

By $l_N \cdot I \preceq \nabla^2 F_N(W) \preceq L_N \cdot I$, similar to the proof in Step 1, we can obtain

$$F_N(\overline{W}_{N,*}) - F_N(W_{N,*}) \leq H_N \|W_{N,*} - \overline{W}_{N,*}\|_F \tag{119}$$

Further we have

$$\begin{aligned}\gamma_N \frac{1}{\sigma_{l_0}(\widehat{W}_N^w)} \sigma_{l_0}(W_{N,*}) &\leq F_N(\overline{W}_{N,*}) - F_N(W_{N,*}) \\ &\leq H_N \|W_{N,*} - \overline{W}_{N,*}\|_F \\ &= H_N \|U_{N,*} \left(\Sigma_{N,*} - \overline{\Sigma}_{N,*}\right) V_{N,*}^T\|_F \\ &= H_N \|\Sigma_{N,*} - \overline{\Sigma}_{N,*}\|_F = H_N \cdot \sigma_{l_0}(W_{N,*}).\end{aligned} \tag{120}$$

On one hand, since $\sigma_{l_0}(W_{N,*}) \neq 0$, we deduce that

$$\gamma_N \frac{1}{\sigma_{l_0}(\widehat{W}_N)} \leq H_N. \tag{121}$$

On the other hand, by Theorem 6.1, we have $\sigma_{l_0}(\widehat{W}_N^w) \leq C_\delta \left(\frac{\log N}{\sqrt{N}}\right)^{1/\nu}$ and by the error bound of $W_{N,*}$, the definition of $\overline{W}_{N,*}$ and Lemma A.12, we can get $H_N \to 0$ as $N \to \infty$ with probability at least $1-\delta$. By the choice of γ_N, there exists $N_1(\delta)$ for $\forall N \geq N_1(\delta)$ we can obtain

$$\gamma_N \frac{1}{\sigma_{l_0}(\widehat{W}_N^w)} \geq \frac{1}{C_{\delta_1}} \frac{\gamma_N}{\left(\frac{\log N}{\sqrt{N}}\right)^{1/\nu}} > H_N \tag{122}$$

with probability at least $1-\delta$. It contradicts equation (121). Now we complete the proof of the low-rank consistency. ■

A.8 Experimental Details on Real Datasets

Our implementation is based on Pytorch 1.8. The hyper-parameters are listed in Table 3.

Table 3: Details of Experimental Setting on Real Datasets in Section 7.2.

Experiments	MNIST	CIFAR-10			Tiny-ImageNet
Network	LeNet-300-100	VGG16	ResNet20	ResNet56	ResNet18
Epochs	30	100	100	100	100
Batch Size	60	128	128	128	128
Optimizer	Adam	SGD	SGD	SGD	SGD
Momentum	-	0.9	0.9	0.9	0.9
Learning Rate	1.2e−3	0.005	0.005	0.005	0.005
Scheduler	-	CosineAnnealingLR	CosineAnnealingLR	CosineAnnealingLR	CosineAnnealingLR
Weight Decay	0	0	0	0	0
ϵ_w	0.005	0.0005	0.005	0.005	0.005
ϵ_s	0.05	0.05	0.05	0.05	0.05
λ_N	1	1	1	1	1
γ_N in L1 and Weighted L1	15	30	8	10	50
γ_N in Nuclear and Weighted Nuclear	8	25	8	10	50