Theoretical insights into the optimization landscape of over-parameterized shallow neural networks

Mahdi Soltanolkotabi[*]   Adel Javanmard[†]   Jason D. Lee[†]

July 15, 2017; Revised July 2022

Abstract

In this paper we study the problem of learning a shallow artificial neural network that best fits a training data set. We study this problem in the over-parameterized regime where the number of observations are fewer than the number of parameters in the model. We show that with quadratic activations the optimization landscape of training such shallow neural networks has certain favorable characteristics that allow globally optimal models to be found efficiently using a variety of local search heuristics. This result holds for an arbitrary training data of input/output pairs. For differentiable activation functions we also show that gradient descent, when suitably initialized, converges at a linear rate to a globally optimal model. This result focuses on a realizable model where the inputs are chosen i.i.d. from a Gaussian distribution and the labels are generated according to planted weight coefficients.

Dedicated to the memory of Maryam Mirzakhani.

1   Introduction

1.1   Motivation

Neural network architectures (a.k.a. deep learning) have recently emerged as powerful tools for automatic knowledge extraction from raw data. These learning architectures have lead to major breakthroughs in applications such as visual object classification [27], speech recognition [36] and natural language processing [16]. Despite their wide empirical use the mathematical success of these architectures remains a mystery. Although the expressive ability of neural networks is relatively well-understood [6], computational tractability of training such networks remains a major challenge. In fact, training neural nets is known to be NP-hard even for very small networks [9]. The main challenge is that training neural networks correspond to extremely high-dimensional and nonconvex optimization problems and it is not clear how to provably solve them to global optimality. Worst-case pessimism aside, these networks are trained successfully in practice via local search heuristics on real or randomly generated data. In particular, over-parameterized neural networks-where the number of parameters exceed the number of data samples-can be optimized to global optimality

---

[*]Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA
[†]Data Sciences and Operations Department, University of Southern California, Los Angeles, CA

1

using local search heuristics such as gradient or stochastic gradient methods [55]. In this paper we wish to provide theoretical insights into this phenomenon by developing a better understanding of optimization landscape of such over-parameterized shallow neural networks.

1.2 Problem formulation and models

A fully connected artificial neural network is composed of computational units called neurons. The neurons are decomposed into layers consisting of one input layer, one output layer and a few hidden layers with the output of each layer fed in (as input) to the next layer. In this paper we shall focus on neural networks with only a single hidden layer with d inputs, k hidden neurons and a single output. The overall input-output relationship of the neural network in this case is a function $f : \mathbb{R}^d \to \mathbb{R}$ that maps the input vector $\boldsymbol{x} \in \mathbb{R}^d$ into a scalar output via the following equation

$$\boldsymbol{x} \mapsto f_{\boldsymbol{v},\boldsymbol{W}}(\boldsymbol{x}) := \sum_{\ell=1}^{k} \boldsymbol{v}_\ell \phi(\langle \boldsymbol{w}_\ell, \boldsymbol{x} \rangle). \tag{1.1}$$

In the above the vectors $\boldsymbol{w}_\ell \in \mathbb{R}^d$ contains the weights of the edges connecting the input to the ℓth hidden node and $\boldsymbol{v}_\ell \in \mathbb{R}$ is the weight of the edge connecting the ℓth hidden node to the output. Finally, $\phi : \mathbb{R} \to \mathbb{R}$ denotes the activation function applied to each hidden node. For more compact notation we gathering the weights $\boldsymbol{w}_\ell/\boldsymbol{v}_\ell$ into larger matrices $\boldsymbol{W} \in \mathbb{R}^{k \times d}$ and $\boldsymbol{v} \in \mathbb{R}^k$ of the form

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{w}_1^T \\ \boldsymbol{w}_2^T \\ \vdots \\ \boldsymbol{w}_k^T \end{bmatrix} \quad \text{and} \quad \boldsymbol{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_k \end{bmatrix}.$$

We can now rewrite our input-output model (1.1) in the more succinct form

$$\boldsymbol{x} \mapsto f_{\boldsymbol{v},\boldsymbol{W}}(\boldsymbol{x}) := \boldsymbol{v}^T \phi(\boldsymbol{W}\boldsymbol{x}). \tag{1.2}$$

Here, we have used the convention that when ϕ is applied to a vector is corresponds to applying ϕ to each entry of that vector. When training a neural network, one typically has access to a data set consisting of n feature/label pairs $(\boldsymbol{x}_i, y_i)$ with $\boldsymbol{x}_i \in \mathbb{R}^d$ representing the feature and y_i the associated label. We wish to infer the best weights $\boldsymbol{v}, \boldsymbol{W}$ such that the mapping $f_{\boldsymbol{v},\boldsymbol{W}}$ best fits the training data. More specifically, we wish to minimize the misfit between $f_{\boldsymbol{v},\boldsymbol{W}}(\boldsymbol{x}_i)$ and y_i via an optimization problem of the form

$$\min_{\boldsymbol{v},\boldsymbol{W}} \mathcal{L}(\boldsymbol{v}, \boldsymbol{W}), \tag{1.3}$$

where

$$\mathcal{L}(\boldsymbol{v}, \boldsymbol{W}) := \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - f_{\boldsymbol{v},\boldsymbol{W}}(\boldsymbol{x}_i) \right)^2 = \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \boldsymbol{v}^T \phi(\boldsymbol{W}\boldsymbol{x}_i) \right)^2. \tag{1.4}$$

In this paper we wish to understand the landscape of optimization problems of the form (1.4).

2

2   Main results

In this section we discuss the main results of this paper. The first set of results focus on understanding the global landscape of neural network optimization with one hidden layer with a particular activation function. We also discuss how this landscape characterization enables algorithms to find a global optima in polynomial time. The second set of results focuses on the local convergence of gradient descent but applies to a broad set of activation functions.

2.1   Global landscape analysis with quadratic activations

We begin by discussing some global properties of the loss function of training neural networks.

Theorem 2.1 *Assume we have an arbitrary data set of input/label pairs $\boldsymbol{x}_i \in \mathbb{R}^d$ and y_i for $i = 1, 2, \ldots, n$. Consider a neural network of the form*

$$\boldsymbol{x} \mapsto \boldsymbol{v}^T \phi(\boldsymbol{W}\boldsymbol{x}),$$

with $\phi(z) = z^2$ a quadratic activation and $\boldsymbol{W} \in \mathbb{R}^{k \times d}$, $\boldsymbol{v} \in \mathbb{R}^k$ denoting the weights connecting input to hidden and hidden to output layers. We assume $k \geq 2d$ and set the weights of the output layer $\boldsymbol{v}$ so as to have at least d positive entries and at least d negative entries. Then, the training loss as a function of the weights $\boldsymbol{W}$ of the hidden layer

$$\mathcal{L}(\boldsymbol{W}) = \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \boldsymbol{v}^T \phi(\boldsymbol{W}\boldsymbol{x}_i) \right)^2,$$

obeys the following two properties.

- *There are no spurious local minima, i.e. all local minima are global.*

- *All saddle points have a direction of strictly negative curvature. That is, at a saddle point $\boldsymbol{W}_s$ there is a direction $\boldsymbol{U} \in \mathbb{R}^{k \times d}$ such that*

$$\text{vect}(\boldsymbol{U})^T \nabla^2 \mathcal{L}(\boldsymbol{W}_s) \text{vect}(\boldsymbol{U}) < 0.$$

Furthermore, for almost every data inputs $\{\boldsymbol{x}_i\}_{i=1}^n$, as long as

$$d \leq n \leq cd^2,$$

the global optimum of $\mathcal{L}(\boldsymbol{W})$ is zero. Here, $c > 0$ is a fixed numerical constant.

The above result states that given an arbitrary data set, the optimization landscape of fitting neural networks have favorable properties that facilitate finding globally optimal models. In particular, by setting the weights of the last layer to have diverse signs all local minima are global minima and all saddles have a direction of negative curvature. This in turn implies that gradient descent on the input-to-hidden weights, when initialized at random, converges to a global optima. All of this holds as long as the neural network is sufficiently wide in the sense that the number of hidden units exceed the dimension of the inputs by a factor of two ($k \geq 2d$).

An interesting and perhaps surprising aspect of the first part of this theorem is its generality: it applies to any arbitrary data set of input/label pairs of any size! However, this result only

guarantees convergence to a global optima but does not explain how good this global model is at fitting the data. Stated differently, it does not provide a bound on the optimal value. Such a bound may not be possible with adversarial data. However, at least intuitively, one expects the optimal value to be small when the input data $\boldsymbol{x}_i$ are sufficiently diverse and the neural network is sufficiently over-parameterized. The second part of Theorem 2.1 confirms that this is indeed true. In particular assuming the input data $\boldsymbol{x}_i$ are distributed i.i.d. $\mathcal{N}(0, \boldsymbol{I}_d)$, and the total number of weights (kd) exceeds the number of data samples (n), the globally optimal model perfectly fits the labels (has optimal value of 0). The interesting part of this result is that it still holds with arbitrary labels. Thus, this theorem shows that with random inputs, and a sufficiently diverse choice of the hidden-output weights, using gradient descent to iteratively update the input-hidden layer weights converges to a model that provably fits arbitrary labels! This result is also inline with recent numerical evidence in [55] that demonstrates that stochastic gradient descent learns deep, over-parametrized models with zero training error even with an arbitrary choice of labels.

While the above theorem shows that the saddles are strict and there is a direction of negative curvature at every saddle point, the margin of negativity is not quantified. Thus, the above theorem does not provide explicit convergence guarantees. In the theorem below we provide such a guarantee albeit for a more restrictive data model.

Theorem 2.2 *Assume we have a data set of input/label pairs $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ with the labels $y_i \in \mathbb{R}$ generated according to a planted two layer neural network model of the form*

$$y_i = \langle \boldsymbol{v}^*, \phi(\boldsymbol{W}^* \boldsymbol{x}_i) \rangle,$$

with $\phi(z) = z^2$ a quadratic activation and $\boldsymbol{W}^* \in \mathbb{R}^{k \times d}$, $\boldsymbol{v}^* \in \mathbb{R}^k$ the weights of the input-hidden and hidden-output layer with $\sigma_{\min}(\boldsymbol{W}^*) > 0$. Furthermore, assume $k \geq d$ and that all the non-zero entries of $\boldsymbol{v}^*$ have the same sign (positive or negative). We set the hidden-output weights to a vector $\boldsymbol{v} \in \mathbb{R}^k$ with non-zero entries also having the same sign with at least d entries strictly nonzero (positive if the nonzero entries of $\boldsymbol{v}^*$ are positive and negative otherwise). Then, as long as

$$d \leq n \leq cd^2,$$

with c a fixed numerical constant, then the training loss as a function of the input-output weights $\boldsymbol{W}$ of the hidden layer

$$\mathcal{L}(\boldsymbol{W}) = \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \boldsymbol{v}^T \phi(\boldsymbol{W} \boldsymbol{x}_i) \right)^2,$$

obeys the following two properties, for almost every input data $\{\boldsymbol{x}_i\}_{i=1}^{n}$.

- *There are no spurious local minima, i.e. all local minima are global optima.*

- *All saddle points have a direction of negative curvature. That is, at a saddle point $\boldsymbol{W}_s$ there is a direction $\boldsymbol{U} \in \mathbb{R}^{k \times d}$ such that*

$$\text{vect}(\boldsymbol{U})^T \nabla^2 \mathcal{L}(\boldsymbol{W}_s) \text{vect}(\boldsymbol{U}) < 0.$$

- *The global optima has loss value equal to zero ($\mathcal{L}(\boldsymbol{W}) = 0$).*

4

Furthermore, suppose that the inputs $\boldsymbol{x}_i \in \mathbb{R}^d$ are distributed i.i.d. $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$. Assume $k > d$ and $cd \log d \leq n \leq Cd^2$ for fixed constants c and C. Also, assume we set all entries of $\boldsymbol{v}$ to ν (i.e. $\boldsymbol{v} = \nu \boldsymbol{1}$) with ν having the same sign as the nonzero entries of $\boldsymbol{v}^*$.

- *Then all points $\boldsymbol{W}$ satisfying the approximate local minima condition*

$$\|\nabla \mathcal{L}(\boldsymbol{W})\|_F \leq \epsilon_g \quad \text{and} \quad \nabla^2 \mathcal{L}(\boldsymbol{W}) \succeq -\epsilon_H \boldsymbol{I}, \tag{2.1}$$

  obey

$$\mathcal{L}(\boldsymbol{W}) \leq \frac{\epsilon_g}{\sqrt{\nu}} \max\left(\sqrt{1 + 14 \left\|\boldsymbol{W}^{*T} \boldsymbol{D}_{\boldsymbol{v}^*} \boldsymbol{W}^*\right\|_*}, 4 \frac{\epsilon_g}{\sqrt{\nu}}\right) + \frac{\epsilon_H}{2\nu} \left\|\boldsymbol{W}^{*T} \boldsymbol{D}_{\boldsymbol{v}^*} \boldsymbol{W}^*\right\|_*,$$

  with probability at least $1 - 12de^{-\gamma d} - 8/d$. Here γ is a fixed numerical constant.

The above result considers the setting where the input-output data set is generated according to a neural network model with Gaussian random input vectors. We show that if the data is generated according this model, then as long as the neural network is over-parameterized ($n \lesssim kd$) then all points obeying condition (2.1) have small objective value.

The reason this result is useful is that points obeying the approximate local minimum condition (2.1) can be found in time depending polynomially on $\frac{1}{\epsilon_g}$ and $\frac{1}{\epsilon_H}$. Algorithms that provably work in this setting include cubic regularization [37], trust region methods [17], approximate cubic regularization schemes [4, 13, 14], randomly initialized and perturbed (stochastic) gradients [23, 18, 30, 31, 23]. Therefore, the above theorem demonstrates that a weight matrix with small training error(i.e. $\mathcal{L}(\boldsymbol{W}) \leq \epsilon$) can be found in a computationally tractable manner (specifically with poly($\frac{1}{\epsilon}$) computational effort).

2.2 Local convergence analysis with general activations

We now extend our analysis to general activations functions. However, our results in this section require a sufficiently close initialization to the global optimum. Starting from such a sufficiently close initialization we apply gradient descent updates based on the loss function (1.4). Our results apply to any differentiable activation function with bounded first and second order derivatives. We believe our result will eventually extend also to non-differentiable activations by smoothing techniques but we do not pursue this direction in this paper.

Before we state our theorem, we make a technical assumption regarding the activation ϕ.

Assumption 2.3 *For $\sigma \in \mathbb{R}_{\geq 0}$, define $\mu(\sigma) = E[\phi'(\sigma g)]$ and $\gamma(\sigma) = E[\phi''(\sigma g)]$ where $g \sim \mathcal{N}(0, 1)$. We consider activations ϕ such that $\mu(\sigma)$ is zero/nonzero everywhere. Likewise, we assume that one of the following holds true for $\gamma(\sigma)$:*

(a) *Function $\gamma(\sigma)$ is nonzero everywhere.*

(b) *Function $\gamma(\sigma)$ is identical to zero.*

We note that $\mu(\sigma)$ and $\gamma(\sigma)$ can be thought of as the average slope and curvature of the activation function. Thus the assumption on $\mu(\sigma)$ can be interpreted as the activation should have a nonzero average slope for all $\sigma > 0$ (i.e. the mapping is somewhat correlated with the identity mapping $\phi(x) = x$, under any gaussian measure) or has average slope equal to zero for all $\sigma > 0$ (i.e. the mapping has no correlation with the identity mapping, under any gaussian measure).
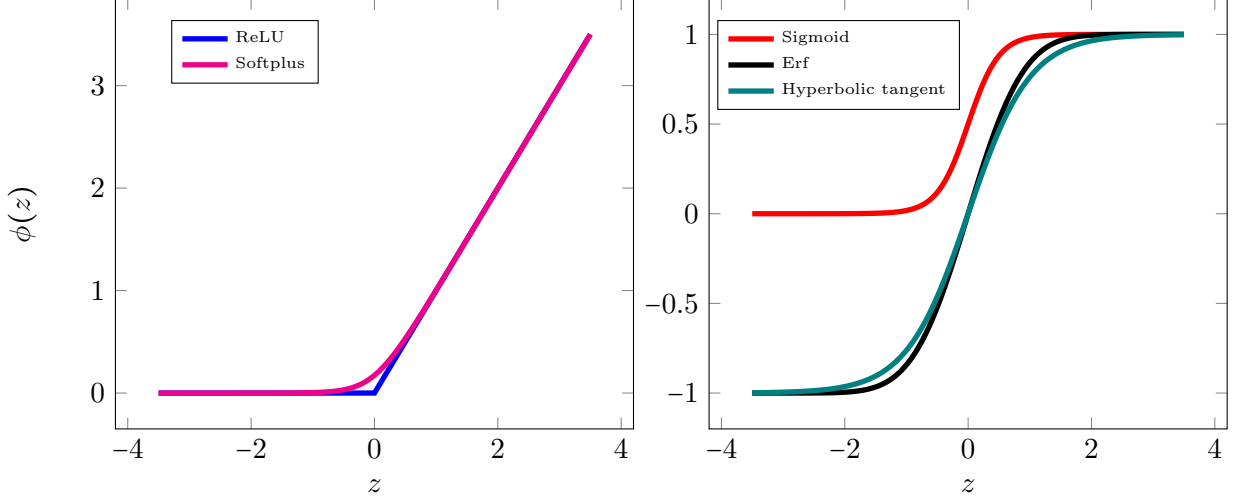
Figure 1: Different activations from Example 2.4 with $b = 4$ along with the ReLU activation.

Example 2.4 *We provide several examples of an activation function that satisfy Assumption 2.3.*

1. *(Softplus)* $\phi_b(z) = \frac{1}{b}\log(1 + e^{bz})$, *for a fixed $b > 0$.*

2. *(Sigmoid)* $\phi_b(z) = \frac{1}{1+e^{-bz}}$, *for a fixed $b > 0$.*

3. *(Erf)* $\phi(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2/2}\mathrm{d}t$.

4. *(Hyperbolic Tangent)* $\phi(z) = \tanh(z)$.

Note that all of these activations obey $\mu(\sigma) > 0$ as they are all strictly increasing. Furthermore, the Softplus activation satisfies Assumption 2.3 (a) because it is strictly convex, while the other three activations satisfy Assumption 2.3 (b), because ϕ'' is an odd function in these cases. These activations along with the popular ReLU activation ($\phi(z) = \max(0, z)$) are depicted in Figure 1.

We now have all the elements to state our local convergence result.

Theorem 2.5 *Assume we have a data set of input/label pairs $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ with the inputs $\boldsymbol{x}_i \in \mathbb{R}^d$ distributed i.i.d. $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ and the labels $y_i \in \mathbb{R}$ generated according to a planted two layer neural network model of the form*

$$y_i = \langle \boldsymbol{v}^*, \phi(\boldsymbol{W}^*\boldsymbol{x}_i)\rangle,$$

Here, $\phi : \mathbb{R} \to \mathbb{R}$ is any activation function with bounded first and second derivatives, satisfying Assumption 2.3(a). Further, $\boldsymbol{W}^* \in \mathbb{R}^{k \times d}$, $\boldsymbol{v}^* \in \mathbb{R}^k$ are the weights of the input-hidden and hidden-output layer. We also assume that the planted weight vector/matrix $\boldsymbol{v}^*/\boldsymbol{W}^*$ obey

$$0 < v_{\min} \le |v_\ell^*| \le v_{\max} \quad 0 < w_{\min} \le \|\boldsymbol{w}_\ell^*\|_{\ell_2} \le w_{\max} \quad \text{for } \ell = 1, 2, \ldots, k, \tag{2.2}$$

for some fixed constants $v_{\min}, v_{\max}, w_{\min},$ *and* $w_{\max}$. *Further, assume that* $k \geq d$ *and*

$$d \leq n \leq c_0 \frac{\sigma_{\min}^4(\boldsymbol{W}^*)}{\sigma_{\max}^8(\boldsymbol{W}^*)} d^2 \tag{2.3}$$

for a sufficiently small constant $c_0 > 0$.

To estimate the weights $\boldsymbol{v}^*$ *and* $\boldsymbol{W}^*$, *we start from initial weights* $\boldsymbol{v}_0$ *and* $\boldsymbol{W}_0$

$$\|\boldsymbol{W}_0 - \boldsymbol{W}^*\|_F \leq C_0 \frac{\sigma_{\min}^3(\boldsymbol{W}^*)}{\sigma_{\max}(\boldsymbol{W}^*)} \cdot \frac{d^{2.5}}{n^{1.5}k}, \tag{2.4}$$

$$\|\boldsymbol{v}_0 - \boldsymbol{v}^*\|_\infty \leq C_0 \frac{\sigma_{\min}^3(\boldsymbol{W}^*)}{\sigma_{\max}(\boldsymbol{W}^*)} \cdot \frac{d^{2.5}}{n^{1.5}k^{1.5}}, \tag{2.5}$$

and apply gradient descent updates of the form

$$\boldsymbol{v}_{\tau+1} = \boldsymbol{v}_\tau - \alpha \nabla_{\boldsymbol{v}} \mathcal{L}(\boldsymbol{v}_\tau, \boldsymbol{W}_\tau),$$
$$\boldsymbol{W}_{\tau+1} = \boldsymbol{W}_\tau - \alpha \nabla_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{v}_\tau, \boldsymbol{W}_\tau), \tag{2.6}$$

with the learning rate obeying $\alpha \leq 1/\beta$. *Then there is an event of probability at least* $1 - n^{-1} - 2ne^{-b\sqrt{d}}$, *such that on this event starting from any initial point obeying* (2.4)-(2.5) *the gradient descent updates* (2.6) *satisfy*

$$\mathcal{L}(\boldsymbol{v}_\tau, \boldsymbol{W}_\tau) \leq \left(1 - c\alpha \frac{d}{n}\right)^\tau \mathcal{L}(\boldsymbol{v}_0, \boldsymbol{W}_0).$$

Here, β *is given by*

$$\beta := C \left(\sigma_{\max}^2(\boldsymbol{W}^*) + 1\right) k, \tag{2.7}$$

and $b, C_0, c_0, C, c > 0$ *are fixed numerical constants.*

Remark 2.6 *We note that the above theorem is still valid if Assumption 2.3 (b) holds in lieu of Assumption 2.3 (a). The only change is that the term* $\sigma_{\min}(\boldsymbol{W})$ *should be now replaced by one in Equations* (2.3), (2.4) *and* (2.5).

The above result considers the setting where the input-output data set is generated according to a neural network model with a general activation and Gaussian random input vectors. We show that if the data is generated according this model, then as long as the neural network is over-parameterized ($n \lesssim kd$) then gradient descent converges to the planted model when initialized close to this planted model. This implies that a training error of zero can be achieved locally, using gradient descent. We would like to note that assumptions (2.2) on the weights of the planted neural networks are only made to avoid unnecessarily complicated expressions in the statement of the theorem. As it will become clear in the proofs (Section 6) our result continues to hold without these assumptions.

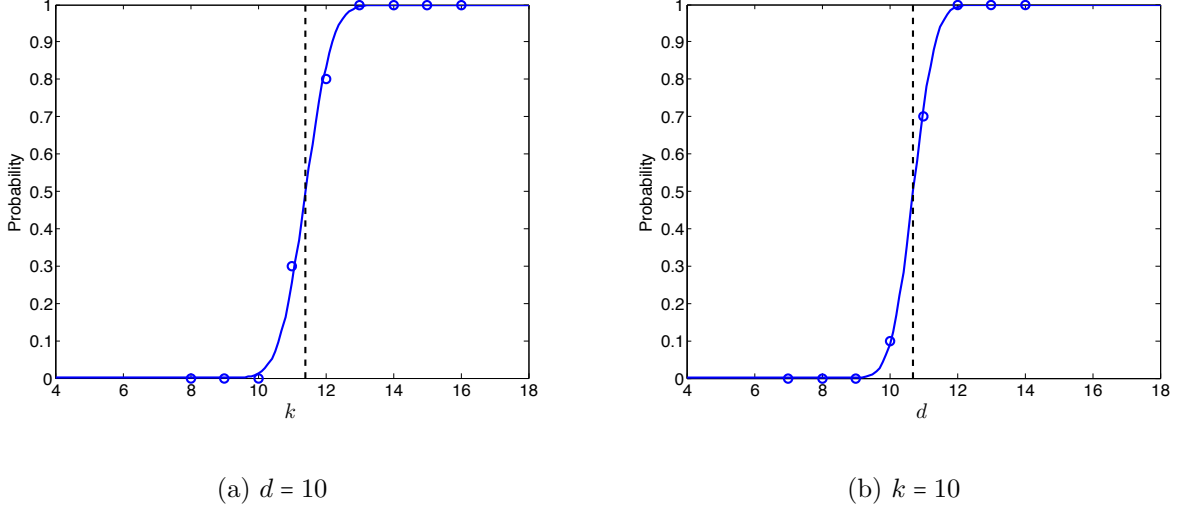|	
(a) $d = 10$	(b) $k = 10$

Figure 2: Empirical probability of having no spurious local minimizers for a one-hidden layer neural network with softplus activation. Dotted points depict the empirical probabilities, the solid line is obtained by fitting a logistic model to the dotted points, and the dashed line depicts the point where the probability is 1/2. The number of samples are equal to $n = 100$. In (a) we fix $d = 10$ and vary k and in (b) we fix $k = 10$ and vary d. In both experiments as the number of parameters (kd) increases beyond the sample size n, the probability of having no spurious local minimizer increases.

3 Numerical experiments

In this section, we numerically investigate whether gradient descent finds the global optimum for various configurations of n, k, d. In our first experiment, the data is generated from a planted Gaussian model of the form

$$y_i = \langle \boldsymbol{v}^*, \phi(\boldsymbol{W}^* \boldsymbol{x}_i) \rangle \quad \text{with} \quad \boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d),$$

and $\boldsymbol{v}^*$ the all-one vector. We consider two activations, namely the softplus $\phi(z) = \frac{1}{10} \log \left(1 + e^{10z}\right)$ and the quadratic activation $\phi(z) = z^2$.

In Figure 2, we show the results for the softplus activation with $n = 100$. In Figure 2a, we fix the input dimension at $d = 10$ and vary the number of hidden nodes k. For each value of (n, k, d) we carryout 10 experiments. In each experiment, we generate $\boldsymbol{W}^*$ at random with i.i.d. $\mathcal{N}(0,1)/\sqrt{d}$ entries. For each value of $\boldsymbol{W}^*$ we carry out 10 trials where in each trial we run gradient descent starting from a random initialization pair $(\boldsymbol{v}_0, \boldsymbol{W}_0)$ generated at random with $\boldsymbol{v}_0$ consisting of i.i.d. Rademacher ± 1 entries and $\boldsymbol{W}_0$ with i.i.d. $\mathcal{N}(0,1)/\sqrt{d}$ entries. If a global minimum was obtained for every single initialization, we declare that the loss function has no spurious local minima for the corresponding $\boldsymbol{W}^*$. In Figure 2 we plot the empirical probability that the loss function has no spurious local optima for the softplus activation. Dots correspond to the simulation results and the solid curve is obtained by fitting a logistic model to the results. In Figure 2a we focus on different number of hidden units, with the input dimension fixed at $d = 10$. The dashed line indicates the value of k ($k = 11.40$) at which the fitted logistic model crosses the probability 1/2. As k grows we enter a region that the gradient descent frequently converges to
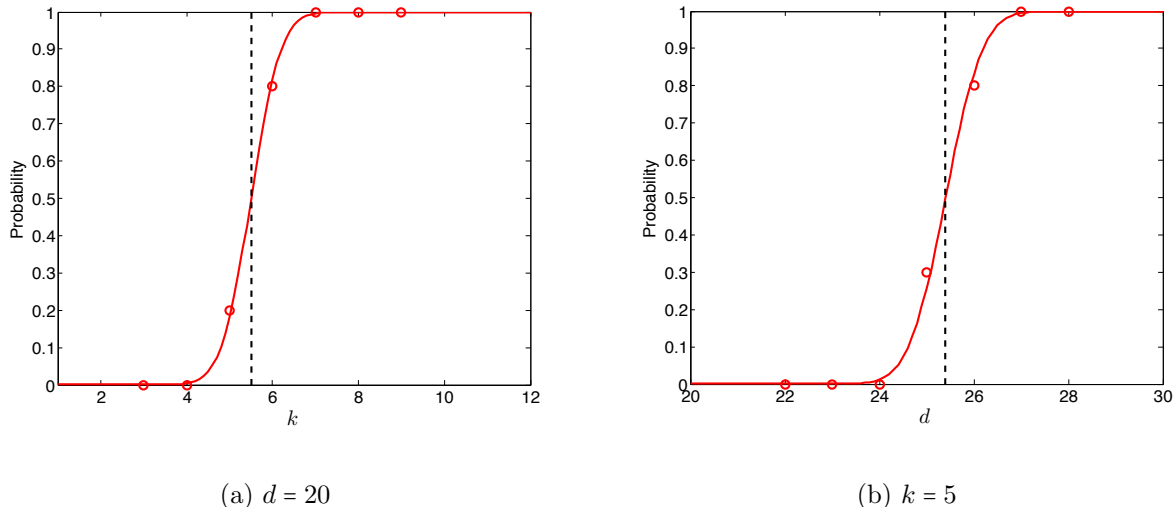
(a) $d = 20$ (b) $k = 5$

Figure 3: Empirical probability of having no spurious local minimizers for a one-hidden layer neural network with quadratic activations. Dotted points depict the empirical probabilities, the solid line is obtained by fitting a logistic model to the dotted points, and the dashed line depicts the point where the probability is $1/2$. The number of samples is $n = 100$. In (a) we fix $d = 10$ and vary k and in (b) we fix $k = 10$ and vary d. In both experiments as the number of parameters (kd) increases beyond the sample size n, the probability of having no spurious local minimizer increases.

global minimizers, and for $k \geq 13$ every single local minimizer found by gradient descent is a global minimizer. Note that in this regime the number of parameters (kd) exceeds the sample size n. This suggests that a modest amount of over-parameterization is sufficient for the no spurious local optima conclusion to hold and the global convergence to occur. In Figure 2b, we fix the number of hidden nodes at $k = 10$ and vary the input dimension d. We observe a similar trend: As d grows we enter a region where every single local minimizer found by gradient descent is a global minimizer. Here, the dashed line is located at $d = 10.70$, corresponding to a $1/2$ probability of converging to a global minima. For $d \geq 12$, gradient descent always finds a global minimizer.

We also repeat a similar experiment with quadratic activations and report the results in Figure 3. In Figure 3a, we fix the input dimension at $d = 20$ and the number of samples at $n = 100$ and vary k. The dashed line here is located at $k = 5.5$. In Figure 3b, we fix the number of hidden units at $k = 5$ and vary d. The dashed line in this experiment is located at $d = 25.40$. These experiments further corroborate our theory that over-parameterization helps gradient descent find a global minimizer with quadratic activations.

In the next set of experiments, we generate the labels and features i.i.d. at random

$$\boldsymbol{x}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d) \quad \text{and} \quad y_i \sim \mathcal{N}(0, 1).$$

We fit a softplus one-hidden layer network of the form $\boldsymbol{x} \to \mathbf{1}^T \phi(\boldsymbol{W}\boldsymbol{x})$ to this data with varying number of hidden units k. In Figure 4 we plot the square root of the objective function $\mathcal{L}$ as a function of iterations. Interestingly, we observe that even for randomly labeled data when $kd > 1.5n$, gradient descent is able to find a global minimizer. These experiments show that simple gradient descent is almost always able to find a global minimizer in the sufficiently over-parametrized regime,

9
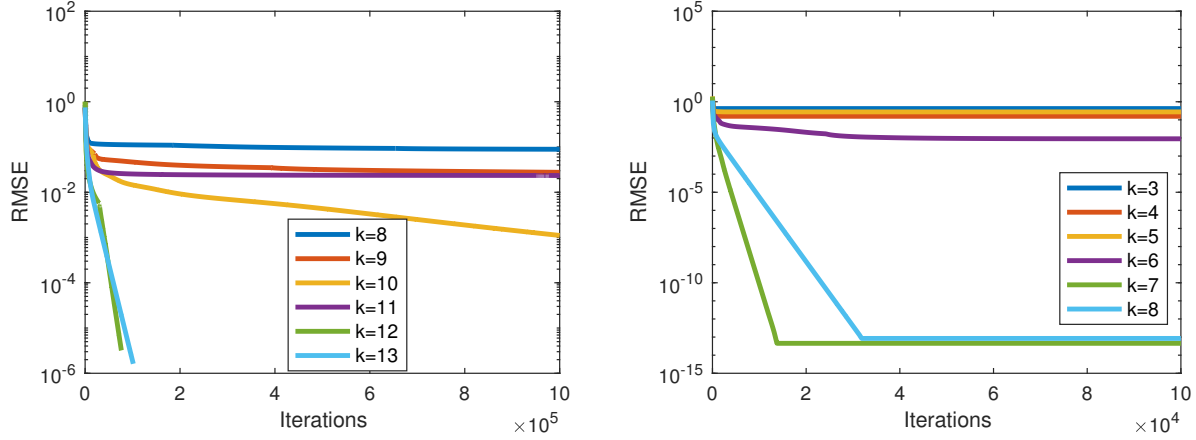
(a) Softplus with random data $(d, n) = (10, 100)$. (b) Quadratic with random data $(d, n) = (20, 100)$.

Figure 4: Plot of RMSE$=\sqrt{\mathcal{L}(\boldsymbol{v}, \boldsymbol{W})}$ vs. the number of gradient descent iterations. As the number of hidden units k increases, gradient descent converges to a solution with zero RMSE.

regardless of whether the data is generated from a planted model.

4   Related Work

As mentioned earlier, neural networks have enjoyed great empirical success [27, 36, 16]. To explain this success, many papers have studied the expressive ability of shallow neural networks dating back to the 80s (e.g. see [6]). More recently, interesting results have focused on the expressive ability of deeper and sparser architectures [49, 10, 28, 53]. Computational tractability of training networks however is still a major challenge. In fact, training neural nets is known to be NP-hard even for very small networks [9]. Despite this worst-case pessimism, local search heuristics such as gradient descent are surprisingly effective. For instance, [55] empirically demonstrated that sufficiently over-parametrized networks can be efficiently optimized to near global optimality with stochastic gradient descent.

For a neural network with zero hidden units and a single output with a monotonic activation function σ, numerous authors [35, 21, 24, 25, 45] have shown that gradient-based methods converge to the global optimum under various assumptions and models. As a result of this literature a good understanding of the optimization landscape of learning single activations have emerged (at least when the data set is generic). Roughly stated, in this case the loss function only has a single local optima that coincides with the global optima and the gradient descent direction is sufficiently correlated with the direction pointing to the global optimum. Thus these results are able to explain the success of gradient-based methods. However, when there are hidden units, the loss function exhibits a completely different landscape, so that such analyses do not easily generalize.

We now turn our attention to global convergence results known for deeper architectures. For a 2-layered network with leaky ReLU activation, [47] showed that gradient descent on a modified loss function can obtain a global minimum of the modified loss function; however, this does not

10

imply reaching a global minimum of the original loss function. Under the same setting, [54] showed that critical points with large "diversity" are near global optimality. [15] used several assumptions to simplify the loss function to a polynomial with i.i.d. Gaussian coefficients. They then showed that every local minima of the simplified loss has objective value comparable to the global minima. [26] used similar assumptions to show that all local minimum are global minimum in a nonlinear network. However, [15, 26] require an *independent activations* assumption meaning the activations of the hidden units are independent of the input and/or mutually independent, which is violated in practice. In comparison, our global convergence results hold for any arbitrary data set, without any additional assumptions. However, our global convergence results (see Section 2.1) only focus on quadratic activations with a single hidden layer.

We would like to mention a few interesting results regarding asymptotic characterizations of the dynamics of training shallow neural networks with Gaussian inputs using ideas from statistical physics [41, 8, 52]. Saad and Solla [41] study the online dynamics of learning a fully connected neural network with one-hidden layer in a Gaussian planted model (or Gaussian student-teacher model). In this model the output labels are generated according to planted weight vectors (a.k.a. a teacher network) with Gaussian input data. Then a new "student" network is trained to find the teacher weights. This result focuses on the case where the hidden-to-output weights are fixed to +1 (a.k.a. soft committee machines) and the asymptotic regime where the number of inputs and the iterations tend to infinity ($d \to +\infty$ and $\tau \to \infty$). The authors provide certain dynamical equations that describes the asymptotic behavior of the training process for one-hidden layer neural networks using an "online SGD" or "resampled SGD" procedure where the SGD update is performed using a fresh and independent data point per new iteration. These dynamical equations do not admit a closed form solution, but do enable the analysis of the convergence behavior close to stationarity. Furthermore, by simulating these dynamical equations it is possible to gain some insights about the evolution of the online SGD updates. In particular, the authors use such simulations to gain insights into the convergence/divergence of such neural networks in the special case where diagonal input-hidden weights ($\boldsymbol{W}$) are used. The paper [8] also studies studies soft-committee machines under similar assumptions to those of [41]. The authors of [8] use tools from statistical physics (and in particular the aforementioned dynamical equations) to demonstrate that the dynamics of training such neural networks has several fixed points and find learning-rate dependent phenomena, such as splitting and disappearing of fixed points. This paper also differs from [41] in that it also applies to transient iterations (finite τ). Finally, the paper [52] also studies soft committee machines but instead of online SGD, they present the analysis for a locally optimized online learning algorithm that focuses on extracting the largest possible amount of information from each new sample. The authors also demonstrate numerically that this choice leads to faster escape from the plateaux and hence faster decay of the generalization error. Our results are not directly comparable to this literature as it differs in terms of assumptions and conclusions in a variety of ways. In particular we focus on (1) analytic formulas, (2) gradient descent without resampling, (3) a non-asymptotic regime (finite d) and transient dynamics (finite τ), (4) the over-parameterized case $n < kd$, and (5) general hidden-output weights. Some of the results presented in this paper also apply without assuming a planted model or random/Gaussian input data (e.g. see Theorem 2.1).

We would like to note that there is also interesting growing literature on learning shallow neural networks with a single hidden layer with i.i.d. inputs, and under a realizable model (i.e. the labels are generated from a network with planted weights) [50, 11, 56, 33, 22, 58]. For isotropic Gaussian inputs, [50] shows that with two hidden unites ($k = 2$) there are no critical points for configurations

where both weight vectors fall into (or outside) the cone of ground truth weights. With the same assumptions, [11] proves that for a single-hidden ReLU network with a single non-overlapping convolutional filter, all local minimizers of the population loss are global; they also give counter-examples in the overlapping case and prove the problem is NP-hard when inputs are not Gaussian. [56] show that for single-hidden layer networks with non-standard activation functions gradient descent converges to global minimizers. [33] focuses on a Recursive Neural Net (RNN) model where the hidden to output weights are close to the identity and shows that stochastic methods converge to the planted model over the population objective. [58] studies general single-hidden layer networks and shows that a version of gradient descent which uses a fresh batch of samples in each iteration converges to the planted model. This holds using an initialization obtained via a tensor decomposition method. Our approach and local convergence results differ from this literature in a variety of different ways. First, unlike some of these results such as [11, 33], we study the optimization properties of the empirical function, not its expected value. Second, we focus on the over-parametrized regime ($n << kd$) which is the regime where most popular neural networks are trained. Mathematically, this is an important distinction as in this data-poor regime the empirical loss is no longer close to the population loss and therefore one can no longer infer the convergence behavior of gradient descent based on connecting the empirical loss to the population loss. Third, we optimize over both weights v and W, while most of this literature assumes $v = v^* = \mathbf{1}$ [33, 58]. Finally, our framework does not require a fresh batch of samples per new gradient iteration as in [58].

Several publications study the effect of over-parametrization on the training of neural networks [40, 20, 38]. These results require a large amount of over-parametrization, mainly that the width of one of the hidden layers to be greater than the number of training samples, which is unrealistic for commonly used networks. For instance, for the case of a single hidden layer using the notation of this paper these publications require $n \lesssim k$. These results also require additional technical assumptions which we do not detail here. However, these results work for fairly general activations and also deep architectures. In comparison, our global optimality results only allow for quadratic activations and a single hidden layer. However, our results are completely deterministic and allow for modest over-parameterization ($n \lesssim kd$). Generalizing these result to other activations and deeper architectures is an interesting future direction. While discussing quadratic activations we would like to mention the interesting paper [34]. This paper is perhaps the first result to clearly state that over-parameterization is helpful for optimization purposes. The authors further showed that the Frank-Wolfe algorithm can optimize a neural net with quadratic activations of the form $\phi(\boldsymbol{x}) = b + \langle \boldsymbol{w}_0, x \rangle + \sum_{\ell=1}^{K} \alpha_k(\langle w_\ell, x \rangle)^2$ under norm constraints of the form $|\alpha_\ell| \le 1$ and $\|\boldsymbol{w}_\ell\|_{\ell_2} = 1$. In comparison we focus on slightly less general activations (no linear term) and study landscape properties that are directly useful for analyzing gradient descent/local search methods in lieu of Franke-Wolfe type schemes. Finally, in the case of over-parameterized neural networks not all global optima may generalize to new data instances. Understanding the generalization capability of the solutions reached by (stochastic) gradient descent is an important research direction. See [7] for an interesting result in this direction under margin assumptions.

Another line of research [19, 43, 57] focuses on improper learning models using kernel-based approaches. These results hold under much more general data models than the realizable case. However, the practical success of deep learning is intricately tied to using gradient-based training procedures, and the learnability of these networks using improper learning does not explain the success of gradient-based methods. Related, [22] proposes a method of moments estimator using

tensor decomposition.

Finally, we would like to mention a few interesting recent developments that appeared after the initial arxiv submission of this paper. The authors of [42] demonstrated analytically (with variable precision arithmetics) that without over-parameterization the landscape of one-hidden layer neural networks has bad local minima as soon as the number of hidden unites exceeds six ($k \geq 6$). This interesting result clearly demonstrates that the over-parametrization is necessary to ensure a favorable optimization landscape when training one-hidden layer neural networks. Another recent paper [48] by Soudry and Hoffer suggests that "most" local minima are global in the over-parameterized regime. This paper provides further evidence that over-parametrizing neural networks leads to more favorable optimization landscapes when training such neural nets. Finally, the recent paper [32] proves that gradient descent when initialized at zero can lead to training errors close to (but not equal to) zero. This result holds for one-hidden layer neural networks with quadratic activations and Gaussian inputs. Developing rigorous convergence guarantees to a global optima with training error equal to zero which also applies to other activations is an important future research direction.

5   Preliminaries and notations

Before we dive into the proofs in the next two sections we collect some useful results and notations in this section.

5.1   Notations

For two matrices $\boldsymbol{A} = [\boldsymbol{A}_1, \dots, \boldsymbol{A}_p] \in \mathbb{R}^{m \times p}$ and $\boldsymbol{B} = [\boldsymbol{B}_1, \dots, \boldsymbol{B}_p] \in \mathbb{R}^{n \times p}$, we define their Khatri-Rao product as $\boldsymbol{A} * \boldsymbol{B} = [\boldsymbol{A}_1 \otimes \boldsymbol{B}_1, \dots, \boldsymbol{A}_p \otimes \boldsymbol{B}_p] \in \mathbb{R}^{mn \times p}$, where $\otimes$ denotes the Kronecher product. For two matrices $\boldsymbol{A}, \boldsymbol{B}$, we denote their Hadamard (entrywise) product by $\boldsymbol{A} \circ \boldsymbol{B}$. For a matrix $\boldsymbol{A}$, we denote its maximum and minim singular values by $\sigma_{\max}(\boldsymbol{A})$ and $\sigma_{\min}(\boldsymbol{A})$, respectively. For a random variable Y, its sub-exponential norms is defined as

$$\|Y\|_{\psi_1} = \inf\{C > 0 : \ \mathbb{E}\exp(|Y|/C) \leq 2\} \,.$$

Further, for a centered random vector $\boldsymbol{x} \in \mathbb{R}^d$, we define its sub-exponential norm as

$$\|\boldsymbol{x}\|_{\psi_1} = \sup_{\boldsymbol{y} \in S^{d-1}} \|\langle \boldsymbol{x}, \boldsymbol{y} \rangle\|_{\psi_1} \tag{5.1}$$

Throughout the paper we use c, C to refer to constants whose values may change from line to line.

5.2   Derivative calculations

In this section we gather some derivative calculations that we will use throughout the proof. As a reminder the loss function is given by

$$\mathcal{L}(\boldsymbol{v}, \boldsymbol{W}) = \frac{1}{2n} \sum_{i=1}^{n} \left(\boldsymbol{v}^T \phi(\boldsymbol{W} \boldsymbol{x}_i) - y_i\right)^2 \,.$$

To continue let us define the residual vector $\boldsymbol{r} \in \mathbb{R}^n$ with the ith entry given by

$$r_i = \boldsymbol{v}^T \phi(\boldsymbol{W} \boldsymbol{x}_i) - y_i.$$

13

5.2.1 First order derivatives

We begin by calculating the gradient with respect to the weight matrix $\boldsymbol{W}$. To this aim we begin by calculating the gradient with respect to the qth row of $\boldsymbol{W}$ denoted by $\boldsymbol{w}_q$. This is equal to

$$\nabla_{\boldsymbol{w}_q}\mathcal{L}(\boldsymbol{v},\boldsymbol{W}) = \frac{v_q}{n}\sum_{i=1}^n \left(\boldsymbol{v}^T\phi(\boldsymbol{W}\boldsymbol{x}_i) - y_i\right)\phi'(\langle\boldsymbol{w}_q,\boldsymbol{x}_i\rangle)\boldsymbol{x}_i = \frac{v_q}{n}\sum_{i=1}^n r_i\phi'(\langle\boldsymbol{w}_q,\boldsymbol{x}_i\rangle)\boldsymbol{x}_i. \tag{5.2}$$

Aggregating these gradients as a row vector and setting $\boldsymbol{D}_{\boldsymbol{v}} = \operatorname{diag}(v_1,\ldots,v_k)$ we arrive at

$$\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{v},\boldsymbol{W}) = \boldsymbol{D}_{\boldsymbol{v}}\left(\frac{1}{n}\sum_{i=1}^n r_i\phi'(\boldsymbol{W}\boldsymbol{x}_i)\boldsymbol{x}_i^T\right). \tag{5.3}$$

We also define the Jacobian matrix $\boldsymbol{J} = \begin{bmatrix} \boldsymbol{J}_1 & \boldsymbol{J}_2 & \ldots & \boldsymbol{J}_n \end{bmatrix} \in \mathbb{R}^{kd\times n}$ with

$$\boldsymbol{J}_i = \begin{bmatrix} v_1\phi'(\boldsymbol{w}_1^T\boldsymbol{x}_i)\boldsymbol{x}_i \\ v_2\phi'(\boldsymbol{w}_2\boldsymbol{x}_i)\boldsymbol{x}_i \\ \vdots \\ v_k\phi'(\boldsymbol{w}_k^T\boldsymbol{x}_i)\boldsymbol{x}_i \end{bmatrix}.$$

Let $\boldsymbol{X} \in \mathbb{R}^{d\times n}$ be the data matrix with the ith column equal to $\boldsymbol{x}_i$. Using the Khatri-Rao product we can rewrite the Jacobian matrix in the form

$$\boldsymbol{J} = \boldsymbol{D}_{\boldsymbol{v}}\phi'(\boldsymbol{W}\boldsymbol{X}) * \boldsymbol{X}. \tag{5.4}$$

Using this Jacobian matrix we can write the vectorized version of the gradient, i.e. gradient with respect to $\operatorname{vect}(\boldsymbol{W})$ as

$$\nabla_{\operatorname{vect}(\boldsymbol{W})}\mathcal{L}(\boldsymbol{v},\boldsymbol{W}) = \frac{1}{n}\boldsymbol{J}\boldsymbol{r}. \tag{5.5}$$

Taking the derivative with respect to v_q we arrive at

$$\frac{\partial}{\partial v_q}\mathcal{L}(\boldsymbol{v},\boldsymbol{W}) = \frac{1}{n}\sum_{i=1}^n \left(\boldsymbol{v}^T\phi(\boldsymbol{W}\boldsymbol{x}_i) - y_i\right)\phi(\langle\boldsymbol{w}_q,\boldsymbol{x}_i\rangle) = \frac{1}{n}\sum_{i=1}^n r_i\phi(\langle\boldsymbol{w}_q,\boldsymbol{x}_i\rangle).$$

Thus the gradient with respect to $\boldsymbol{v}$ is equal to

$$\nabla_{\boldsymbol{v}}\mathcal{L}(\boldsymbol{v},\boldsymbol{W}) = \frac{1}{n}\phi(\boldsymbol{W}\boldsymbol{X})\boldsymbol{r}. \tag{5.6}$$

5.2.2 Second order derivatives

Using (5.2) we have

$$\frac{\partial^2}{\boldsymbol{w}_p^2}\mathcal{L}(\boldsymbol{v},\boldsymbol{W}) = \frac{v_p}{n}\sum_{i=1}^n \left(\boldsymbol{v}^T\phi(\boldsymbol{W}\boldsymbol{x}_i) - y_i\right)\phi''(\langle\boldsymbol{w}_p,\boldsymbol{x}_i\rangle)\boldsymbol{x}_i\boldsymbol{x}_i^T + \frac{v_p^2}{n}\sum_{i=1}^n \left(\phi'(\langle\boldsymbol{w}_p,\boldsymbol{x}_i\rangle)\right)^2\boldsymbol{x}_i\boldsymbol{x}_i^T.$$

Also using (5.2), for $p \neq q$

$$\frac{\partial^2}{\partial\boldsymbol{w}_p\boldsymbol{w}_q}\mathcal{L}(\boldsymbol{v},\boldsymbol{W}) = \frac{v_p v_q}{n}\sum_{i=1}^n \phi'(\langle\boldsymbol{w}_p,\boldsymbol{x}_i\rangle)\phi'(\langle\boldsymbol{w}_q,\boldsymbol{x}_i\rangle)\boldsymbol{x}_i\boldsymbol{x}_i^T.$$

Thus

$$
(\text{vect}(\boldsymbol{U}))^T \nabla_{\boldsymbol{W}}^2 \mathcal{L}(\boldsymbol{v}, \boldsymbol{W}) \, (\text{vect}(\boldsymbol{U}))
$$

$$
= \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{v}^T \phi(\boldsymbol{W}\boldsymbol{x}_i) - y_i \right) \left(\sum_{p=1}^{k} v_p \phi''(\langle \boldsymbol{w}_p, \boldsymbol{x}_i \rangle)(\langle \boldsymbol{u}_p, \boldsymbol{x}_i \rangle)^2 \right)
$$

$$
+ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{p=1}^{k} v_p \phi'(\langle \boldsymbol{w}_q, \boldsymbol{x}_i \rangle) \langle \boldsymbol{u}_p, \boldsymbol{x}_i \rangle \right)^2
$$

$$
= \frac{1}{n} \sum_{i=1}^{n} r_i \left(\boldsymbol{x}_i^T \boldsymbol{U}^T \boldsymbol{D}_{\boldsymbol{v}} \boldsymbol{D}_{\phi''(\boldsymbol{W}\boldsymbol{x}_i)} \boldsymbol{U} \boldsymbol{x}_i \right) + \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{v}^T \boldsymbol{D}_{\phi'(\boldsymbol{W}\boldsymbol{x}_i)} \boldsymbol{U} \boldsymbol{x}_i \right)^2. \tag{5.7}
$$

Using the expression for the Jacobian matrix the latter can also be written in the form

$$
(\text{vect}(\boldsymbol{U}))^T \nabla_{\boldsymbol{W}}^2 \mathcal{L}(\boldsymbol{v}, \boldsymbol{W}) \, (\text{vect}(\boldsymbol{U}))
$$

$$
= \frac{1}{n} \sum_{i=1}^{n} \langle \boldsymbol{U}^T \boldsymbol{D}_{\boldsymbol{v}} \boldsymbol{D}_{\phi''(\boldsymbol{W}\boldsymbol{x}_i)} \boldsymbol{U}, r_i \boldsymbol{x}_i \boldsymbol{x}_i^T \rangle + \frac{1}{n} \text{vect}(\boldsymbol{U})^T \boldsymbol{J}\boldsymbol{J}^T \text{vect}(\boldsymbol{U}). \tag{5.8}
$$

Finally, note that

$$
\nabla_{\boldsymbol{v}}^2 \mathcal{L}(\boldsymbol{v}, \boldsymbol{W}) = \phi(\boldsymbol{W}\boldsymbol{X}) \phi(\boldsymbol{W}\boldsymbol{X})^T. \tag{5.9}
$$

5.3   Useful identities involving matrix products

In this section we gather a few preliminary results regarding matrix products that will be useful throughout our proofs. Most of these identities are trivial and we thus skip their proof. The only exception is Lemma 5.3 which is proved in Appendix A.

Lemma 5.1 *For two vectors $\boldsymbol{u} \in \mathbb{R}^{m \times 1}$ and $\boldsymbol{v} \in \mathbb{R}^{n \times 1}$, and a matrix $\boldsymbol{M} \in \mathbb{R}^{k \times n}$, we have*

$$
\boldsymbol{u} \otimes \boldsymbol{M}\boldsymbol{v} = \boldsymbol{D}_{\boldsymbol{M}}(\boldsymbol{u} \otimes \boldsymbol{v}),
$$

where $\boldsymbol{D}_{\boldsymbol{M}} \in \mathbb{R}^{mk \times mn}$ is the block diagonal matrix with m copies of $\boldsymbol{M}$ in the diagonal blocks.

Lemma 5.2 *For vectors $\boldsymbol{a} \in \mathbb{R}^{m \times 1}$ and $\boldsymbol{b} \in \mathbb{R}^{n \times 1}$, and a matrix $\boldsymbol{M} \in \mathbb{R}^{n \times m}$, we have*

$$
(\boldsymbol{a} \otimes \boldsymbol{b})^T \text{vec}(\boldsymbol{M}) = \boldsymbol{b}^T \boldsymbol{M}\boldsymbol{a},
$$

where $\text{vec}(M)$ denotes the vectorization of the matrix $\boldsymbol{M}$ formed by stacking the columns of $\boldsymbol{M}$ into a single column vector.

Lemma 5.3 *Consider an arbitrary matrix $\boldsymbol{A} \in \mathbb{R}^{k \times d}$ and an arbitrary vector $\boldsymbol{v} \in \mathbb{R}^k$ and define $\boldsymbol{D}_{\boldsymbol{v}} := \text{diag}(\boldsymbol{v})$. Then the following identity holds*

$$
\|\boldsymbol{v}\|_\infty \sigma_{\min}(\boldsymbol{A}) \le \sigma_{\max}(\boldsymbol{D}_{\boldsymbol{v}}\boldsymbol{A}). \tag{5.10}
$$

Lemma 5.4 *For any two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we have*

$$
(\boldsymbol{A} * \boldsymbol{B})^T (\boldsymbol{A} * \boldsymbol{B}) = (\boldsymbol{A}^T \boldsymbol{A}) \circ (\boldsymbol{B}^T \boldsymbol{B}). \tag{5.11}
$$

5.4 Useful probabilistic identities involving random vectors and matrices

In this Section we gather some useful probabilistic identities that will be used throughout our proofs. We defer the proof of these results to Appendix A. The first result, proven in Appendix A.2, relates the tail of a random variable to a proper Orlicz norm.

Lemma 5.5 *Suppose that a random variable Y satisfies the following tail bound*

$$\mathbb{P}(|Y| \ge t) \le 2e^{-c\min(t^2/A, t/B)}. \tag{5.12}$$

Then, $\|Y\|_{\psi_1} \le 9\max(\sqrt{A/(2c)}, B/c)$.

The second result concerns the minimum eigenvalue of the Khatrio-Rao product of a generic matrix with itself.

Lemma 5.6 *For almost every data input matrix $\boldsymbol{X} \in \mathbb{R}^{d\times n}$, as long as $d \le n \le d(d+1)/2$, the following holds*

$$\sigma_{\min}(\boldsymbol{X} * \boldsymbol{X}) > 0.$$

We refer to Appendix A.3 for the proof of Lemma 5.6.

Finally, we state a Lemma based on [12, 44, 46] that allows us to lower bound the loss function $\mathcal{L}(\boldsymbol{v}, \boldsymbol{W})$ in terms of how close $(\boldsymbol{v}, \boldsymbol{W})$ is to $(\boldsymbol{v}^*, \boldsymbol{W}^*)$. We prove this Lemma in Appendix A.4.

Lemma 5.7 *Let $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ be a symmetric positive semidefinite matrix. Also for $i = 1, 2, \ldots, n$, let $\boldsymbol{x}_i$ be i.i.d. random vectors distributed as $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$. Furthermore, assume*

$$n \ge c(\delta)d\log d,$$

with $c(\delta)$ a constant only depending on δ. Then for all matrices $\boldsymbol{M} \in \mathbb{R}^{d\times d}$ we have

$$\left\| \frac{1}{n}\sum_{i=1}^n \left(\boldsymbol{x}_i^T \boldsymbol{A}\boldsymbol{x}_i\right)\boldsymbol{x}_i\boldsymbol{x}_i^T - (2\boldsymbol{A} + \mathrm{trace}(\boldsymbol{A})\boldsymbol{I}) \right\|_* \le \delta \cdot \|\boldsymbol{A}\|_*,$$

holds with probability at least $1 - 10de^{-\gamma d} - 8/d$. Here, $\|\cdot\|_*$ denotes the nuclear norm i.e. sum of singular values of the input matrix.

6 Proof of global landscape results

In this section we prove the two global theorems.

6.1 Derivative calculations for quadratic activations

We begin by gathering the derivatives of the loss function for quadratic activations. We note that for quadratic activations the loss function (1.4) as a function of $\boldsymbol{W}$ takes the form

$$\begin{aligned}
\mathcal{L}(\boldsymbol{W}) &= \frac{1}{2n}\sum_{i=1}^n \left(y_i - \sum_{\ell=1}^k \boldsymbol{v}_\ell \left|\langle \boldsymbol{w}_\ell, \boldsymbol{x}_i\rangle\right|^2\right)^2, \\
&= \frac{1}{2n}\sum_{i=1}^n \left(\boldsymbol{x}_i^T \boldsymbol{W}^T \mathrm{diag}(\boldsymbol{v})\boldsymbol{W}\boldsymbol{x}_i - y_i\right)^2, \\
&:= \frac{1}{2n}\sum_{i=1}^n r_i^2.
\end{aligned}$$

16

Using (5.3) we have

$$\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{v},\boldsymbol{W}) = 2\boldsymbol{D_v}\boldsymbol{W}\left(\frac{1}{n}\sum_{i=1}^{n}r_i\boldsymbol{x}_i\boldsymbol{x}_i^T\right). \tag{6.1}$$

Using (5.4) the Jacobian matrix is given by

$$\boldsymbol{J} = 2\boldsymbol{D_v}(\boldsymbol{W}\boldsymbol{X}) * \boldsymbol{X}. \tag{6.2}$$

Using this Jacobian matrix we can write the vectorized version of the gradient, i.e. gradient with respect to vect($\boldsymbol{W}$) as

$$\nabla_{\text{vect}(\boldsymbol{W})}\mathcal{L}(\boldsymbol{v},\boldsymbol{W}) = \frac{1}{n}\boldsymbol{J}\boldsymbol{r}. \tag{6.3}$$

Also (5.8) specialized to quadratic activations results in the partial Hessian

$$(\text{vect}(\boldsymbol{U}))^T\nabla_{\boldsymbol{W}}^2\mathcal{L}(\boldsymbol{v},\boldsymbol{W})(\text{vect}(\boldsymbol{U})) = \frac{2}{n}\sum_{i=1}^{n}r_i\left(\boldsymbol{x}_i^T\boldsymbol{U}^T\boldsymbol{D_v}\boldsymbol{U}\boldsymbol{x}_i\right) + \frac{2}{n}\sum_{i=1}^{n}\left(\boldsymbol{x}_i^T\boldsymbol{W}^T\boldsymbol{D_v}\boldsymbol{U}\boldsymbol{x}_i\right)^2. \tag{6.4}$$

6.2 Proof of Theorem 2.1

We begin by proving a simple lemma.

Lemma 6.1 *Any point $\widetilde{\boldsymbol{W}} \in \mathbb{R}^{k \times d}$ obeying*

$$\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{x}_i^T\widetilde{\boldsymbol{W}}^T\text{diag}(\boldsymbol{v})\widetilde{\boldsymbol{W}}\boldsymbol{x}_i - y_i\right)\boldsymbol{x}_i\boldsymbol{x}_i^T = 0, \tag{6.5}$$

is a global optimum of the loss function

$$\mathcal{L}(\boldsymbol{W}) = \frac{1}{2n}\sum_{i=1}^{n}\left(\boldsymbol{x}_i^T\boldsymbol{W}^T\text{diag}(\boldsymbol{v})\boldsymbol{W}\boldsymbol{x}_i - y_i\right)^2.$$

Proof Consider the optimization problem

$$\min_{\boldsymbol{M}\in\mathbb{R}^{d\times d}} f(\boldsymbol{M}) := \frac{1}{2n}\sum_{i=1}^{n}\left(\boldsymbol{x}_i^T\boldsymbol{M}\boldsymbol{x}_i - y_i\right)^2.$$

The objective value is convex in terms of $\boldsymbol{M}$ (in fact its a quadratic). Therefore, at any point where the gradient with respect to $\widetilde{\boldsymbol{M}}$ is zero i.e. $\nabla f(\widetilde{\boldsymbol{M}}) = 0$, that point must be a global optimum of f. More specifically, any point $\widetilde{\boldsymbol{M}}$ obeying

$$\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{x}_i^T\widetilde{\boldsymbol{M}}\boldsymbol{x}_i - y_i\right)\boldsymbol{x}_i\boldsymbol{x}_i^T = 0, \tag{6.6}$$

is a global optimum of $f(\boldsymbol{M})$. That is, for any arbitrary matrix $\boldsymbol{M} \in \mathbb{R}^{d\times d}$ and any point $\widetilde{\boldsymbol{M}} \in \mathbb{R}^{d\times d}$ obeying (6.6) we have

$$f(\boldsymbol{M}) \geq f(\widetilde{\boldsymbol{M}}). \tag{6.7}$$

Now note that if (6.5) holds for some $\widetilde{\boldsymbol{W}}$, then (6.6) holds with $\widetilde{\boldsymbol{M}} = \widetilde{\boldsymbol{W}}^T\text{diag}(\boldsymbol{v})\widetilde{\boldsymbol{W}}$. Thus for any $\boldsymbol{W} \in \mathbb{R}^{k\times d}$, using (6.7) with $\boldsymbol{M} = \boldsymbol{W}^T\text{diag}(\boldsymbol{v})\boldsymbol{W}$ we have

$$\mathcal{L}(\boldsymbol{W}) = f\left(\boldsymbol{W}^T\text{diag}(\boldsymbol{v})\boldsymbol{W}\right) \geq f(\widetilde{\boldsymbol{M}}) = \mathcal{L}(\widetilde{\boldsymbol{W}}).$$

Thus any $\widetilde{\boldsymbol{W}}$ obeying (6.5) is a global optima of $\mathcal{L}(\boldsymbol{W})$, concluding the proof. ∎

6.2.1 Proof of no spurious local minima and strict saddle property

We will prove the first two conclusions of the theorem (i.e. no spurious local optima and saddles have a direction of negative curvature) simultaneously. To show this note that since the function $\mathcal{L}$ is twice differentiable all local optima or saddles that do not have a direction of negative curvature obey

$$\nabla \mathcal{L}(\boldsymbol{W}) = \boldsymbol{0} \quad \text{and} \quad \nabla^2 \mathcal{L}(\boldsymbol{W}) \succeq \boldsymbol{0}. \tag{6.8}$$

We will prove that all points obeying (6.8) are a global optima of $\mathcal{L}(\boldsymbol{W})$. To this aim let us define the sets

$$\mathcal{S}_+ = \{i : \boldsymbol{v}_i > 0\} \quad \text{and} \quad \mathcal{S}_- = \{i : \boldsymbol{v}_i < 0\},$$

i.e. the indices of the positive and negative entries of $\boldsymbol{v}$. Now note that two cases are possible

- **Case I:** At least one of the sub-matrices $(\boldsymbol{D_v W})_{\mathcal{S}_+}$ and $(\boldsymbol{D_v W})_{\mathcal{S}_-}$ is not rank deficient. That is,

$$\text{rank}\,(\boldsymbol{D_v W})_{\mathcal{S}_+} = d \quad \text{or} \quad \text{rank}\,(\boldsymbol{D_v W})_{\mathcal{S}_-} = d. \tag{6.9}$$

- **Case II:** Both of the sub-matrices $(\boldsymbol{D_v W})_{\mathcal{S}_+}$ and $(\boldsymbol{D_v W})_{\mathcal{S}_-}$ are rank deficient. That is,

$$\text{rank}\,\big((\boldsymbol{D_v W})_{\mathcal{S}_+}\big) < d \quad \text{or} \quad \text{rank}\,\big((\boldsymbol{D_v W})_{\mathcal{S}_-}\big) < d. \tag{6.10}$$

In both cases we will show that for any $\boldsymbol{W}$ obeying (6.8), we have

$$\sum_{i=1}^{n} r_i \boldsymbol{x}_i \boldsymbol{x}_i^T = \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{x}_i^T \boldsymbol{W}^T \text{diag}(\boldsymbol{v}) \boldsymbol{W} \boldsymbol{x}_i - y_i \right) \boldsymbol{x}_i \boldsymbol{x}_i^T = 0.$$

The latter together with Lemma 6.1 immediately implies that any $\boldsymbol{W}$ obeying (6.8) is a global optimum, proving that there are no spurious local minima and no saddles which do not have a direction of strict negative curvature. We now proceed to show that indeed $\sum_{i=1}^{n} r_i \boldsymbol{x}_i \boldsymbol{x}_i^T = 0$ holds in both cases mentioned above.

Case I: Note that (6.9) together with the fact that $k \geq d$ implies that the matrix $\boldsymbol{D_v W}$ has a left inverse i.e. there exists a matrix $\boldsymbol{M} \in \mathbb{R}^{d \times k}$ such that $\boldsymbol{M} \boldsymbol{D_v W} = \boldsymbol{I}$. Furthermore, note that by (6.1), $\nabla \mathcal{L}(\boldsymbol{W}) = \boldsymbol{0}$ is equivalent to

$$\boldsymbol{D_v W} \left(\frac{1}{n} \sum_{i=1}^{n} r_i \boldsymbol{x}_i \boldsymbol{x}_i^T \right) = 0.$$

Multiplying both sides of the above equality on the left by $\boldsymbol{M}$ we conclude that

$$\frac{1}{n} \sum_{i=1}^{n} r_i \boldsymbol{x}_i \boldsymbol{x}_i^T = 0,$$

concluding the proof in case I.

Case II: First, note that for any $\boldsymbol{W}$ obeying $\nabla^2 \mathcal{L}(\boldsymbol{W}) \succeq \boldsymbol{0}$ and any $\boldsymbol{U} \in \mathbb{R}^{k \times d}$ by (5.8) we have

$$
\begin{aligned}
0 &\leq \frac{1}{2} \left(\text{vect}(\boldsymbol{U})\right)^T \nabla^2_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{v}, \boldsymbol{W}) \left(\text{vect}(\boldsymbol{U})\right), \\
&= \frac{1}{n} \sum_{i=1}^{n} r_i \left(\boldsymbol{x}_i^T \boldsymbol{U}^T \boldsymbol{D_v} \boldsymbol{U} \boldsymbol{x}_i\right) + \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{x}_i^T \boldsymbol{W}^T \boldsymbol{D_v} \boldsymbol{U} \boldsymbol{x}_i\right)^2.
\end{aligned} \tag{6.11}
$$

Let us choose $\boldsymbol{U}$ of the form $\boldsymbol{U} = \boldsymbol{a}\boldsymbol{b}^T$ with $\boldsymbol{a} \in \mathbb{R}^k$ obeying $\boldsymbol{W}^T \boldsymbol{D_v} \boldsymbol{a} = 0$ (i.e. $\boldsymbol{D_v}\boldsymbol{a} \in \text{Null}(\boldsymbol{W}^T)$) and $\boldsymbol{b} \in \mathbb{R}^d$ an arbitrary vector. Plugging such a $\boldsymbol{U}$ into (6.11) we conclude that

$$
\begin{aligned}
0 &\leq \frac{1}{n} \sum_{i=1}^{n} r_i \left(\boldsymbol{x}_i^T \boldsymbol{U}^T \boldsymbol{D_v} \boldsymbol{U} \boldsymbol{x}_i\right) + \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{x}_i^T \boldsymbol{W}^T \boldsymbol{D_v} \boldsymbol{U} \boldsymbol{x}_i\right)^2, \\
&= (\boldsymbol{a}^T \boldsymbol{D_v} \boldsymbol{a}) \boldsymbol{b}^T \left(\sum_{i=1}^{n} r_i \boldsymbol{x}_i \boldsymbol{x}_i^T\right) \boldsymbol{b} + \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{x}_i^T \boldsymbol{W}^T \boldsymbol{D_v} \boldsymbol{a} \boldsymbol{b}^T \boldsymbol{x}_i\right)^2, \\
&= \left(\boldsymbol{a}^T \boldsymbol{D_v} \boldsymbol{a}\right) \boldsymbol{b}^T \left(\sum_{i=1}^{n} r_i \boldsymbol{x}_i \boldsymbol{x}_i^T\right) \boldsymbol{b},
\end{aligned} \tag{6.12}
$$

holds for all $\boldsymbol{b} \in \mathbb{R}^d$ and all $\boldsymbol{a} \in \mathbb{R}^k$ obeying $\boldsymbol{W}^T \boldsymbol{D_v} \boldsymbol{a} = 0$. Next, we note that by the assumptions of Theorem 2.1 $|\mathcal{S}_+| \geq d$ and $|\mathcal{S}_+| \geq d$. This together with (6.10) implies that there exists non-zero vectors $\boldsymbol{u}, \boldsymbol{w} \in \mathbb{R}^k$ with non-zero entries belonging to $\mathcal{S}_+$ and $\mathcal{S}_-$ such that

$$
\boldsymbol{W}^T \boldsymbol{D_v} \boldsymbol{u} = \boldsymbol{W}^T \boldsymbol{D_v} \boldsymbol{w} = 0.
$$

Furthermore,

$$
\boldsymbol{u}^T \boldsymbol{D_v} \boldsymbol{u} = \sum_{i \in \mathcal{S}_+} v_i u_i^2 > 0.
$$

Thus using (6.12) with $\boldsymbol{a} = \boldsymbol{u}$ we conclude that for all $\boldsymbol{b} \in \mathbb{R}^d$ we have

$$
\boldsymbol{b}^T \left(\sum_{i=1}^{n} r_i \boldsymbol{x}_i \boldsymbol{x}_i^T\right) \boldsymbol{b} \geq 0. \tag{6.13}
$$

Similarly,

$$
\boldsymbol{w}^T \boldsymbol{D_v} \boldsymbol{w} = \sum_{i \in \mathcal{S}_-} v_i w_i^2 < 0.
$$

Thus using (6.12) with $\boldsymbol{a} = \boldsymbol{w}$ we conclude that for all $\boldsymbol{b} \in \mathbb{R}^d$ we have

$$
\boldsymbol{b}^T \left(\sum_{i=1}^{n} r_i \boldsymbol{x}_i \boldsymbol{x}_i^T\right) \boldsymbol{b} \leq 0. \tag{6.14}
$$

Equations (6.13) and (6.14) together imply that

$$
\sum_{i=1}^{n} r_i \boldsymbol{x}_i \boldsymbol{x}_i^T = 0,
$$

concluding the proof in case II.

19

6.2.2 Proof of zero training error

Note that in the previous section we proved that all global optima of $\mathcal{L}(\boldsymbol{W})$ obey

$$\sum_{i=1}^{n} r_i \boldsymbol{x}_i \boldsymbol{x}_i^T = 0.$$

The latter identity can be rewritten in the form

$$(\boldsymbol{X} * \boldsymbol{X}) \boldsymbol{r} = 0. \tag{6.15}$$

Using Lemma 5.6, $\sigma_{\min}(\boldsymbol{X} * \boldsymbol{X}) > 0$ for almost every data input matrix $\boldsymbol{X} \in \mathbb{R}^{d \times n}$, as long as $n \le c d^2$ with c a fixed numerical constant. This combined with (6.15) implies that $\boldsymbol{r} = \boldsymbol{0}$, completing the proof for zero training error.

6.3 Proof of Theorem 2.2

First note that since we assume both $\boldsymbol{v}$ and $\boldsymbol{v}^*$ have the same sign pattern we can without loss of generality assume both $\boldsymbol{v}$ and $\boldsymbol{v}^*$ have non-negative entries. We will first prove that there are no spurious local minima, and all saddles have a direction of negative curvature, and all global optima have zero loss value. We then proceed to show all approximate local minima have small objective value in Section 6.3.2.

6.3.1 Proof of no spurious local minima, zero training error, and strict saddle

We begin by noting that the function $\mathcal{L}$ is twice differentiable and thus all local optima or saddle points that do not have a direction of negative curvature obey

$$\nabla \mathcal{L}(\boldsymbol{W}) = \boldsymbol{0} \quad \text{and} \quad \nabla^2 \mathcal{L}(\boldsymbol{W}) \succeq \boldsymbol{0}. \tag{6.16}$$

Note that $\nabla \mathcal{L}(\boldsymbol{W}) = \boldsymbol{0}$ implies that

$$\boldsymbol{D}_{\boldsymbol{v}} \boldsymbol{W} \left(\sum_{i=1}^{n} r_i \boldsymbol{x}_i \boldsymbol{x}_i^T \right) = \boldsymbol{0}. \tag{6.17}$$

Define the set

$$\mathcal{S} = \{i : v_i > 0\}.$$

Since we assume $k \ge d$, two cases can occur.

- **Case I:** $\mathrm{rank}((\boldsymbol{D}_{\boldsymbol{v}} \boldsymbol{W})_{\mathcal{S}}) = d$.

- **Case II:** $\mathrm{rank}((\boldsymbol{D}_{\boldsymbol{v}} \boldsymbol{W})_{\mathcal{S}}) < d$.

In both cases we will show that any $\boldsymbol{W}$ obeying (6.8), must be a global optimum.
Case I: $\mathrm{rank}((\boldsymbol{D}_{\boldsymbol{v}} \boldsymbol{W})_{\mathcal{S}}) = d$.
In this case the matrix $\boldsymbol{D}_{\boldsymbol{v}} \boldsymbol{W}$ has a left inverse, i.e. there exists a $\boldsymbol{M} \in \mathbb{R}^{d \times k}$ such that $\boldsymbol{M} \boldsymbol{D}_{\boldsymbol{v}} \boldsymbol{W} = \boldsymbol{I}$. Multiplying both sides of (6.17) on the left by $\boldsymbol{M}$ yields

$$\sum_{i=1}^{n} r_i \boldsymbol{x}_i \boldsymbol{x}_i^T = (\boldsymbol{X} * \boldsymbol{X}) \boldsymbol{r} = \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{x}_i^T \boldsymbol{W}^T \mathrm{diag}(\boldsymbol{v}) \boldsymbol{W} \boldsymbol{x}_i - y_i \right) \boldsymbol{x}_i \boldsymbol{x}_i^T = 0. \tag{6.18}$$

The latter together with Lemma 6.1 immediately implies that any $\boldsymbol{W}$ obeying (6.16) and the condition of Case I is a global optimum, proving that there are no spurious local minima. Furthermore using Lemma 5.6, $\sigma_{\min}(\boldsymbol{X} * \boldsymbol{X}) > 0$ as long as $n \le cd^2$ with c a fixed numerical constant. This combined with (6.18) implies that $\boldsymbol{r} = \boldsymbol{0}$, completing the proof for zero training error in this case.

Case II: $\text{rank}((\boldsymbol{D_v W})_{\mathcal{S}}) < d$.

First, note that for any $\boldsymbol{W}$ obeying $\nabla^2 \mathcal{L}(\boldsymbol{W}) \ge \boldsymbol{0}$ and any $\boldsymbol{U} \in \mathbb{R}^{k \times d}$ by (5.8) we have

$$
\begin{aligned}
0 &\le \frac{1}{2} \left(\text{vect}(\boldsymbol{U})\right)^T \nabla_{\boldsymbol{W}}^2 \mathcal{L}(\boldsymbol{v}, \boldsymbol{W}) \left(\text{vect}(\boldsymbol{U})\right), \\
&= \frac{1}{n} \sum_{i=1}^{n} r_i \left(\boldsymbol{x}_i^T \boldsymbol{U}^T \boldsymbol{D_v} \boldsymbol{U} \boldsymbol{x}_i\right) + \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{x}_i^T \boldsymbol{W}^T \boldsymbol{D_v} \boldsymbol{U} \boldsymbol{x}_i\right)^2 .
\end{aligned} \tag{6.19}
$$

In this case since $(\mathcal{D_v})_{\mathcal{S}}$ is invertible and $(\boldsymbol{D_v W})_{\mathcal{S}}$ is rank deficient there exists a nonzero vector $\boldsymbol{a} \in \mathbb{R}^k$ supported on $\mathcal{S}$ obeying $\boldsymbol{W}^T \boldsymbol{D_v a} = 0$ (i.e. $\boldsymbol{D_v a} \in \text{Null}(\boldsymbol{W}^T)$). Thus plugging $\boldsymbol{U}$ of the form $\boldsymbol{U} = \boldsymbol{ab}^T$ with $\boldsymbol{b} \in \mathbb{R}^d$ in (6.19) we conclude that

$$
\begin{aligned}
0 &\le \frac{1}{n} \sum_{i=1}^{n} r_i \left(\boldsymbol{x}_i^T \boldsymbol{U}^T \boldsymbol{D_v} \boldsymbol{U} \boldsymbol{x}_i\right) + \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{x}_i^T \boldsymbol{W}^T \boldsymbol{D_v} \boldsymbol{U} \boldsymbol{x}_i\right)^2, \\
&= (\boldsymbol{a}^T \boldsymbol{D_v a}) \boldsymbol{b}^T \left(\sum_{i=1}^{n} r_i \boldsymbol{x}_i \boldsymbol{x}_i^T\right) \boldsymbol{b} + \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{x}_i^T \boldsymbol{W}^T \boldsymbol{D_v} \boldsymbol{ab}^T \boldsymbol{x}_i\right)^2, \\
&= \left(\boldsymbol{a}^T \boldsymbol{D_v a}\right) \boldsymbol{b}^T \left(\sum_{i=1}^{n} r_i \boldsymbol{x}_i \boldsymbol{x}_i^T\right) \boldsymbol{b},
\end{aligned} \tag{6.20}
$$

holds for all $\boldsymbol{b} \in \mathbb{R}^d$. Also note that

$$
\boldsymbol{a}^T \boldsymbol{D_v a} = \sum_{i \in \mathcal{S}} v_i a_i^2 > 0.
$$

The latter together with (6.20) implies that

$$
\sum_{i=1}^{n} r_i \boldsymbol{x}_i \boldsymbol{x}_i^T \ge \boldsymbol{0}. \tag{6.21}
$$

Thus for any $\boldsymbol{W}$ obeying (6.16) and the condition of Case II we have

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{W}) &= \frac{1}{2n} \langle \boldsymbol{W}^T \boldsymbol{D_v W} - \boldsymbol{W}^{*T} \boldsymbol{D_{v^*}} \boldsymbol{W}^*, \sum_{i=1}^{n} r_i \boldsymbol{x}_i \boldsymbol{x}_i^T \rangle, \\
&\stackrel{(a)}{=} -\frac{1}{2n} \langle \boldsymbol{W}^{*T} \boldsymbol{D_{v^*}} \boldsymbol{W}^*, \sum_{i=1}^{n} r_i \boldsymbol{x}_i \boldsymbol{x}_i^T \rangle, \\
&\stackrel{(b)}{\le} 0.
\end{aligned}
$$

where (a) follows from (6.17) and (b) follows from (6.21) combined with the fact that $\boldsymbol{v}^*$ has nonnegative entries. Thus, $\mathcal{L}(\boldsymbol{W}) = 0$ and therefore such a $\boldsymbol{W}$ must be a global optimum. This completes the proof in Case II.

6.3.2 Bounding the objective value for approximate local minima

Assume $\boldsymbol{W}$ is an approximate local minima. That is it obeys

$$\|\nabla\mathcal{L}(\boldsymbol{W})\|_F \le \epsilon_g \quad \text{and} \quad \nabla^2\mathcal{L}(\boldsymbol{W}_s) \ge -\epsilon_H \boldsymbol{I}. \tag{6.22}$$

For such a point by Cauchy-Schwarz we have

$$\langle \boldsymbol{W}, \nabla\mathcal{L}(\boldsymbol{W}) \rangle \le \epsilon_g \|\boldsymbol{W}\|_F. \tag{6.23}$$

Furthermore, for such a point using (5.8)

$$-\epsilon_H \|\boldsymbol{U}\|_F^2 \le \frac{1}{2} \left(\text{vect}(\boldsymbol{U})\right)^T \nabla_{\boldsymbol{W}}^2 \mathcal{L}(\boldsymbol{v}, \boldsymbol{W}) \left(\text{vect}(\boldsymbol{U})\right),$$
$$= \frac{1}{n} \sum_{i=1}^n r_i \left(\boldsymbol{x}_i^T \boldsymbol{U}^T \boldsymbol{D}_{\boldsymbol{v}} \boldsymbol{U} \boldsymbol{x}_i\right) + \frac{1}{n} \sum_{i=1}^n \left(\boldsymbol{x}_i^T \boldsymbol{W}^T \boldsymbol{D}_{\boldsymbol{v}} \boldsymbol{U} \boldsymbol{x}_i\right)^2, \tag{6.24}$$

holds for all $\boldsymbol{U} \in \mathbb{R}^{k \times d}$. Since $\boldsymbol{v} = \nu\boldsymbol{1}$ there exists a unit norm, non-zero vector $\boldsymbol{a} \in \mathbb{R}^k$ supported on $\mathcal{S}$ such that $\boldsymbol{W}^T \boldsymbol{D}_{\boldsymbol{v}} \boldsymbol{a} = 0$. We now set $\boldsymbol{U} = \boldsymbol{a}\boldsymbol{b}^T$ in (6.24) with $\boldsymbol{b} \in \mathbb{R}^d$. We conclude that

$$\left(\boldsymbol{a}^T \boldsymbol{D}_{\boldsymbol{v}} \boldsymbol{a}\right) \left(\boldsymbol{b}^T \left(\frac{1}{n} \sum_{i=1}^n r_i \boldsymbol{x}_i \boldsymbol{x}_i^T\right) \boldsymbol{b}\right) \ge -\epsilon_H \|\boldsymbol{b}\|_{\ell_2}^2. \tag{6.25}$$

holds for all $\boldsymbol{b} \in \mathbb{R}^d$. Also note that

$$\boldsymbol{a}^T \boldsymbol{D}_{\boldsymbol{v}} \boldsymbol{a} = \nu \|\boldsymbol{a}\|_{\ell_2}^2 = \nu.$$

Thus, (6.25) implies that

$$\frac{1}{n} \sum_{i=1}^n r_i \boldsymbol{x}_i \boldsymbol{x}_i^T \ge -\frac{\epsilon_H}{\nu} \boldsymbol{I}. \tag{6.26}$$

Now note that the gradient is equal to

$$\nabla\mathcal{L}(\boldsymbol{W}) = \boldsymbol{D}_{\boldsymbol{v}} \boldsymbol{W} \left(\sum_{i=1}^n r_i \boldsymbol{x}_i \boldsymbol{x}_{.i}^T\right). \tag{6.27}$$

Thus using (6.27), (6.23), and (6.26) we conclude that for any point $\boldsymbol{W}$ obeying (6.22) we have

$$\mathcal{L}(\boldsymbol{W}) = \frac{1}{2n} \langle \boldsymbol{W}^T \boldsymbol{D}_{\boldsymbol{v}} \boldsymbol{W} - \boldsymbol{W}^{*T} \boldsymbol{D}_{\boldsymbol{v}^*} \boldsymbol{W}^*, \sum_{i=1}^n r_i \boldsymbol{x}_i \boldsymbol{x}_i^T \rangle,$$
$$\overset{(a)}{=} \langle \boldsymbol{W}, \nabla\mathcal{L}(\boldsymbol{W}) \rangle - \frac{1}{2n} \langle \boldsymbol{W}^{*T} \boldsymbol{D}_{\boldsymbol{v}^*} \boldsymbol{W}^*, \sum_{i=1}^n r_i \boldsymbol{x}_i \boldsymbol{x}_i^T \rangle,$$
$$\overset{(b)}{\le} \epsilon_g \|\boldsymbol{W}\|_F - \frac{1}{2} \langle \boldsymbol{W}^{*T} \boldsymbol{D}_{\boldsymbol{v}^*} \boldsymbol{W}^*, \frac{1}{n} \sum_{i=1}^n r_i \boldsymbol{x}_i \boldsymbol{x}_i^T \rangle,$$
$$\overset{(c)}{\le} \epsilon_g \|\boldsymbol{W}\|_F + \frac{\epsilon_H}{2\nu} \|\boldsymbol{W}^{*T} \boldsymbol{D}_{\boldsymbol{v}^*} \boldsymbol{W}^*\|_*. \tag{6.28}$$

Here, (a) follows from (6.27), (b) from (6.23) and (c) from Holder's inequality combined with (6.26).

We proceed by showing that for a point $\boldsymbol{W}$ obeying $\|\nabla\mathcal{L}(\boldsymbol{W})\|_F \le \epsilon_g$ it's Frobenius norm norm is also sufficiently small. To this aim note that for such a point by Cauchy-Schwarz have

$$\langle \nabla\mathcal{L}(\boldsymbol{W}), \boldsymbol{W}\rangle \le \Delta \|\boldsymbol{W}\|_F. \tag{6.29}$$

Also

$$
\begin{aligned}
\langle \nabla\mathcal{L}(\boldsymbol{W}), \boldsymbol{W}\rangle &= \frac{1}{n}\langle \boldsymbol{D_v}\boldsymbol{W}\sum_i \left(\boldsymbol{x}_i^T\left(\boldsymbol{W}^T\boldsymbol{D_v}\boldsymbol{W} - \boldsymbol{W}^{*T}\boldsymbol{D_{v^*}}\boldsymbol{W}^*\right)\boldsymbol{x}_i\right)\boldsymbol{x}_i\boldsymbol{x}_i^T, \boldsymbol{W}\rangle \\
&= \frac{1}{n}\sum_{i=1}^n \left(\boldsymbol{x}_i^T\boldsymbol{W}^T\boldsymbol{D_v}\boldsymbol{W}\boldsymbol{x}_i\right)^2 - \frac{1}{n}\sum_{i=1}^n \left(\boldsymbol{x}_i^T\boldsymbol{W}^T\boldsymbol{D_v}\boldsymbol{W}\boldsymbol{x}_i\right)\left(\boldsymbol{x}_i^T\boldsymbol{W}^{*T}\boldsymbol{D_{v^*}}\boldsymbol{W}^*\boldsymbol{x}_i\right) \\
&\ge \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i^T\boldsymbol{W}^T\boldsymbol{D_v}\boldsymbol{W}\boldsymbol{x}_i\right)^2 - \frac{1}{n}\sum_{i=1}^n \left(\boldsymbol{x}_i^T\boldsymbol{W}^T\boldsymbol{D_v}\boldsymbol{W}\boldsymbol{x}_i\right)\left(\boldsymbol{x}_i^T\boldsymbol{W}^{*T}\boldsymbol{D_{v^*}}\boldsymbol{W}^*\boldsymbol{x}_i\right). \quad (6.30)
\end{aligned}
$$

Using Holder's inequality for matrices we have

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i^T\boldsymbol{W}^T\boldsymbol{D_v}\boldsymbol{W}\boldsymbol{x}_i - \operatorname{trace}\left(\boldsymbol{W}^T\boldsymbol{D_v}\boldsymbol{W}\right) &= \langle \boldsymbol{W}^T\boldsymbol{D_v}\boldsymbol{W}, \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T - \boldsymbol{I}\rangle \\
&\le \operatorname{trace}(\boldsymbol{W}^T\boldsymbol{D_v}\boldsymbol{W})\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T - \boldsymbol{I}\right\| \\
&= \operatorname{trace}(\boldsymbol{W}^T\boldsymbol{D_v}\boldsymbol{W})\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T - \boldsymbol{I}\right\|.
\end{aligned}
$$

Thus by standard concentration of sample covariance matrices we conclude that for any $\delta > 0$, as long as $n \ge \frac{c}{\delta^2}d$ for a fixed numerical constant c, then for all $\boldsymbol{W}$

$$\left|\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i^T\boldsymbol{W}^T\boldsymbol{D_v}\boldsymbol{W}\boldsymbol{x}_i - \operatorname{trace}\left(\boldsymbol{W}^T\boldsymbol{D_v}\boldsymbol{W}\right)\right| \le \delta \cdot \operatorname{trace}(\boldsymbol{W}^T\boldsymbol{D_v}\boldsymbol{W}),$$

holds with probability at least $1 - 2e^{-bd}$, for a fixed constant $b > 0$. Thus with high probability for all $\boldsymbol{W}$ we have

$$\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i^T\boldsymbol{W}^T\boldsymbol{D_v}\boldsymbol{W}\boldsymbol{x}_i \ge (1-\delta)\operatorname{trace}(\boldsymbol{W}^T\boldsymbol{D_v}\boldsymbol{W}).$$

Plugging the latter into (6.30) we conclude that

$$\langle \nabla\mathcal{L}(\boldsymbol{W}), \boldsymbol{W}\rangle \ge (1-\delta)^2\left(\operatorname{trace}(\boldsymbol{W}^T\boldsymbol{D_v}\boldsymbol{W})\right)^2 - \frac{1}{n}\sum_{i=1}^n \left(\boldsymbol{x}_i^T\boldsymbol{W}^T\boldsymbol{D_v}\boldsymbol{W}\boldsymbol{x}_i\right)\left(\boldsymbol{x}_i^T\boldsymbol{W}^{*T}\boldsymbol{D_{v^*}}\boldsymbol{W}^*\boldsymbol{x}_i\right). \tag{6.31}$$

We now apply Lemma 5.7 to conclude that as long as $n \ge c(\delta)d\log d$

$$\left\|\frac{1}{n}\sum_{i=1}^n \left(\boldsymbol{x}_i^T\boldsymbol{W}^{*T}\boldsymbol{D_{v^*}}\boldsymbol{W}^*\boldsymbol{x}_i\right)\boldsymbol{x}_i\boldsymbol{x}_i^T - \left(2\boldsymbol{W}^{*T}\boldsymbol{D_{v^*}}\boldsymbol{W}^* + \operatorname{trace}\left(\boldsymbol{W}^{*T}\boldsymbol{D_{v^*}}\boldsymbol{W}^*\right)\boldsymbol{I}\right)\right\|$$

$$\le \delta \cdot \operatorname{trace}\left(\boldsymbol{W}^{*T}\boldsymbol{D_{v^*}}\boldsymbol{W}^*\right),$$

holds with probability at least $1 - 10de^{-bd} - 8/d$. The latter implies that with high probability for all $\boldsymbol{W}$ we have

$$\left| \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{x}_i^T \boldsymbol{W}^T \boldsymbol{D_v} \boldsymbol{W} \boldsymbol{x}_i \right) \left(\boldsymbol{x}_i^T \boldsymbol{W}^{*T} \boldsymbol{D_{v^*}} \boldsymbol{W}^* \boldsymbol{x}_i \right) \right.$$
$$\left. - \left(2 \langle \boldsymbol{W}^T \boldsymbol{D_v} \boldsymbol{W}, \boldsymbol{W}^{*T} \boldsymbol{D_{v^*}} \boldsymbol{W}^* \rangle + \operatorname{trace}\left(\boldsymbol{W}^{*T} \boldsymbol{D_{v^*}} \boldsymbol{W}^* \right) \operatorname{trace}\left(\boldsymbol{W}^T \boldsymbol{D_v} \boldsymbol{W} \right) \right) \right|$$
$$\leq \delta \cdot \operatorname{trace}\left(\boldsymbol{W}^{*T} \boldsymbol{D_{v^*}} \boldsymbol{W}^* \right) \operatorname{trace}\left(\boldsymbol{W}^T \boldsymbol{D_v} \boldsymbol{W} \right).$$

Plugging the latter into (6.31) and we conclude that with high probability for all $\boldsymbol{W}$ we have

$$\langle \nabla \mathcal{L}(\boldsymbol{W}), \boldsymbol{W} \rangle \geq (1-\delta)^2 \left(\operatorname{trace}(\boldsymbol{W}^T \boldsymbol{D_v} \boldsymbol{W}) \right)^2 - \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{x}_i^T \boldsymbol{W}^T \boldsymbol{D_v} \boldsymbol{W} \boldsymbol{x}_i \right) \left(\boldsymbol{x}_i^T \boldsymbol{W}^{*T} \boldsymbol{D_{v^*}} \boldsymbol{W}^* \boldsymbol{x}_i \right).$$
$$\geq (1-\delta)^2 \left(\operatorname{trace}(\boldsymbol{W}^T \boldsymbol{D_v} \boldsymbol{W}) \right)^2 - \delta \cdot \operatorname{trace}\left(\boldsymbol{W}^{*T} \boldsymbol{D_{v^*}} \boldsymbol{W}^* \right) \operatorname{trace}\left(\boldsymbol{W}^T \boldsymbol{D_v} \boldsymbol{W} \right)$$
$$- \left(2 \langle \boldsymbol{W}^T \boldsymbol{D_v} \boldsymbol{W}, \boldsymbol{W}^{*T} \boldsymbol{D_{v^*}} \boldsymbol{W}^* \rangle + \operatorname{trace}\left(\boldsymbol{W}^{*T} \boldsymbol{D_{v^*}} \boldsymbol{W}^* \right) \operatorname{trace}\left(\boldsymbol{W}^T \boldsymbol{D_v} \boldsymbol{W} \right) \right)$$
$$\geq (1-\delta)^2 \left(\operatorname{trace}(\boldsymbol{W}^T \boldsymbol{D_v} \boldsymbol{W}) \right)^2 - (3+\delta) \cdot \operatorname{trace}\left(\boldsymbol{W}^{*T} \boldsymbol{D_{v^*}} \boldsymbol{W}^* \right) \operatorname{trace}\left(\boldsymbol{W}^T \boldsymbol{D_v} \boldsymbol{W} \right).$$

Plugging the latter into (6.29) implies that

$$(1-\delta)^2 \left(\operatorname{trace}(\boldsymbol{W}^T \boldsymbol{D_v} \boldsymbol{W}) \right)^2 - (3+\delta) \cdot \operatorname{trace}\left(\boldsymbol{W}^{*T} \boldsymbol{D_{v^*}} \boldsymbol{W}^* \right) \operatorname{trace}\left(\boldsymbol{W}^T \boldsymbol{D_v} \boldsymbol{W} \right) \leq \epsilon_g \|\boldsymbol{W}\|_F.$$

Using the assumption that $\boldsymbol{v} = \nu \boldsymbol{1}$, the latter is equivalent to

$$\|\boldsymbol{W}\|_F \left((1-\delta)^2 \nu^2 \|\boldsymbol{W}\|_F^2 - \nu(3+\delta)\operatorname{trace}\left(\boldsymbol{W}^{*T} \boldsymbol{D_{v^*}} \boldsymbol{W}^* \right) \right)$$
$$= (1-\delta)^2 \nu^2 \|\boldsymbol{W}\|_F^3 - (3+\delta)\nu \|\boldsymbol{W}\|_F \operatorname{trace}\left(\boldsymbol{W}^{*T} \boldsymbol{D_{v^*}} \boldsymbol{W}^* \right) \leq \epsilon_g.$$

From the latter we can conclude that for all $a > 0$ we have

$$\|\boldsymbol{W}\|_F \leq \max\left(\sqrt{\frac{\epsilon_g}{a\nu^2(1-\delta)^2} + \frac{3+\delta}{\nu(1-\delta)^2}\operatorname{trace}\left(\boldsymbol{W}^{*T} \boldsymbol{D_{v^*}} \boldsymbol{W}^* \right)}, a \right)$$

Setting $a = \frac{\epsilon_g}{\nu(1-\delta)^2}$ and $\delta = 1/2$, the latter inequality implies that

$$\|\boldsymbol{W}\|_F \leq \max\left(\frac{1}{\sqrt{\nu}}\sqrt{1 + 14\operatorname{trace}\left(\boldsymbol{W}^{*T} \boldsymbol{D_{v^*}} \boldsymbol{W}^* \right)}, \frac{4}{\nu}\epsilon_g \right).$$

Plugging the latter into (6.28) we conclude that for all $\boldsymbol{W}$ obeying (6.22)

$$\mathcal{L}(\boldsymbol{W}) \leq \epsilon_g \|\boldsymbol{W}\|_F + \frac{\epsilon_H}{2\nu} \left\| \boldsymbol{W}^{*T} \boldsymbol{D_{v^*}} \boldsymbol{W}^* \right\|_*,$$
$$\leq \frac{\epsilon_g}{\sqrt{\nu}} \max\left(\sqrt{1 + 14 \left\| \boldsymbol{W}^{*T} \boldsymbol{D_{v^*}} \boldsymbol{W}^* \right\|_*}, 4\frac{\epsilon_g}{\sqrt{\nu}} \right) + \frac{\epsilon_H}{2\nu} \left\| \boldsymbol{W}^{*T} \boldsymbol{D_{v^*}} \boldsymbol{W}^* \right\|_*,$$

holds with high probability, concluding the proof.

24

7 Proof of local convergence results

In this section we prove Theorem 2.5. We begin by explaining a crucial component of our proof which involves bounding the spectrum of the Jacobian matrix. These results are the subject of the next section with the corresponding proofs appearing in Section 7.2. We then utilize these intermediate results to finalize the proof of Theorem 2.5 in Section 7.3.

7.1 Key lemmas on spectrum of the Jacobian Matrix

We start with a few definitions. As defined previously in Assumption 2.3 for $\sigma \in \mathbb{R}$ and $g \sim \mathcal{N}(0, \sigma^2)$ we define $\mu_\phi(\sigma) := \mathbb{E}[\phi'(g)]$, and let

$$\boldsymbol{\mu} = \left(\mu_\phi(\|\boldsymbol{w}_1\|_{\ell_2}), \mu_\phi(\|\boldsymbol{w}_2\|_{\ell_2}), \ldots, \mu_\phi(\|\boldsymbol{w}_k\|_{\ell_2}) \right). \tag{7.1}$$

Similarly,

$$\gamma_\phi(\sigma) := \frac{1}{\sigma^2} \mathbb{E}[\phi'(g)g],$$

and note that by Stein's lemma we have

$$\gamma_\phi(\sigma) = \frac{1}{\sigma^2} \mathbb{E}[\phi'(g)g] = \mathbb{E}[\phi''(g)].$$

Also define

$$\boldsymbol{\Gamma} = \text{diag} \left(\gamma_\phi(\|\boldsymbol{w}_1\|_{\ell_2}), \gamma_\phi(\|\boldsymbol{w}_2\|_{\ell_2}), \ldots, \gamma_\phi(\|\boldsymbol{w}_k\|_{\ell_2}) \right). \tag{7.2}$$

Recall that the Jacobian matrix is given by $\boldsymbol{J} = \boldsymbol{D_v}\phi'(\boldsymbol{WX}) * \boldsymbol{X}$. To bound the spectrum of the Jacobian we first introduce a new matrix obtained by centering the $\phi'(\boldsymbol{WX})$ component. Precisely, we let

$$\tilde{\boldsymbol{J}} = \boldsymbol{X} * \boldsymbol{D_v}(\phi'(\boldsymbol{WX}) - \mathbb{E}[\phi'(\boldsymbol{WX})]). \tag{7.3}$$

Our first result, proven in in Appendix B, relates the spectrum of $\boldsymbol{J}$ to that of $\tilde{\boldsymbol{J}}$. Note that in case of $\mu(\sigma) = 0$ everywhere (See Assumption 2.3), then $\mathbb{E}[\phi'(\boldsymbol{WX})] = 0$ and $\boldsymbol{J} = \tilde{\boldsymbol{J}}$, up to a permutation of the rows. Therefore, this step becomes superfluous in this case.

Proposition 7.1 *Let $\phi : \mathbb{R} \to \mathbb{R}$ be a nonlinear activation obeying $|\phi''| \le L$. Assume the inputs $\boldsymbol{x}_i \in \mathbb{R}^d$ are distributed i.i.d. $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ for $i = 1, 2, \ldots, n$.*

- *There exists an absolute constant $c > 0$ such that for $n \le cd^2(L \|\boldsymbol{v}\|_{\ell_\infty} \|\boldsymbol{W}\|)^4$, we have*

$$\sigma_{\min}^2(\boldsymbol{J}) \ge \sigma_{\min}^2(\tilde{\boldsymbol{J}}) - \left(4\sqrt{3}n + Cd^{\frac{3}{2}} + C\sqrt{dn} \right) L^2 \|\boldsymbol{v}\|_{\ell_\infty}^2 \|\boldsymbol{W}\|^2, \tag{7.4}$$

  with probability at least $1 - 2e^{-c'\sqrt{d}}$, where $c', C > 0$ are constants (depending on c).

- *In addition, with probability at least $1 - 2e^{-n/2}$, we have*

$$\sigma_{\max}(\boldsymbol{J}) \le \sigma_{\max}(\tilde{\boldsymbol{J}}) + 3\sqrt{n} \|\boldsymbol{D_v}\boldsymbol{\mu}\|_{\ell_2}, \tag{7.5}$$

Now that we have established a connection between the spectrum of $\boldsymbol{J}$ and $\tilde{\boldsymbol{J}}$, we focus on bounding $\sigma_{\min}(\tilde{\boldsymbol{J}})$ and $\sigma_{\max}(\tilde{\boldsymbol{J}})$. The following proposition is at the core of our analysis and may be of independent interest. We defer the proof of this result to Section 7.2.

Proposition 7.2 *Let $\phi : \mathbb{R} \to \mathbb{R}$ be a general activation obeying $|\phi''| \leq L$. Assume the inputs $\boldsymbol{x}_i \in \mathbb{R}^d$ are distributed i.i.d. $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ for $i \in [n]$. Suppose that $n \geq d$ and $k \geq d$. Furthermore, assume*

$$\sqrt{n} \leq c \frac{\sigma_{\min}^2 (\boldsymbol{\Gamma W})}{L^4 \sigma_{\max}^4 (\boldsymbol{W}) + 1} d, \tag{7.6}$$

holds for a sufficiently small constant $c > 0$. Then, there exist constants $C > 0$, such that,

$$\sigma_{\min}(\tilde{\boldsymbol{J}}) \geq \frac{d}{2} \sigma_{\min}(\boldsymbol{D_v \Gamma W}), \tag{7.7}$$

$$\sigma_{\max}(\tilde{\boldsymbol{J}}) \leq C \left(d + \sqrt{nd} \, \sigma_{\max}(\boldsymbol{D_v \Gamma W}) \right), \tag{7.8}$$

holds with probability at least $1 - ne^{-b_1\sqrt{n}} - n^{-1} - 2ne^{-b_2 d}$ with $b_1, b_2 > 0$ fixed numerical constants.

Remark 7.3 *As discussed in the proof of Proposition 7.2, when Assumption 2.3 (b) holds in lieu of Assumption 2.3 (a), then $\boldsymbol{\Gamma} = \boldsymbol{0}$. In this case the statement of the above proposition must be modified as follows. Equation (7.6) should be replaced with*

$$\sqrt{n} \leq c \frac{1}{L^4 \sigma_{\max}^4 (\boldsymbol{W}) + 1} d. \tag{7.9}$$

Further, we have the following bounds in lieu of equations (7.7) and (7.8).

$$\sigma_{\min}(\tilde{\boldsymbol{J}}) \geq \frac{d}{2}.$$
$$\sigma_{\max}(\tilde{\boldsymbol{J}}) \leq Cd.$$

(See Equations (7.35) and (7.40)).

We now combine Propositions 7.1 and 7.2 to characterize the spectrum of $\boldsymbol{J}$. We defer the proof of this result to Appendix C.

Proposition 7.4 *Let $\phi : \mathbb{R} \to \mathbb{R}$ be a general activation obeying Assumption 2.3 (a) and $|\phi''| \leq L$. Assume the inputs $\boldsymbol{x}_i \in \mathbb{R}^d$ are distributed i.i.d. $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ for $i \in [n]$. Suppose that*

$$0 < v_{\min} \leq |v_\ell^*| \leq v_{\max} \quad 0 < w_{\min} \leq \|\boldsymbol{w}_\ell^*\|_{\ell_2} \leq w_{\max} \quad \text{for } \ell = 1, 2, \ldots, k, \tag{7.10}$$

for some fixed constants $v_{\min}, v_{\max}, w_{\min}, w_{\max}$. Further, assume that $k \geq d$ and

$$d \leq n \leq \frac{c_0 \sigma_{\min}^4 (\boldsymbol{W})}{\sigma_{\max}^8 (\boldsymbol{W})} d^2 \tag{7.11}$$

for a sufficiently small constant $c_0 > 0$. Then, there exist constants $C \geq c > 0$, such that,

$$\sigma_{\min}(\boldsymbol{J}) \geq c \, \sigma_{\min}(\boldsymbol{W}) d, \tag{7.12}$$

$$\sigma_{\max}(\boldsymbol{J}) \leq C \, \sigma_{\max}(\boldsymbol{W}) \sqrt{nk}, \tag{7.13}$$

holds with probability at least $1 - n^{-1} - 2ne^{-b\sqrt{d}}$ with $b > 0$ a fixed numerical constant.

Also when Assumption 2.3 (b) holds in lieu of Assumption 2.3 (a), the term $\sigma_{\min}(\boldsymbol{W})$ should be replaced by one in Equations (7.11) and (7.12).

A noteworthy case that will be used in our analysis for both local and global convergence results is $\phi(z) = z^2/2$. Note that for this choice of ϕ, we have $\mathbb{E}[\phi'(\boldsymbol{WX})] = 0$ and therefore the centering step (7.4) becomes superfluous as $\tilde{\boldsymbol{J}} = \boldsymbol{J}$. Further, $\boldsymbol{\Gamma} = \boldsymbol{I}$. Applying Proposition 7.4 to this case with $\boldsymbol{W} = \boldsymbol{I}$ and $\boldsymbol{v} = (1, 1, \ldots, 1, 1)$, we obtain the following result.

Corollary 7.5 *Let $\boldsymbol{X} \in \mathbb{R}^{d \times n}$ be a matrix with i.i.d. $\mathcal{N}(0, 1)$ entries. Suppose that $d \le n \le c_1 d^2$, for a sufficiently small constant $c_1 > 0$. Then, there exist constants $C \ge c > 0$, such that,*

$$\sigma_{\min}(\boldsymbol{X} * \boldsymbol{X}) \ge cd, \tag{7.14}$$

$$\sigma_{\max}(\boldsymbol{X} * \boldsymbol{X}) \le C\sqrt{nd}, \tag{7.15}$$

holds with probability at least $1 - ne^{-b_1\sqrt{n}} - n^{-1} - 2ne^{-b_2 d}$ with $b_1, b_2 > 0$ fixed numerical constants.

7.2   Proof of Proposition 7.2

Bounding $\sigma_{\min}(\tilde{\boldsymbol{J}})$ is involved and requires several technical lemmas that are of independent interest. We first present an outline of our approach and then discuss the steps in details. We then focus on bounding $\sigma_{\max}(\tilde{\boldsymbol{J}})$, which follows along the same lines.
Our proof strategy consists of four main steps:

1. *(Whitening).* We let $\boldsymbol{M} \in \mathbb{R}^{d \times k}$ be the left inverse of $\boldsymbol{D_v}\boldsymbol{\Gamma}\boldsymbol{W}$ and construct a block-diagonal matrix $\boldsymbol{D_M}$ with d copies of $\boldsymbol{M}$ on its diagonal. We construct the whitened matrix $\boldsymbol{D_M}\tilde{\boldsymbol{J}} \in \mathbb{R}^{d^2 \times n}$. The reason this operation is useful is that it acts as a partial centering of the entries of $\tilde{\boldsymbol{J}}$ so that almost all entries of the resulting matrix have zero mean.

2. *(dropping rows)* Let us index the rows of $\boldsymbol{D_M}\tilde{\boldsymbol{J}} \in \mathbb{R}^{d^2 \times n}$ by (i, j) for $1 \le i \le j \le d$. We next construct the matrix $\tilde{\boldsymbol{J}}^c$ which is obtained from $\boldsymbol{D_M}\tilde{\boldsymbol{J}}$ by dropping rows indexed by (i, i). The reason this operation is useful is that it removes the entries of $\tilde{\boldsymbol{J}}$ that are not centered. Indeed, the notation $\tilde{\boldsymbol{J}}^c$ is used to cue that the columns of $\tilde{\boldsymbol{J}}^c$ are centered (have zero mean).

3. *(Bounding sub-exponential norms).* In this step we show that the columns of $\tilde{\boldsymbol{J}}^c$ have bounded sub-exponential norm.

4. *(Bounding singular values).* We prove bounds on singular values of a matrix with independent columns and bounded sub-exponential norms.

Steps one and two collectively act as a nontrivial "centering" of the columns of $\tilde{\boldsymbol{J}}$. The reason this centering is required is that the columns of the matrix $\tilde{\boldsymbol{J}}$ have non-zero mean which lead to rather large sub-exponential norms. By centering the columns we are able to reduce the sub-exponential norm. The reader may be puzzled as to why we do not use the trivial centering of subtracting the mean from each column of $\tilde{\boldsymbol{J}}$. The reason we do not pursue this path is that we can not directly relate the minimum eigenvalues of the resulting matrix to that of $\tilde{\boldsymbol{J}}$ in a useful way. Steps one and two allow us to center the columns of $\tilde{\boldsymbol{J}}$, while being able to relate the minimum eigenvalue of $\tilde{\boldsymbol{J}}$ to that of the centered matrix $\tilde{\boldsymbol{J}}^c$. We note that under Assumption 2.3 (b), the whitening and droppings step are superfluous because the matrix $\tilde{\boldsymbol{J}}$ is centered. We are now ready to discuss each of these steps in greater detail.

7.2.1 Whitening $\tilde{J}$

In this step we whiten the matrix $\tilde{J}$ in such a way as most of the entries of the corresponding matrix have zero mean. To explain our whitening procedure we begin this section by computing the expectation of $\tilde{J}$.

Lemma 7.6 *Let $\mathbf{\Gamma}$ be given by* (7.2) *and set $\boldsymbol{A} \coloneqq \boldsymbol{D_v\Gamma W}$. Construct the block diagonal matrix $\boldsymbol{D_A} \in \mathbb{R}^{kd \times d^2}$ with d copies of $\boldsymbol{A}$ on its diagonal. Further, consider $\boldsymbol{Q} \in \mathbb{R}^{d^2 \times n}$, where its rows are indexed by (i, j), for $1 \le i \le j \le d$. The rows (i, i) of $\boldsymbol{Q}$ are all-one, while the other rows are all-zero. Then,*

$$\mathbb{E}[\tilde{J}] = \boldsymbol{D_A Q} \,, \tag{7.16}$$

where the expectation is with respect to the randomness in the inputs $\boldsymbol{x}_i$.

Lemma 7.6 above is proven in Appendix E. To center, most of the entries of $\tilde{J}$ we use the structure of $\boldsymbol{Q}$ and whiten it from the left-hand side. We will also show that this whitening procedure will not significantly decrease the minimum eigenvalue.

Note that we can assume that $\sigma_{\min}(\boldsymbol{D_v\Gamma W}) > 0$ otherwise, Claim (7.7) becomes trivial. Now let $\boldsymbol{M} \in R^{d \times k}$ be the left inverse of $\boldsymbol{D_v\Gamma W}$ and construct a block-diagonal matrix $\boldsymbol{D_M}$ with d copies of $\boldsymbol{M}$ on its diagonal. The lemma below, relates the minimum singular value of $\tilde{J}$ to that of the whitened matrix $\boldsymbol{D_M}\tilde{J}$.

Lemma 7.7 *We have*

$$\sigma_{\min}(\tilde{J}) \ge \sigma_{\min}(\boldsymbol{D_M}\tilde{J})\sigma_{\min}(\boldsymbol{D_v\Gamma W}) \,.$$

Proof We prove this lemma by contradiction i.e. assume

$$\sigma_{\min}(\tilde{J}) < \sigma_{\min}(\boldsymbol{D_M}\tilde{J})\sigma_{\min}(\boldsymbol{D_v\Gamma W}) \,.$$

Now let $\boldsymbol{v}$ be the bottom singular vector of $\tilde{J}$. Then,

$$\left\|\tilde{J}\boldsymbol{v}\right\|_{\ell_2} = \sigma_{\min}(\tilde{J})\left\|\boldsymbol{v}\right\|_{\ell_2} < \sigma_{\min}(\boldsymbol{D_v\Gamma W})\sigma_{\min}(\boldsymbol{D_M}\tilde{J})\left\|\boldsymbol{v}\right\|_{\ell_2} \le \frac{\left\|\boldsymbol{D_M}\tilde{J}\boldsymbol{v}\right\|_{\ell_2}}{\sigma_{\max}(\boldsymbol{M})} \le \left\|\tilde{J}\boldsymbol{v}\right\|_{\ell_2} \,,$$

which is a contradiction. Note that in the penultimate inequality we used the fact that $\boldsymbol{MD_v\Gamma W} = \boldsymbol{I}$ and the last inequality holds because $\left\|\boldsymbol{D_M}\right\| = \sigma_{\max}(\boldsymbol{M})$. ∎

7.2.2 Dropping rows

By Lemma 7.6, we have

$$\mathbb{E}[\boldsymbol{D_M}\tilde{J}] = \boldsymbol{D_M D_A Q} = \boldsymbol{Q} \,.$$

Therefore, the whitened matrix $\boldsymbol{D_M}\tilde{J}$ is almost centered. To reach a completely centered matrix we drop the rows corresponding to the nonzero rows of $\boldsymbol{Q}$. Specifically, let $\tilde{J}^c$ be the matrix obtained after dropping rows (i, i) from $\boldsymbol{D_M}\tilde{J}$, for $1 \le i \le d$. By dropping these rows, $\boldsymbol{Q}$ vanishes and hence, $\mathbb{E}[\tilde{J}^c] = \boldsymbol{0}$. Note that $\tilde{J}^c$ is obtained by dropping d of the rows from $\boldsymbol{D_M}\tilde{J}$. Hence, $\sigma_{\min}(\boldsymbol{D_M}\tilde{J}) \ge \sigma_{\min}(\tilde{J}^c)$. Combining the latter with Lemma 7.7 we arrive at

$$\sigma_{\min}(\tilde{J}) \ge \sigma_{\min}(\tilde{J}^c)\sigma_{\min}(\boldsymbol{D_v\Gamma W}) \,. \tag{7.17}$$

Thus, in the remainder we shall focus on lower bounding $\sigma_{\min}(\tilde{J}^c)$.

7.2.3 Bounding sub-exponential norms

Let $\tilde{\boldsymbol{J}}_{\boldsymbol{x}}$ be the column of $\tilde{\boldsymbol{J}}$ corresponding to data point $\boldsymbol{x}$. By Lemma 5.1, we can write

$$\boldsymbol{D_M}\tilde{\boldsymbol{J}}_{\boldsymbol{x}} = \boldsymbol{x} \otimes \boldsymbol{M}\boldsymbol{D_v}(\phi'(\boldsymbol{W}\boldsymbol{x}) - \mathbb{E}[\phi'(\boldsymbol{W}\boldsymbol{x})]).$$

We recall that $\tilde{\boldsymbol{J}}^{\boldsymbol{c}}$ is obtained by dropping d of the rows from $\boldsymbol{D_M}\tilde{\boldsymbol{J}}$. By the structure of $\boldsymbol{Q} = \mathbb{E}[\boldsymbol{D_M}\tilde{\boldsymbol{J}}]$, we can alternatively obtain $\tilde{\boldsymbol{J}}^{\boldsymbol{c}}$ by dropping the same set of rows from $\boldsymbol{D_M}\tilde{\boldsymbol{J}} - \mathbb{E}[\boldsymbol{D_M}\tilde{\boldsymbol{J}}]$. Note that dropping entries from a vector can only reduce the Orlicz norm. Therefore,

$$\|\tilde{\boldsymbol{J}}^{\boldsymbol{c}}_{\boldsymbol{x}}\|_{\psi_1} \le \left\|\boldsymbol{D_M}\tilde{\boldsymbol{J}}_{\boldsymbol{x}} - \mathbb{E}[\boldsymbol{D_M}\tilde{\boldsymbol{J}}_{\boldsymbol{x}}]\right\|_{\psi_1} = \left\|\boldsymbol{x} \otimes \boldsymbol{M}\boldsymbol{D_v}\boldsymbol{z} - \mathbb{E}[\boldsymbol{x} \otimes \boldsymbol{M}\boldsymbol{D_v}\boldsymbol{z}]\right\|_{\psi_1}. \tag{7.18}$$

In order to bound the right-hand-side of (7.18), we need to study the tail of the random variable

$$\left\langle \boldsymbol{x} \otimes \boldsymbol{M}\boldsymbol{D_v}\boldsymbol{z} - \mathbb{E}[\boldsymbol{x} \otimes \boldsymbol{M}\boldsymbol{D_v}\boldsymbol{z}], \boldsymbol{u} \right\rangle$$

for any unit-norm vector $\boldsymbol{u} \in \mathbb{R}^{kd}$. To simplify this random variable define $\widetilde{\boldsymbol{U}} \in \mathbb{R}^{d \times d}$ by arranging every k entries of $\boldsymbol{u}$ as the rows of $\widetilde{\boldsymbol{U}}$. Note that $\boldsymbol{u} = \text{vect}(\widetilde{\boldsymbol{U}})$. Thus Lemma 5.2 allows us to rewrite this random variable in the simplified form

$$\langle \boldsymbol{x} \otimes \boldsymbol{M}\boldsymbol{D_v}\boldsymbol{z} - \mathbb{E}[\boldsymbol{x} \otimes \boldsymbol{M}\boldsymbol{D_v}\boldsymbol{z}], \boldsymbol{u} \rangle = \boldsymbol{x}^T\widetilde{\boldsymbol{U}}\boldsymbol{M}\boldsymbol{D_v}\boldsymbol{z} - \mathbb{E}[\boldsymbol{x}^T\widetilde{\boldsymbol{U}}\boldsymbol{M}\boldsymbol{D_v}\boldsymbol{z}], \tag{7.19}$$

with $\boldsymbol{z} \in \mathbb{R}^k$ defined as $\boldsymbol{z} := \phi'(\boldsymbol{W}\boldsymbol{x}) - \mathbb{E}[\phi'(\boldsymbol{W}\boldsymbol{x})]$. To characterize the tails of the latter random variable we use the following lemma proven in Appendix D.

Proposition 7.8 *Let $\boldsymbol{x} \in \mathbb{R}^d$ be a random Gaussian vector distributed as $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$. Also define $\boldsymbol{z} := \phi'(\boldsymbol{W}\boldsymbol{x}) - \mathbb{E}[\phi'(\boldsymbol{W}\boldsymbol{x})]$. Then,*

$$\mathbb{P}\left\{\left|\boldsymbol{x}^T\boldsymbol{U}\boldsymbol{z} - \mathbb{E}[\boldsymbol{x}^T\boldsymbol{U}\boldsymbol{z}]\right| \ge t\right\} \le 2\exp\left(-\frac{1}{C}\min\left(\frac{t^2}{K^4\|\boldsymbol{U}\|_F^2}, \frac{t}{K^2\|\boldsymbol{U}\|}\right)\right), \tag{7.20}$$

holds with C a fixed numerical constant and $K = \sqrt{L^2\sigma_{\max}^2(\boldsymbol{W}) + 1}$.

We apply Proposition 7.8, with $\boldsymbol{U} = \widetilde{\boldsymbol{U}}\boldsymbol{M}\boldsymbol{D_v}$ and use Lemma 5.5, to conclude that

$$\left\|\boldsymbol{x}^T\widetilde{\boldsymbol{U}}\boldsymbol{M}\boldsymbol{D_v}\boldsymbol{z} - \mathbb{E}[\boldsymbol{x}^T\widetilde{\boldsymbol{U}}\boldsymbol{M}\boldsymbol{D_v}\boldsymbol{z}]\right\|_{\psi_1} \le C(L^2\sigma_{\max}^2(\boldsymbol{W}) + 1)\left\|\widetilde{\boldsymbol{U}}\boldsymbol{M}\boldsymbol{D_v}\right\|_F,$$
$$\le C(L^2\sigma_{\max}^2(\boldsymbol{W}) + 1)\,\sigma_{\max}(\boldsymbol{M}\boldsymbol{D_v})\left\|\widetilde{\boldsymbol{U}}\right\|_F,$$
$$\le \frac{C}{\sigma_{\min}(\boldsymbol{\Gamma}\boldsymbol{W})}(L^2\sigma_{\max}^2(\boldsymbol{W}) + 1). \tag{7.21}$$

In the last inequality we used the fact that $\left\|\widetilde{\boldsymbol{U}}\right\|_F = \|\boldsymbol{u}\|_{\ell_2} = 1$. Now since $\boldsymbol{u}$ was arbitrary, recalling Definition 5.1 and combining equations (7.18), (7.19) and (7.21) we conclude that

$$\|\tilde{\boldsymbol{J}}^{\boldsymbol{c}}_{\boldsymbol{x}}\|_{\psi_1} \le \frac{C}{\sigma_{\min}(\boldsymbol{\Gamma}\boldsymbol{W})}(L^2\sigma_{\max}^2(\boldsymbol{W}) + 1). \tag{7.22}$$

7.2.4 Bounding minimum singular value

Now that we have a bound on the sub-exponential norm of the columns of $\tilde{J}^c$, we are now ready to bound $\sigma_{\min}(\tilde{J}^c)$. To this aim we first state a lemma on the spectrum of matrices with independent sub-exponential columns. The Lemma is similar to [2, Theorem 3.2] and we give a proof in Appendix F.

Proposition 7.9 *Let $u_1, u_2, \ldots, u_n$ be independent sub-exponential random vectors and also let $\psi = \max_{1 \le i \le n} \|u_i\|_{\psi_1}$. Furthermore, let U be a matrix with $u_1, \ldots, u_n$ as columns. Define $\eta_{\min} = \min_{1 \le i \le n} \|u_i\|$ and $\eta_{\max} = \max_{1 \le i \le n} \|u_i\|$. Further, set $\xi = \psi K + K'$, where $K, K' \ge 1$ are arbitrary but fixed. Also define*

$$\Delta = C\xi^2 \eta_{\min} \sqrt{n} \log\left(\frac{2\eta_{\min}}{\sqrt{n}}\right).$$

Then

$$\sigma_{\min}^2(U) \ge \eta_{\min}^2 - \Delta \quad and \quad \sigma_{\max}^2(U) \le \eta_{\max}^2 + \Delta,$$

hold with probability larger than

$$1 - C\exp\left(-cK\sqrt{n}\log\left(\frac{2\eta_{\min}}{\sqrt{n}}\right)\right) - \mathbb{P}\left(\eta_{\max} \ge K'\eta_{\min}\right).$$

Here, $c, C > 0$ are universal constants.

To apply Proposition 7.9, we only need to control norms of columns $\tilde{J}^c_x$ (and thus the parameters $\eta_{\min}$ and $\eta_{\max}$ in the proposition statement). To lighten the notation, we let $z = MD_v(\phi'(Wx) - \mathbb{E}[\phi'(Wx)])$. The column of $D_M\tilde{J}$ corresponding to data point x is given by $x \otimes z$ and the column $\tilde{J}^c_x$ is obtained after dropping the entries $x_i z_i$, for $1 \le i \le d$. Hence,

$$\|\tilde{J}^c_x\|_{\ell_2}^2 = \sum_{i=1}^d z_i^2 \left(\sum_{j \ne i} x_j^2\right). \tag{7.23}$$

We continue by stating a lemma that bounds the Euclidean norm of the random vector z. We defer the proof of this lemma to Appendix G.

Lemma 7.10 *Assume x is a Gaussian random vector distributed as $\mathcal{N}(0, I)$. Furthuremore, assume $\phi : \mathbb{R} \to \mathbb{R}$ is an activation function with bounded second derivative, i.e. $|\phi''| \le L$. Define*

$$\rho(W) := L\frac{\sigma_{\max}(W)}{\sigma_{\min}(\Gamma W)}.$$

Then for $z = MD_v(\phi'(Wx) - \mathbb{E}[\phi'(Wx)])$,

$$\sqrt{d} - t\rho(W) \le \|z\|_{\ell_2} \le (\sqrt{d} + t)\rho(W), \tag{7.24}$$

holds with probability at least $1 - 2e^{-\frac{t^2}{2}}$.

Further, by concentration of χ^2 random variables, with probability at least $1 - 2e^{-(d-1)/32}$, the following bounds hold. For any fixed $1 \le i \le n$,

$$\frac{d-1}{2} \le \left| \sum_{j \ne i} x_j^2 \right| \le \frac{3(d-1)}{2}. \tag{7.25}$$

Plugging bound (7.24), with $t = 2\sqrt{\log n}$, and bound (7.25) into (7.23), the followings inequalities hold with probability at least $1 - 2n^{-2} - 2e^{-(d-1)/32}$,

$$\|\tilde{\boldsymbol{J}}_{\boldsymbol{x}}^{\boldsymbol{c}}\|_{\ell_2} \ge \sqrt{\frac{d-1}{2}}(\sqrt{d} - 2\rho(\boldsymbol{W})\sqrt{\log n}) \ge \frac{1}{3}\sqrt{d(d-1)}, \tag{7.26}$$

$$\|\tilde{\boldsymbol{J}}_{\boldsymbol{x}}^{\boldsymbol{c}}\|_{\ell_2} \le \sqrt{\frac{3(d-1)}{2}}(\sqrt{d} + 2\sqrt{\log n})\rho(\boldsymbol{W}) \le 3\sqrt{d(d-1)}\rho(\boldsymbol{W}). \tag{7.27}$$

Note that the second inequality in (7.26) holds because by our assumption (7.6),

$$\rho^2(\boldsymbol{W}) < \frac{L^2\sigma_{\max}^2(\boldsymbol{W}) + 1}{\sigma_{\min}^2(\boldsymbol{\Gamma}\boldsymbol{W})} \le c_1\frac{d}{\sqrt{n}} < \frac{d}{16\log n}. \tag{7.28}$$

Therefore, by union bounding over n columns with probability at least $1 - 2n^{-1} - 2ne^{-(d-1)/32}$ we have

$$\eta_{\min} \ge \frac{1}{3}\sqrt{d(d-1)}, \tag{7.29}$$

$$\eta_{\max} \le 3\sqrt{d(d-1)}\rho(\boldsymbol{W}). \tag{7.30}$$

Also, by (7.22), the maximum sub-exponential norms of the columns is bounded by

$$\psi \le \frac{C}{\sigma_{\min}(\boldsymbol{\Gamma}\boldsymbol{W})}(L^2\sigma_{\max}^2(\boldsymbol{W}) + 1). \tag{7.31}$$

We now have all the elements to apply Proposition 7.9. In particular we use $K = 1$ and $K' = 10\rho(\boldsymbol{W})$, to conclude that

$$\sigma_{\min}^2(\tilde{\boldsymbol{J}}^{\boldsymbol{c}}) \ge \eta_{\min}^2\left(1 - C(\psi + \rho(\boldsymbol{W}))^2\frac{\sqrt{n}}{\eta_{\min}}\log\left(\frac{2\eta_{\min}}{\sqrt{n}}\right)\right), \tag{7.32}$$

holds with probability at least

$$1 - C\exp\left(-c\sqrt{n}\log\left(\frac{2\eta_{\min}}{\sqrt{n}}\right)\right).$$

Using the derived lower bound on $\eta_{\min}$ and the upper bound on ψ, we obtain that with high probability $\sigma_{\min}^2(\tilde{\boldsymbol{J}}^{\boldsymbol{c}}) \ge d/2$, as long as

$$\frac{L^4\sigma_{\max}^4(\boldsymbol{W}) + 1}{\sigma_{\min}^2(\boldsymbol{\Gamma}\boldsymbol{W})} \le c_1\frac{d}{\sqrt{n}}, \tag{7.33}$$

which holds true by the assumption given in Equation (7.6).

31

Finally by (7.17), we conclude that

$$\sigma_{\min}(\tilde{J}) \geq \frac{d}{2}\sigma_{\min}(D_v\Gamma W). \tag{7.34}$$

Note that under Assumption 2.3 (b), we skip the whitening step and by following the same proof, we obtain

$$\sigma_{\min}(\tilde{J}) \geq \frac{d}{2}. \tag{7.35}$$

7.2.5 Bounding maximum singular value

The argument we used for the minimum eigenvalue does not apply to the largest eigenvalue. This is because D_M is a fat matrix and the maximum eigenvalue of $D_M\tilde{J}$ does not provide any information about the maximum eigenvalue of $\tilde{J}$. To see this clearly, consider the case where the column space of $\tilde{J}$ intersects with the null space of D_M e.g. when maximum eigenvector of $\tilde{J}$ is in the null space of D_M. For bounding $\sigma_{\max}(\tilde{J})$, instead of whitening and dropping some of the row, we center $\tilde{J}$ in the most natural way. Specifically, in this section we let $\tilde{J}^c := \tilde{J} - \mathbb{E}[\tilde{J}]$. Then,

$$\tilde{J}_x^c = \tilde{J}_x - \mathbb{E}[\tilde{J}_x] = x \otimes D_v z - \mathbb{E}[x \otimes D_v z],$$

with $z = \phi'(Wx) - \mathbb{E}[\phi'(Wx)]$. Applying Proposition 7.8, we get

$$\|\tilde{J}_x^c\|_{\psi_1} \leq C\|v\|_\infty(L^2\sigma_{\max}^2(W) + 1). \tag{7.36}$$

Similar to Section 7.9, we can bound the maximum and the minimum norms of the columns of $\tilde{J}^c$. With probability at least $1 - 2n^{-1} - 2ne^{-(d-1)/32}$,

$$\eta_{\min} \geq d/3, \quad \eta_{\max} \leq 3d\rho(W). \tag{7.37}$$

By employing Proposition 7.9,

$$\begin{aligned}
\sigma_{\max}^2(\tilde{J}^c) &\leq C\left(d^2 + \sqrt{n}d\|v\|_\infty^2\left(L^4\sigma_{\max}^4(W) + 1\right) + \sqrt{n}d\rho^2(W)\right) \\
&\leq Cd^2\left(1 + c_1\|v\|_\infty^2\sigma_{\min}^2(\Gamma W) + c_1\right) \\
&\leq Cd^2\left(1 + \|v\|_\infty^2\sigma_{\min}^2(\Gamma W)\right),
\end{aligned} \tag{7.38}$$

where the second inequality follows from our assumption given by Equation (7.6).

We next bound the maximum eigenvalue of $\mathbb{E}[\tilde{J}]$. Invoking Lemma 7.6, we have

$$\mathbb{E}[\tilde{J}] = D_A Q = \text{vec}(A)\mathbf{1}_n^T,$$

where $\mathbf{1}_n \in \mathbb{R}^n$ is the all-one vector. Therefore,

$$\sigma_{\max}(\mathbb{E}[\tilde{J}]) \leq \sqrt{n}\|A\|_F \leq \sqrt{nd}\,\sigma_{\max}(D_v\Gamma W). \tag{7.39}$$

Combining Equations (7.38) and (7.39) via the triangular inequality we arrive at:

$$\begin{aligned}
\sigma_{\max}(\tilde{J}) &\leq Cd\left(1 + \|v\|_\infty\sigma_{\min}(\Gamma W)\right) + \sqrt{nd}\,\sigma_{\max}(D_v\Gamma W), \\
&\leq C\left(d + \sqrt{nd}\,\sigma_{\max}(D_v\Gamma W)\right),
\end{aligned}$$

where the second inequality holds because $n \geq d$ and $\|v\|_\infty \sigma_{\min}(\mathbf{\Gamma W}) \leq \sigma_{\max}(\mathbf{D_v \Gamma W})$ as per Lemma 5.3.

Note that under Assumption 2.3 (b), matrix $\tilde{\mathbf{J}}$ is centered and the whitening step should be skipped. Indeed, by a similar argument for inequality (7.38), we have

$$
\begin{aligned}
\sigma_{\max}^2(\tilde{\mathbf{J}}) &\leq C\left(d^2 + \sqrt{n}dv_{\max}^2(L^4\sigma_{\max}^4(\mathbf{W}) + 1) + \sqrt{n}d\sigma_{\max}^2(\mathbf{W})\right), \\
&\leq Cd^2(1 + c_1 v_{\max}^2 + c_1), \\
&\leq Cd^2.
\end{aligned}
\tag{7.40}
$$

7.3 Proof of Theorem 2.5

We begin by stating a few useful lemmas. We defer the proofs of these lemmas to the Appendices. Throughout, we assume $|\phi'| < B$ and $|\phi''| < L$. We present the proof under Assumption 2.3 (a). The proof under Assumption 2.3 (b) follows by an analogous argument. We begin by a lemma that bound the perturbation of the Jacobian matrix as a function of its inputs. We defer the proof to Appendix H.

Lemma 7.11 *Assume $\mathbf{x}_i$ are i.i.d. Gaussian random vectors distributed as $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Further, suppose that $n \leq cd^2$, and let $\mathbf{J}(v, \mathbf{W}) = \mathbf{D_v}\phi'(\mathbf{WX}) * \mathbf{X}$ be the Jacobian matrix associated to weights v and $\mathbf{W}$. Then, for any two fixed matrices $\widetilde{\mathbf{W}}, \mathbf{W} \in \mathbb{R}^{k \times d}$ and any two fixed vectors $v, \widetilde{v} \in \mathbb{R}^k$,*

$$
\left\|J(\widetilde{v}, \widetilde{\mathbf{W}}) - J(v, \mathbf{W})\right\| \leq C'\sqrt{nd}\left(\|v\|_\infty\left\|\widetilde{\mathbf{W}} - \mathbf{W}\right\| + \|\widetilde{v} - v\|_\infty\right),
\tag{7.41}
$$

holds with probability at least $1 - ne^{-b_1\sqrt{n}} - n^{-1} - 2ne^{-b_2 d}$. Here, $b_1, b_2 > 0$ are fixed numerical constants.

Now note that by Proposition 7.4 for v^* and $\mathbf{W}^*$, we have

$$
\sigma_{\min}(\mathbf{J}(v^*, \mathbf{W}^*)) \geq c\,\sigma_{\min}(\mathbf{W}^*)d.
\tag{7.42}
$$

Using this value of c we define the radius

$$
R := \frac{c}{4C'}\sigma_{\min}(\mathbf{W}^*)\sqrt{\frac{d}{n}},
$$

where C' is the same quantity that appears in Equation (7.41). Without loss of generality, we can assume $c < 4C'w_{\max}v_{\max}$. Plugging this into the definition of R together with the fact that $d \leq n$, allows us to conclude that $R \leq v_{\max}$.

Define the set $\Omega \subseteq \mathbb{R}^{k \times d} \times \mathbb{R}^k$ as follows:

$$
\Omega := \left\{(v, \mathbf{W}) : \|v - v^*\|_\infty \leq R,\ \|\mathbf{W} - \mathbf{W}^*\|_F \leq R/\|v^*\|_\infty\right\}
\tag{7.43}
$$

In the next lemma, proven in Appendix I, we relate the gradients $\nabla_{(\mathbf{W})}\mathcal{L}(v, \mathbf{W})$ and $\nabla_v\mathcal{L}(v, \mathbf{W})$ to the function value $\mathcal{L}(v, \mathbf{W})$.

Lemma 7.12 *For $(\boldsymbol{v}, \boldsymbol{W}) \in \Omega$, the following inequalities hold with probability at least $1 - n^{-1} - 2ne^{-b\sqrt{d}}$ for some constant $b > 0$.*

$$\|\nabla_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{v}, \boldsymbol{W})\|_F^2 \geq m_{\mathrm{L}} \mathcal{L}(\boldsymbol{v}, \boldsymbol{W}) \qquad m_{\mathrm{L}} := \frac{1}{2} c^2 \sigma_{\min}^2(\boldsymbol{W}^*) \frac{d^2}{n}, \tag{7.44}$$

$$\|\nabla_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{v}, \boldsymbol{W})\|_F^2 \leq m_{\mathrm{U}} \mathcal{L}(\boldsymbol{v}, \boldsymbol{W}) \qquad m_{\mathrm{U}} := \frac{9}{2} C^2 \sigma_{\max}^2(\boldsymbol{W}^*) k, \tag{7.45}$$

$$\|\nabla_{\boldsymbol{v}} \mathcal{L}(\boldsymbol{v}, \boldsymbol{W})\|_\infty^2 \leq \widetilde{m}_{\mathrm{U}} \mathcal{L}(\boldsymbol{v}, \boldsymbol{W}) \qquad \widetilde{m}_{\mathrm{U}} := \left(4\phi^2(0) + 128 B^2 + 128 B^2 w_{\max}^2 \right). \tag{7.46}$$

Our next lemma, proven in Appendix J, upper bounds the function value $\mathcal{L}(\boldsymbol{v}, \boldsymbol{W})$ in terms of distance of $(\boldsymbol{v}, \boldsymbol{W})$ to the planted solution $(\boldsymbol{v}, \boldsymbol{W})$.

Lemma 7.13 *The following bound holds with probability at least $1 - 2e^{-b_0 d}$ for some constant $b_0 > 0$.*

$$\mathcal{L}(\boldsymbol{v}, \boldsymbol{W}) \leq \|\boldsymbol{v}^*\|_{\ell_2}^2 \|\boldsymbol{W} - \boldsymbol{W}^*\|_F^2 + 2 \left(\phi^2(0) + 2 B^2 w_{\max}^2 \right) k^2 \|\boldsymbol{v} - \boldsymbol{v}^*\|_\infty^2. \tag{7.47}$$

Finally, the lemma below, proven in Appendix K, controls the second order derivative of the loss function $\mathcal{L}(\boldsymbol{v}, \boldsymbol{W})$.

Lemma 7.14 *The function $\mathcal{L}(\boldsymbol{v}, \boldsymbol{W})$ is β-smooth on Ω. Namely, there exists an event of probability at least $1 - n^{-1} - 2ne^{-b\sqrt{d}}$, such that on this event, for any $(\boldsymbol{v}, \boldsymbol{W}) \in \Omega$, we have*

$$\nabla^2 \mathcal{L}(\boldsymbol{v}, \boldsymbol{W}) \leq \beta \boldsymbol{I}, \tag{7.48}$$

where $\nabla^2 \mathcal{L}(\boldsymbol{v}, \boldsymbol{W}) \in \mathbb{R}^{(kd+k) \times (kd+k)}$ denotes the Hessian w.r.t both $\boldsymbol{v}, \boldsymbol{W}$. Further, the smoothness parameter β is given by

$$\beta := \left(3 C^2 \sigma_{\max}^2(\boldsymbol{W}^*) + 8 v_{\max}^2 B L + 4 B^2 w_{\max}^2 + 2\phi^2(0) \right) k. \tag{7.49}$$

With this lemmas in place we are now ready to present our local convergence analysis. To this aim note that Lemma 7.12 (Equation (7.44)) implies that for $(\boldsymbol{v}, \boldsymbol{W}) \in \Omega$, the function $\mathcal{L}(\boldsymbol{v}, \boldsymbol{W})$ satisfies the Polyak-Lojasiewicz (PL) inequality [39]

$$\|\nabla_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{v}, \boldsymbol{W})\|_F^2 + \|\nabla_{\boldsymbol{v}} \mathcal{L}(\boldsymbol{v}, \boldsymbol{W})\|_F^2 \geq m_{\mathrm{L}} \mathcal{L}(\boldsymbol{v}, \boldsymbol{W}). \tag{7.50}$$

We shall now focus on how the loss function value changes in one iteration. Using the β-smoothness condition for $\alpha \leq 1/\beta$ we have

$$\mathcal{L}\left(\boldsymbol{v} - \alpha \nabla_{\boldsymbol{v}} \mathcal{L} \boldsymbol{v}, \boldsymbol{W} - \alpha \nabla_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{W}) \right),$$

$$\leq \mathcal{L}(\boldsymbol{v}, \boldsymbol{W}) - \alpha \left(\|\nabla_{\boldsymbol{v}} \mathcal{L}(\boldsymbol{v}, \boldsymbol{W})\|_{\ell_2}^2 + \|\nabla_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{v}, \boldsymbol{W})\|_F^2 \right) + \frac{\alpha^2 \beta}{2} \left(\|\nabla_{\boldsymbol{v}} \mathcal{L}(\boldsymbol{v}, \boldsymbol{W})\|_{\ell_2}^2 + \|\nabla_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{v}, \boldsymbol{W})\|_F^2 \right),$$

$$= \mathcal{L}(\boldsymbol{v}, \boldsymbol{W}) - \alpha \left(1 - \frac{\alpha \beta}{2} \right) \left(\|\nabla_{\boldsymbol{v}} \mathcal{L}(\boldsymbol{v}, \boldsymbol{W})\|_{\ell_2}^2 + \|\nabla_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{v}, \boldsymbol{W})\|_F^2 \right),$$

$$\leq \left(1 - \alpha m_{\mathrm{L}} \left(1 - \frac{\alpha \beta}{2} \right) \right) \mathcal{L}(\boldsymbol{v}, \boldsymbol{W}) \leq \left(1 - \frac{\alpha}{2} m_{\mathrm{L}} \right) \mathcal{L}(\boldsymbol{v}, \boldsymbol{W}), \tag{7.51}$$

where in the last inequality we used (7.50).

Our assumptions on the initial weights $\boldsymbol{v}_0$ and $\boldsymbol{W}_0$ in equations (2.5) and (2.4) imply the following identities.

$$\|\boldsymbol{W}_0 - \boldsymbol{W}^*\|_F \le \frac{Rm_{\mathrm{L}}}{4\sqrt{2m_{\mathrm{U}}}\,v_{\max}\,\|\boldsymbol{v}^*\|_{\ell_2}},\tag{7.52}$$

$$\|\boldsymbol{v}_0 - \boldsymbol{v}^*\|_\infty \le \frac{Rm_{\mathrm{L}}}{4\sqrt{2m_{\mathrm{U}}}v_{\max}\sqrt{2(\phi^2(0) + 2B^2 w_{\max}^2)\,k}}.\tag{7.53}$$

We next show that starting with $\boldsymbol{v}_0, \boldsymbol{W}_0$ obeying (7.52) and (7.53), the entire trajectory of gradient descent remains in the set Ω. To establish this we use induction on τ. The induction basis $\tau = 0$ is trivial. Assuming the induction hypothesis for $0 \le t \le \tau - 1$, we show that the result continues to hold for $t = \tau$.

By our induction hypothesis $((\boldsymbol{v}_t, \boldsymbol{W}_t) \in \Omega$ for $t = 0, 1, \ldots, \tau - 1)$ and (7.51), we have

$$\mathcal{L}(\boldsymbol{v}_\tau, \boldsymbol{W}_\tau) = \mathcal{L}\left(\boldsymbol{v}_{\tau-1} - \alpha \nabla_{\boldsymbol{v}} \mathcal{L}(\boldsymbol{v}_{\tau-1}, \boldsymbol{W}_{\tau-1}), \boldsymbol{W}_{\tau-1} - \alpha \nabla_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{v}_{\tau-1}, \boldsymbol{W}_{\tau-1})\right),$$

$$\le \left(1 - \frac{\alpha}{2} m_{\mathrm{L}}\right) \mathcal{L}(\boldsymbol{v}_{\tau-1}, \boldsymbol{W}_{\tau-1}).$$

By iterating the above identity we arrive at

$$\mathcal{L}(\boldsymbol{v}_\tau, \boldsymbol{W}_\tau) \le \left(1 - \frac{\alpha}{2} m_{\mathrm{L}}\right)^\tau \mathcal{L}(\boldsymbol{v}_0, \boldsymbol{W}_0).\tag{7.54}$$

To show that $(\boldsymbol{v}_\tau, \boldsymbol{W}_\tau) \in \Omega$ we proceed by quantifying how far the gradient descent trajectory can get from $(\boldsymbol{v}_0, \boldsymbol{W}_0)$.

$$\|\boldsymbol{W}_\tau - \boldsymbol{W}_0\|_F = \left\|\sum_{t=1}^\tau (\boldsymbol{W}_t - \boldsymbol{W}_{t-1})\right\|_F \le \sum_{t=1}^\tau \|\boldsymbol{W}_t - \boldsymbol{W}_{t-1}\|_F$$

$$\le \alpha \sum_{t=1}^\tau \|\nabla_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{v}_{t-1}, \boldsymbol{W}_{t-1})\|_F \overset{(a)}{\le} \alpha \sqrt{m_{\mathrm{U}}} \sum_{t=1}^\tau \sqrt{\mathcal{L}(\boldsymbol{v}_{t-1}, \boldsymbol{W}_{t-1})}$$

$$\overset{(b)}{\le} \alpha \sqrt{m_{\mathrm{U}}} \sum_{t=1}^\tau \sqrt{\left(1 - \frac{\alpha}{2} m_{\mathrm{L}}\right)^{t-1} \mathcal{L}(\boldsymbol{v}_0, \boldsymbol{W}_0)}$$

$$\le \alpha \sqrt{m_{\mathrm{U}} \mathcal{L}(\boldsymbol{v}_0, \boldsymbol{W}_0)} \sum_{t=1}^\tau \left(1 - \frac{\alpha}{4} m_{\mathrm{L}}\right)^{t-1}$$

$$\le \frac{\alpha \sqrt{m_{\mathrm{U}} \mathcal{L}(\boldsymbol{v}_0, \boldsymbol{W}_0)}}{\alpha \frac{m_{\mathrm{L}}}{4}}$$

$$= \frac{4}{m_{\mathrm{L}}} \sqrt{m_{\mathrm{U}} \mathcal{L}(\boldsymbol{v}_0, \boldsymbol{W}_0)},\tag{7.55}$$

where (a) follows from $(\boldsymbol{v}_{t-1}, \boldsymbol{W}_{t-1}) \in \Omega$ and Lemma 7.12 equation (7.45) and (b) follows from (7.54). Likewise, we obtain

$$\|\boldsymbol{v}_\tau - \boldsymbol{v}_0\|_\infty \le \frac{4}{m_{\mathrm{L}}} \sqrt{\widetilde{m}_{\mathrm{U}} \mathcal{L}(\boldsymbol{v}_0, \boldsymbol{W}_0)}\tag{7.56}$$

Using bounds (7.55) and (7.56), in order to show that $(\boldsymbol{v}_\tau, \boldsymbol{W}_\tau) \in \Omega$, it suffices to show that

$$\mathcal{L}(\boldsymbol{v}_0, \boldsymbol{W}_0) \le \frac{R^2 m_{\mathrm{L}}^2}{16 \max(\widetilde{m}_{\mathrm{U}}, m_{\mathrm{U}} v_{\max}^2)}.\tag{7.57}$$

35

The right-hand sides of Equation (7.57) depends on $\max(\widetilde{m}_U, m_U v_{\max}^2)$. The dominant term is $m_U v_{\max}^2$ because it is of order at least k, while m_U is $O(1)$. Therefore, the desired bound in (7.57) is equivalent to

$$\mathcal{L}(\boldsymbol{v}_0, \boldsymbol{W}_0) \le \frac{R^2 m_L^2}{16 m_U v_{\max}^2} \,. \tag{7.58}$$

We can verify Equation (7.58) by using Lemma 7.13 combined with (7.52) and (7.53). This completes the induction argument and shows that $(\boldsymbol{v}_\tau, \boldsymbol{W}_\tau) \in \Omega$ for all $\tau \ge 1$.

Finally, since $(\boldsymbol{v}_\tau, \boldsymbol{W}_\tau) \in \Omega$ for all $\tau \ge 0$, (7.54) holds for all $\tau \ge 0$. Substituting for m_L, we obtain

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{v}_\tau, \boldsymbol{W}_\tau) &\le \left(1 - \frac{\alpha c^2}{4} \sigma_{\min}^2(\boldsymbol{W}^*) \frac{d}{n}\right)^\tau \mathcal{L}(\boldsymbol{v}_0, \boldsymbol{W}_0), \\
&\le \left(1 - c^2 w_{\max} \frac{\alpha d}{4n}\right)^\tau \mathcal{L}(\boldsymbol{v}_0, \boldsymbol{W}_0),
\end{aligned}
\tag{7.59}
$$

where the last step holds because $\sigma_{\min}(\boldsymbol{W}^*) < w_{\max}$.

This concludes the proof under Assumption 2.3 (a). The claim under Assumption 2.3 (b) can be proven by a similar argument. The only required adjustment is that the initial radius R and the term m_L should now be defined via $R = \frac{c}{4C'}\sqrt{d/n}$ and $m_L = c^2 d^2/(2n)$.

Acknowledgements

References

[1] R. Adamczak. A note on the hanson-wright inequality for random vectors with dependencies. *Electronic Communications in Probability*, 20, 2015.

[2] R. Adamczak, A. E. Litvak, A. Pajor, and N. Tomczak-Jaegermann. Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling. *Constructive Approximation*, 34(1):61–88, 2011.

[3] Radosław Adamczak, Alexander E Litvak, Alain Pajor, and Nicole Tomczak-Jaegermann. Sharp bounds on the rate of convergence of the empirical covariance matrix. *Comptes Rendus Mathematique*, 349(3-4):195–200, 2011.

[4] N. Agarwal, Z. Allen-Zhu, B. Bullins, E. Hazan, and T. Ma. Finding approximate local minima for nonconvex optimization in linear time. *arXiv preprint arXiv:1611.01146*, 2016.

[5] G. W. Anderson, A. Guionnet, and O. Zeitouni. *An introduction to random matrices*. Cambridge University Press, 2009.

[6] A. R. Barron. Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 14(1):115–133, 1994.

[7] P. Bartlett, D. J. Foster, and M. Telgarsky. Spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1706.08498*, 2017.

[8] M. Biehl, P. Riegler, and C. Wohler. Transient dynamics of on-line learning in two-layered neural networks. *Journal of Physics A: Mathematical and General*, 29(16):4769, 1996.

[9] A. Blum and R. L. Rivest. Training a 3-node neural network is NPs-complete. In *Proceedings of the 1st International Conference on Neural Information Processing Systems*, pages 494–501. MIT Press, 1988.

[10] H. Bolcskei, P. Grohs, G. Kutyniok, and P. Petersen. Optimal approximation with sparsely connected deep neural networks. *arXiv preprint arXiv:1705.01714*, 2017.

[11] A. Brutzkus and A. Globerson. Globally optimal gradient descent for a convnet with gaussian inputs. *arXiv preprint arXiv:1702.07966*, 2017.

[12] E. J. Candes, X. Li, and M. Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[13] Y. Carmon and J. C. Duchi. Gradient descent efficiently finds the cubic-regularized non-convex newton step. *arXiv preprint arXiv:1612.00547*, 2016.

[14] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. Accelerated methods for non-convex optimization. *arXiv preprint arXiv:1611.00756*, 2016.

[15] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun. The loss surfaces of multilayer networks. In *AISTATS*, 2015.

[16] R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

[17] F. E. Curtis, D. P. Robinson, and M. Samadi. A trust region algorithm with a worst-case iteration complexity of\ mathcal {O}(\ epsilonˆ{-3/2}) for nonconvex optimization. *Mathematical Programming*, pages 1–32, 2014.

[18] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on learning theory*, pages 797–842. PMLR, 2015.

[19] S. Goel, V. Kanade, A. Klivans, and J. Thaler. Reliably learning the relu in polynomial time. *arXiv preprint arXiv:1611.10258*, 2016.

[20] B. D. Haeffele and R. Vidal. Global optimality in tensor factorization, deep learning, and beyond. *arXiv preprint arXiv:1506.07540*, 2015.

[21] E. Hazan, K. Levy, and S. Shalev-Shwartz. Beyond convexity: Stochastic quasi-convex optimization. In *Advances in Neural Information Processing Systems*, pages 1594–1602, 2015.

[22] M. Janzamin, H. Sedghi, and A. Anandkumar. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods. *arXiv preprint arXiv:1506.08473*, 2015.

[23] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan. How to escape saddle points efficiently. *arXiv preprint arXiv:1703.00887*, 2017.

[24] S. M. Kakade, V. Kanade, O. Shamir, and A. Kalai. Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems*, pages 927–935, 2011.

[25] A. T. Kalai and R. Sastry. The isotron algorithm: High-dimensional isotonic regression. In *COLT*, 2009.

[26] K. Kawaguchi. Deep learning without poor local minima. In *Advances In Neural Information Processing Systems*, pages 586–594, 2016.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[28] C. J. Kuo. The CNN as a guided multilayer RECOS transform [lecture notes]. *IEEE Signal Processing Magazine*, 34(3):81–89, 2017.

[29] Michel Ledoux. *The concentration of measure phenomenon*, volume 89. American Mathematical Soc., 2005.

[30] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht. Gradient descent converges to minimizers. *University of California, Berkeley*, 1050:16, 2016.

[31] K. Y. Levy. The power of normalization: Faster evasion of saddle points. *arXiv preprint arXiv:1611.04831*, 2016.

[32] Y. Li, T. Ma, and H. Zhang. Algorithmic regularization in over-parameterized matrix recovery. *arXiv preprint arXiv:1712.09203*, 2017.

[33] Y. Li and Y. Yuan. Convergence analysis of two-layer neural networks with ReLU activation. *arXiv preprint arXiv:1705.09886*, 2017.

[34] R. Livni, S. Shalev-Shwartz, and O. Shamir. On the computational efficiency of training neural networks. In *Advances in Neural Information Processing Systems*, pages 855–863, 2014.

[35] S. Mei, Y. Bai, and A. Montanari. The landscape of empirical risk for non-convex losses. *arXiv preprint arXiv:1607.06534*, 2016.

[36] A. Mohamed, G. E. Dahl, and G. Hinton. Acoustic modeling using deep belief networks. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):14–22, 2012.

[37] Y. Nesterov and B. T. Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[38] Quynh Nguyen and Matthias Hein. The loss surface of deep and wide neural networks. *arXiv preprint arXiv:1704.08045*, 2017.

[39] B. T. Polyak. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.

[40] T. Poston, C-N. Lee, Y. Choie, and Y. Kwon. Local minima and back propagation. In *Neural Networks, 1991., IJCNN-91-Seattle International Joint Conference on*, volume 2, pages 173–176. IEEE, 1991.

[41] D. Saad and S. A. Solla. On-line learning in soft committee machines. *Physical Review E*, 52(4):4225, 1995.

[42] I. Safran and O. Shamir. Spurious local minima are common in two-layer ReLU neural networks. *arXiv preprint arXiv:1712.08968*, 2017.

[43] S. Shalev-Shwartz, O. Shamir, and K. Sridharan. Learning kernel-based halfspaces with the 0-1 loss. *SIAM Journal on Computing*, 40(6):1623–1646, 2011.

[44] M. Soltanolkotabi. *Algorithms and Theory for Clustering and Nonconvex Quadratic Programming*. PhD thesis, Stanford University, 2014.

[45] M. Soltanolkotabi. Learning ReLUs via gradient descent. *arXiv preprint arXiv:1705.04591*, 2017.

[46] M. Soltanolkotabi. Structured signal recovery from quadratic measurements: Breaking sample complexity barriers via nonconvex optimization. *arXiv preprint arXiv:1702.06175*, 2017.

[47] D. Soudry and Y. Carmon. No bad local minima: Data independent training error guarantees for multilayer neural networks. *arXiv preprint arXiv:1605.08361*, 2016.

[48] D. Soudry and E. Hoffer. Exponentially vanishing sub-optimal local minima in multilayer neural networks. *arXiv preprint arXiv:1702.05777*, 2017.

[49] M. Telgarsky. Benefits of depth in neural networks. *arXiv preprint arXiv:1602.04485*, 2016.

[50] Y. Tian. An analytical formula of population gradient for two-layered relu network and its applications in convergence and critical point analysis. *International Conference on Machine Learning (ICML)*, 2017.

[51] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[52] R. Vicente and N. Caticha. Functional optimization of online algorithms in multilayer neural networks. *Journal of Physics A: Mathematical and General*, 30(17):L599, 1997.

[53] T. Wiatowski, P. Grohs, and H. Bolcskei. Energy propagation in deep convolutional neural networks. *arXiv preprint arXiv:1704.03636*, 2017.

[54] B. Xie, Y. Liang, and L. Song. Diversity leads to generalization in neural networks. *arXiv preprint arXiv:1611.03131*, 2016.

[55] C. Zhang, S. Y. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

[56] Q. Zhang, R. Panigrahy, S. Sachdeva, and A. Rahimi. Electron-proton dynamics in deep learning. *arXiv preprint arXiv:1702.00458*, 2017.

[57] Y. Zhang, J. D. Lee, and M. I. Jordan. L1-regularized neural networks are improperly learnable in polynomial time. In *International Conference on Machine Learning*, pages 993–1001, 2016.

[58] K. Zhong, Z. Song, P. Jain, P. L. Bartlett, and I. S. Dhillon. Recovery guarantees for one-hidden-layer neural networks. *arXiv preprint arXiv:1706.03175*, 2017.

A   Proof of preliminary lemmas

A.1   Proof of Lemma 5.3

If $\sigma_{\min}(\boldsymbol{A}) = 0$ then the claim is obvious. We therefore assume that $\boldsymbol{A}$ has full column rank. Let $i^* = \arg\max_{1 \le i \le k} |v_i|$. We choose $\boldsymbol{u}$ such that $\boldsymbol{A}\boldsymbol{u} = e_{i^*}$. We then have

$$\|\boldsymbol{v}\|_{\ell_\infty} = |v_{i^*}| = \|\boldsymbol{D_v}\boldsymbol{A}\boldsymbol{u}\| \le \sigma_{\max}(\boldsymbol{D_v}\boldsymbol{A})\|\boldsymbol{u}\|. \tag{A.1}$$

We also have

$$\sigma_{\min}(\boldsymbol{A})\|\boldsymbol{u}\| \le \|\boldsymbol{A}\boldsymbol{u}\| = 1. \tag{A.2}$$

Equations (A.1) and (A.2) together implies the desired result.

A.2   Proof of Lemma 5.5

We have

$$\begin{aligned}
\mathbb{E}|Y|^p &= \int_0^\infty \mathbb{P}\{|Y|^p \ge u\}\mathrm{d}u \\
&= \int_0^\infty \mathbb{P}\{|Y| \ge t\}pt^{p-1}\mathrm{d}t \\
&= \int_0^{\frac{A}{B}} e^{-c\frac{t^2}{A}}pt^{p-1}\mathrm{d}t + \int_{\frac{A}{B}}^\infty e^{-c\frac{t}{B}}pt^{p-1}\mathrm{d}t \\
&\le p\left(\int_0^\infty e^{-c\frac{t^2}{A}}t^{p-1}\mathrm{d}t + \int_0^\infty e^{-c\frac{t}{B}}t^{p-1}\mathrm{d}t\right) \\
&= \left(\frac{A}{c}\right)^{\frac{p}{2}}\Gamma\left(\frac{p}{2}\right) + \left(\frac{B}{c}\right)^p\Gamma(p) \\
&\le \left(\frac{A}{2c}p\right)^{\frac{p}{2}} + \left(\frac{B}{c}p\right)^p \le 2\max\left(\sqrt{\frac{A}{2c}}p, \frac{B}{c}p\right)^p \le 2p^p\max\left(\sqrt{\frac{A}{2c}}, \frac{B}{c}\right)^p. \tag{A.3}
\end{aligned}$$

Let $\xi = \max(\sqrt{\frac{A}{2c}}, \frac{B}{c})$. This implies that

$$\mathbb{E}\exp(|Y|/C) = 1 + \sum_{p=1}^\infty \frac{\mathbb{E}|Y|^p}{C^p p!} \le 1 + 2\sum_{p=1}^\infty \left(\frac{\xi e}{C}\right)^p = 1 + \frac{2\xi e}{C - \xi e}.$$

The first inequality follows from (A.3); in the second one we use $p! \ge (p/e)^p$. Therefore, we have $\mathbb{E}\exp(|Y|/C) \le 2$, if $C \ge 3e\xi$. This yields $\|Y\|_{\psi_1} \le 9\max(\sqrt{\frac{A}{2c}}, \frac{B}{c})$.

A.3   Proof of Lemma 5.6

To prove this result we define the mapping $\mathcal{O} : \mathbb{R}^{d \times n} \mapsto \mathbb{R}^{d^2 \times n}$ as $\mathcal{O}(\boldsymbol{X}) = \boldsymbol{X} * \boldsymbol{X}$. To show that for almost every $\boldsymbol{X}$, we have $\mathrm{rank}(\mathcal{O}(\boldsymbol{X_0})) = n$ when $n \ge d$ we first construct $\boldsymbol{X_0} \in \mathbb{R}^{d \times n}$, such that $\mathrm{rank}(\mathcal{O}(\boldsymbol{X_0})) = n$. To this aim define the following two sets of vectors in $\mathbb{R}^d$:

$$\begin{aligned}
\mathcal{S}_1 &\equiv \{\boldsymbol{e_i} : 1 \le i \le d\}, \\
\mathcal{S}_2 &\equiv \{\boldsymbol{e_i} + \boldsymbol{e_j} : 1 \le i < j \le d\},
\end{aligned}$$

where $\boldsymbol{e_i}$ is the i-th element of the standard basis with one at the i-th position and zero everywhere else. Let $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$ and note that $|\mathcal{S}| = d(d+1)/2$. We construct $\boldsymbol{X_0} \in \mathbb{R}^{d \times n}$ by choosing n arbitrary vectors in $\mathcal{S}$ as its columns. In order to show that $\mathcal{O}(\boldsymbol{X_0})$ is full rank, we index its rows by pairs (i,j), with $1 \le i \le j \le d$. For any $i \ne j$, and note that $(\boldsymbol{e_i} + \boldsymbol{e_j}) \otimes (\boldsymbol{e_i} + \boldsymbol{e_j})$ is the only vector in $\mathcal{S}^{\otimes} \equiv \{\boldsymbol{v} \otimes \boldsymbol{v} : \boldsymbol{v} \in \mathcal{S}\}$ which is nonzero at the (i,j)-th coordinate. This implies that the vectors in $\mathcal{S}_2^{\otimes}$ are linearly independent of those in $\mathcal{S}^{\otimes}$. Therefore, it suffices to show that $\mathcal{S}_1^{\otimes}$ is a linearly independent set. To see this note that for any $1 \le i \le d$, the vector $\boldsymbol{e_i} \otimes \boldsymbol{e_i}$ is the only vector in $\mathcal{S}_1^{\otimes}$ which is nonzero at the (i,i) position. Hence, $\mathcal{S}_1^{\otimes}$ is also a set of linearly independent vectors. This completes the proof of showing that $\mathcal{O}(\boldsymbol{X_0})$ is full rank.

We now show that $\mathcal{O}(\boldsymbol{X})$ is full rank for almost every $\boldsymbol{X}$. To this aim define the polynomial mapping $g : \mathbb{R}^{d^2 \times n} \mapsto \mathbb{R}$, with $g(\boldsymbol{A})$ denoting the the sum of the squares of the determinants of all possible different subsets of n rows from $\boldsymbol{A}$. Since $n \le d^2$, $\mathcal{O}(\boldsymbol{X})$ is rank deficient (i.e. does not have full column rank) if and only if $g(\mathcal{O}(\boldsymbol{X})) = 0$. Moreover, as we showed above, $\mathcal{O}(\boldsymbol{X_0})$ is full rank and hence $g(\mathcal{O}(\boldsymbol{X_0})) \ne 0$. Hence, $g(\mathcal{O}(\boldsymbol{X}))$ is not identically zero and since it is a polynomial mapping, its zeros –which correspond to matrices $\boldsymbol{X}$ such that $\mathcal{O}(\boldsymbol{X})$ is singular– are a set of measure zero. This completes the proof.

A.4   Proof of Lemma 5.7

We begin by considering the eigenvalue decomposition of the matrix $\boldsymbol{A}$ given by

$$\boldsymbol{A} = \sum_{i=1}^{d} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^T.$$

Note that we have

$$\frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{x}_i^T \boldsymbol{A} \boldsymbol{x}_i \right) \boldsymbol{x}_i \boldsymbol{x}_i^T - (2\boldsymbol{A} + \mathrm{trace}(\boldsymbol{A})\boldsymbol{I}) = \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{x}_i^T \left(\sum_{j=1}^{d} \lambda_j \boldsymbol{v}_j \boldsymbol{v}_j^T \right) \boldsymbol{x}_i \right) \boldsymbol{x}_i \boldsymbol{x}_i^T - \left(2 \sum_{j=1}^{d} \lambda_j \boldsymbol{v}_j \boldsymbol{v}_j^T + \left(\sum_{j=1}^{d} \lambda_j \right) \boldsymbol{I} \right),$$

$$= \sum_{j=1}^{d} \lambda_j \left(\frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{x}_i^T \boldsymbol{v}_i)^2 \boldsymbol{x}_i \boldsymbol{x}_i^T - \left(2\boldsymbol{v}_i \boldsymbol{v}_i^T + \boldsymbol{I} \right) \right). \qquad (A.4)$$

To continue we state a Lemma due to [12] (see also [44, 46] for closely related results).

Lemma A.1 *[12, Lemma 7.4] Let $\boldsymbol{a} \in \mathbb{R}^d$ be a fixed vector and for $i = 1, 2, \ldots, n$, let $\boldsymbol{x}_i$ be distributed i.i.d. $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$. Then as long as*

$$n \ge c(\delta) d \log d,$$

with c a constant depending only on δ. Then

$$\left\| \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{x}_i^T \boldsymbol{a})^2 \boldsymbol{x}_i \boldsymbol{x}_i^T - \left(2\boldsymbol{a}\boldsymbol{a}^T + \|\boldsymbol{a}\|_{\ell_2}^2 \boldsymbol{I} \right) \right\| \le \delta \|\boldsymbol{a}\|_{\ell_2}^2,$$

holds with probability at least $1 - 10e^{-\gamma d} - 8/d^2$ with γ a fixed numerical constant.

Applying the union bound to the above lemma we conclude that for the unit norm vectors $\{\boldsymbol{v}_j\}_{j=1}^{d}$

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \left| \boldsymbol{x}_i^T \boldsymbol{v}_j \right|^2 \boldsymbol{x}_i \boldsymbol{x}_i^T - \left(2\boldsymbol{v}_j \boldsymbol{v}_j^T + \boldsymbol{I} \right) \right\| \le \delta,$$

hold simultaneously with probability at least $1 - 10de^{-\gamma d} - 8/d$. Combining the latter inequalities together with (A.4) we conclude that

$$
\left\| \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{x}_i^T \boldsymbol{A} \boldsymbol{x}_i \right) \boldsymbol{x}_i \boldsymbol{x}_i^T - (2\boldsymbol{A} + \operatorname{trace}(\boldsymbol{A})\boldsymbol{I}) \right\| = \left\| \sum_{j=1}^{d} \lambda_j \left(\frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{x}_i^T \boldsymbol{v}_i)^2 \boldsymbol{x}_i \boldsymbol{x}_i^T - \left(2\boldsymbol{v}_i \boldsymbol{v}_i^T + \boldsymbol{I} \right) \right) \right\|,
$$

$$
\leq \sum_{j=1}^{d} \lambda_j \left\| \frac{1}{n} \sum_{i=1}^{n} \left| \boldsymbol{x}_i^T \boldsymbol{v}_j \right|^2 \boldsymbol{x}_i \boldsymbol{x}_i^T - \left(2\boldsymbol{v}_j \boldsymbol{v}_j^T + \boldsymbol{I} \right) \right\|,
$$

$$
\leq \delta \left(\sum_{j=1}^{d} \lambda_j \right),
$$

$$
= \delta \cdot \operatorname{trace}(\boldsymbol{A}),
$$

completing the proof.

B   Proof of Proposition 7.1

Recalling definition of $\boldsymbol{\mu}$ given by (7.1), we first show that $\mathbb{E}[\phi'(\boldsymbol{W}\boldsymbol{x})] = \boldsymbol{\mu}$. Due to rotational invariance of Gaussian distribution, without loss of generality, we assume that $\boldsymbol{w_i} = \|\boldsymbol{w_i}\|\boldsymbol{e_1}$. Therefore,

$$
\mathbb{E}[\phi'(\boldsymbol{w_i}^T \boldsymbol{x})] = \mathbb{E}[\phi'(\|\boldsymbol{w_i}\| x_1)] = \mu_i \,.
$$

Consequently, we have $\mathbb{E}[\phi'(\boldsymbol{W}\boldsymbol{X})] = \boldsymbol{\mu} \boldsymbol{1}_n^T$.

We shall try to deduce a bound on the minimum singular value of $\boldsymbol{J}$ via a lower bound on the minimum singular value of $\tilde{\boldsymbol{J}} = \boldsymbol{X} * \boldsymbol{D_v}(\phi'(\boldsymbol{W}\boldsymbol{X}) - \boldsymbol{\mu}\boldsymbol{1}_n^T)$. Define $\boldsymbol{\Delta} = \boldsymbol{D_v}(\phi'(\boldsymbol{W}\boldsymbol{X}) - \boldsymbol{\mu}\boldsymbol{1}_n^T)$ and $\widetilde{\boldsymbol{\mu}} = \boldsymbol{D_v}\boldsymbol{\mu}$. By Lemma 5.4, we have

$$
\boldsymbol{J}^T \boldsymbol{J} = \left(\boldsymbol{X} * (\boldsymbol{\Delta} + \widetilde{\boldsymbol{\mu}}\boldsymbol{1}_n^T) \right)^T \left(\boldsymbol{X} * (\boldsymbol{\Delta} + \widetilde{\boldsymbol{\mu}}\boldsymbol{1}_n^T) \right)
$$

$$
= (\boldsymbol{X}^T \boldsymbol{X}) \circ \left(\boldsymbol{\Delta}^T \boldsymbol{\Delta} + \boldsymbol{\Delta}^T \widetilde{\boldsymbol{\mu}}\boldsymbol{1}_n^T + \boldsymbol{1}_n \widetilde{\boldsymbol{\mu}}^T \boldsymbol{\Delta} + \|\widetilde{\boldsymbol{\mu}}\|_{\ell_2}^2 \boldsymbol{1}_n \boldsymbol{1}_n^T \right) \,.
$$

Similarly we have

$$
\widetilde{\boldsymbol{J}}^T \widetilde{\boldsymbol{J}} = (\boldsymbol{X} * \boldsymbol{\Delta})^T (\boldsymbol{X} * \boldsymbol{\Delta}) = (\boldsymbol{X}^T \boldsymbol{X}) \circ \left(\boldsymbol{\Delta}^T \boldsymbol{\Delta} \right) \,.
$$

Subtracting the above two expressions we obtain

$$
\boldsymbol{J}^T \boldsymbol{J} = \widetilde{\boldsymbol{J}}^T \widetilde{\boldsymbol{J}} + (\boldsymbol{X}^T \boldsymbol{X}) \circ \left(\boldsymbol{\Delta}^T \widetilde{\boldsymbol{\mu}}\boldsymbol{1}_n^T + \boldsymbol{1}_n \widetilde{\boldsymbol{\mu}}^T \boldsymbol{\Delta} + \|\widetilde{\boldsymbol{\mu}}\|_{\ell_2}^2 \boldsymbol{1}_n \boldsymbol{1}_n^T \right) \,.
$$

Now note that the trivial identity

$$
\left(\|\widetilde{\boldsymbol{\mu}}\|_{\ell_2} \boldsymbol{1}_n + \frac{\boldsymbol{\Delta}^T \widetilde{\boldsymbol{\mu}}}{\|\widetilde{\boldsymbol{\mu}}\|_{\ell_2}} \right) \left(\|\widetilde{\boldsymbol{\mu}}\|_{\ell_2} \boldsymbol{1}_n + \frac{\boldsymbol{\Delta}^T \widetilde{\boldsymbol{\mu}}}{\|\widetilde{\boldsymbol{\mu}}\|_{\ell_2}} \right)^T \succeq \boldsymbol{0} \,,
$$

along with the fact that the Hadamard (entrywise) product of two positive-semidefinite matrices is positive-semidefinite, implies that

$$
(\boldsymbol{X}^T \boldsymbol{X}) \circ \left(\boldsymbol{\Delta}^T \widetilde{\boldsymbol{\mu}}\boldsymbol{1}_n^T + \boldsymbol{1}_n \widetilde{\boldsymbol{\mu}}^T \boldsymbol{\Delta} + \|\widetilde{\boldsymbol{\mu}}\|_{\ell_2}^2 \boldsymbol{1}_n \boldsymbol{1}_n^T \right) \succeq - (\boldsymbol{X}^T \boldsymbol{X}) \circ \frac{(\boldsymbol{\Delta}^T \widetilde{\boldsymbol{\mu}})(\boldsymbol{\Delta}^T \widetilde{\boldsymbol{\mu}})^T}{\|\widetilde{\boldsymbol{\mu}}\|_{\ell_2}^2} \,.
$$

Thus it suffices to show that the matrix

$$M := \frac{1}{\|\widetilde{\boldsymbol{\mu}}\|_{\ell_2}^2} \left(\boldsymbol{X}^T \boldsymbol{X} \right) \circ \left(\boldsymbol{\Delta}^T \widetilde{\boldsymbol{\mu}} \right) \left(\boldsymbol{\Delta}^T \widetilde{\boldsymbol{\mu}} \right)^T$$

obeys

$$\|\boldsymbol{M}\| \le C d^2 L^2 \|\boldsymbol{v}\|_{\ell_\infty}^2 \|\boldsymbol{W}\|^2$$

with high probability. To this aim note that the latter matrix can be written in the form $\boldsymbol{Y}^T \boldsymbol{Y}$ where

$$\boldsymbol{Y} = \boldsymbol{X} \operatorname{diag}(\boldsymbol{d}) \quad \text{with} \quad \boldsymbol{d} = \frac{1}{\|\widetilde{\boldsymbol{\mu}}\|_{\ell_2}} \boldsymbol{\Delta}^T \widetilde{\boldsymbol{\mu}}.$$

To continue we state the following result from [3].

Theorem B.1 *([3, Theorem 1]) Let $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n \in \mathbb{R}^d$ with $n \ge d$ be independent random vectors satisfying*

$$\sup_{i=1,2,\ldots,n} \|\boldsymbol{y}_i\|_{\psi_1} \le \psi, \tag{B.1}$$

and

$$\mathbb{P}\left\{ \max_{i=1,2,\ldots,n} \|\boldsymbol{y}_i\|_{\ell_2} > K \sqrt[4]{nd} \right\} \le e^{-\sqrt{d}}, \tag{B.2}$$

for some $K \ge 1$. Define $\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{y}_1 & \boldsymbol{y}_2 & \ldots & \boldsymbol{y}_n \end{bmatrix}$ and $\boldsymbol{\Sigma} := \mathbb{E}[\boldsymbol{y}_1 \boldsymbol{y}_1^T]$. Then with probability at least $1 - 2e^{-c\sqrt{d}}$ the following holds:

$$\left\| \boldsymbol{Y} \boldsymbol{Y}^T - n \boldsymbol{\Sigma} \right\| \le C \left(\psi + K \right)^2 \sqrt{dn}$$

To use this theorem we need to verify the assumptions. We first verify the sub-exponential assumption. Observe that $\boldsymbol{d}_i$ are generated i.i.d. according to the distribution

$$f(\boldsymbol{x}) = \frac{1}{\|\widetilde{\boldsymbol{\mu}}\|_{\ell_2}} \widetilde{\boldsymbol{\mu}}^T \left(\boldsymbol{D_v} \phi'(\boldsymbol{W} \boldsymbol{x}) - \widetilde{\boldsymbol{\mu}} \right).$$

The gradient of this function reads as

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \frac{1}{\|\widetilde{\boldsymbol{\mu}}\|_{\ell_2}} \boldsymbol{W}^T \operatorname{diag} \left(\phi''(\boldsymbol{W} \boldsymbol{x}) \right) \left(\boldsymbol{v} \circ \widetilde{\boldsymbol{\mu}} \right).$$

Thus

$$\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x})\|_{\ell_2} \le L \|\boldsymbol{v}\|_{\ell_\infty} \|\boldsymbol{W}\| \tag{B.3}$$

This in turn implies that $f(\boldsymbol{x})$ is an $L \|\boldsymbol{v}\|_{\ell_\infty} \|\boldsymbol{W}\|$-Lipshitz function of a Gaussian vector and is thus a sub-Gaussian random variable with sub-Gaussian norm $cL \|\boldsymbol{v}\|_{\ell_\infty} \|\boldsymbol{W}\|$, for some absolute

constant $c > 0$. Given that $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$, $\boldsymbol{y}_i$ is sub-exponential with sub-exponential norm at most $\psi = cL \|\boldsymbol{v}\|_{\ell_\infty} \|\boldsymbol{W}\|$.

We next verify the norm bound condition (B.2). Note that we have

$$\|\boldsymbol{y}_i\|_{\ell_2} = \|\boldsymbol{x}_i\|_{\ell_2} |\boldsymbol{d}_i|$$

Since $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$, we have $\|\boldsymbol{x}_i\|_{\ell_2} \geq 2\sqrt{d}$ with probability at most e^{-cd}. Hence,

$$\max_{i \in [n]} \|\boldsymbol{x}_i\|_{\ell_2} \geq 2\sqrt{d},$$

holds with probability at most ne^{-cd}. The latter is smaller than $\frac{1}{2}e^{-\sqrt{d}}$ as long as $n \leq e^{c'd}$. In addition, as mentioned in the proof of sub-exponential in the previous part we have that $\boldsymbol{d}_i$ is a sub-Gaussian random variable with sub-Gaussian norm $cL \|\boldsymbol{v}\|_{\ell_\infty} \|\boldsymbol{W}\|$. Thus,

$$\mathbb{P}\left\{ \max_i |\boldsymbol{d}_i| \geq ctL \|\boldsymbol{v}\|_{\ell_\infty} \|\boldsymbol{W}\| \right\} \leq 2ne^{-t^2}.$$

Therefore, when $n \leq \frac{1}{4}e^{\sqrt{d}}$ using

$$t = 2\sqrt[4]{d} \geq \sqrt{\log(4n)} + \sqrt[4]{d} \quad \Rightarrow \quad t^2 \geq \log(4n) + \sqrt{d}$$

we have that

$$\mathbb{P}\left\{ \max_i |\boldsymbol{d}_i| \geq ctL \|\boldsymbol{v}\|_{\ell_\infty} \|\boldsymbol{W}\| \right\} \leq 2ne^{-t^2} \leq \frac{1}{2}e^{-\sqrt{d}}.$$

Putting the latter two together we conclude that

$$\mathbb{P}\left\{ \|\boldsymbol{y}_i\|_{\ell_2} \geq cLd^{\frac{3}{4}} \|\boldsymbol{v}\|_{\ell_\infty} \|\boldsymbol{W}\| \right\} \leq \frac{1}{2}e^{-\sqrt{d}} + \frac{1}{2}e^{-\sqrt{d}} \leq e^{-\sqrt{d}}.$$

Thus we can take

$$K = \frac{c\sqrt{d}}{\sqrt[4]{n}} L \|\boldsymbol{v}\|_{\ell_\infty} \|\boldsymbol{W}\|$$

in (B.2). Note that this choice of K obeys $K \geq 1$ as long as

$$n \leq cd^2 L^4 \|\boldsymbol{v}\|_{\ell_\infty}^4 \|\boldsymbol{W}\|^4 .$$

Thus using Theorem B.1 we conclude that as long as $n \leq cd^2 L^4 \|\boldsymbol{v}\|_{\ell_\infty}^4 \|\boldsymbol{W}\|^4$ we have

$$\left\| \boldsymbol{Y}\boldsymbol{Y}^T - n\boldsymbol{\Sigma} \right\| \leq C\left(d^{\frac{3}{2}} + \sqrt{dn} \right) L^2 \|\boldsymbol{v}\|_{\ell_\infty}^2 \|\boldsymbol{W}\|^2 \tag{B.4}$$

We next upper bound $\|\boldsymbol{\Sigma}\|$. Recall the definition

$$f(\boldsymbol{x}) = \frac{1}{\|\widetilde{\boldsymbol{\mu}}\|_{\ell_2}} \widetilde{\boldsymbol{\mu}}^T \left(\boldsymbol{D}_{\boldsymbol{v}} \phi'(\boldsymbol{W}\boldsymbol{x}) - \widetilde{\boldsymbol{\mu}} \right),$$

and note that

$$\begin{aligned}
\|\mathbf{\Sigma}\| &= \|\boldsymbol{y}_1 \boldsymbol{y}_1^T\| \\
&= \sup_{\boldsymbol{u} \in \mathbb{S}^{d-1}} \boldsymbol{u}^T \left(\mathbb{E}\left[f^2(\boldsymbol{x})\boldsymbol{x}\boldsymbol{x}^T\right]\right) \boldsymbol{u} \\
&= \sup_{\boldsymbol{u} \in \mathbb{S}^{d-1}} \mathbb{E}\left[f^2(\boldsymbol{x})\left(\boldsymbol{x}^T\boldsymbol{u}\right)^2\right] \\
&\leq \sup_{\boldsymbol{u} \in \mathbb{S}^{d-1}} \sqrt{\mathbb{E}\left[f^4(\boldsymbol{x})\right]}\sqrt{\mathbb{E}\left[\left(\boldsymbol{x}^T\boldsymbol{u}\right)^4\right]},
\end{aligned}$$

where the last line follows from Cauchy-Schwarz inequality. To continue note that since $f(\boldsymbol{x})$ is an $R \coloneqq L \|\boldsymbol{v}\|_{\ell_\infty} \|\boldsymbol{W}\|$-Lipshitz function of a Gaussian vector, using the tail bound inequality (see e.g., [51, Proposition 5.34]), we have

$$\mathbb{P}(|f(\boldsymbol{x})| > t) \leq 2e^{-\frac{t^2}{2B^2}}.$$

Thus,

$$\begin{aligned}
\mathbb{E}[f^4(\boldsymbol{x})] &= \int_0^\infty \mathbb{P}(|f(\boldsymbol{x})| \geq t)4t^3 \mathrm{d}t \\
&\leq \int_0^\infty 8e^{-\frac{t^2}{2R^2}}t^3 \mathrm{d}t \\
&= 32R^4 \int_0^\infty e^{-t^2}t^3 \mathrm{d}t \\
&= 16R^4 = 16L^4 \|\boldsymbol{v}\|_{\ell_\infty}^4 \|\boldsymbol{W}\|^4.
\end{aligned}$$

We also have $\mathbb{E}\left[\left(\boldsymbol{x}^T\boldsymbol{u}\right)^4\right] = 3$. Thus,

$$\|\mathbf{\Sigma}\| \leq 4\sqrt{3}L^2 \|\boldsymbol{v}\|_{\ell_\infty}^2 \|\boldsymbol{W}\|^2.$$

Combining the latter with (B.4) we can conclude that as long as $n \leq cd^2$ we have

$$\begin{aligned}
\left\|\boldsymbol{Y}\boldsymbol{Y}^T\right\| &\leq \left\|\boldsymbol{Y}\boldsymbol{Y}^T - n\mathbf{\Sigma}\right\| + n\|\mathbf{\Sigma}\| \\
&\leq C\left(d^{\frac{3}{2}} + \sqrt{dn}\right)L^2 \|\boldsymbol{v}\|_{\ell_\infty}^2 \|\boldsymbol{W}\|^2 + 4\sqrt{3}nL^2 \|\boldsymbol{v}\|_{\ell_\infty}^2 \|\boldsymbol{W}\|^2
\end{aligned}$$

concluding the proof.

We next prove the bound in (7.5). Note that

$$\sigma_{\max}(\boldsymbol{J}) \leq \sigma_{\max}(\tilde{\boldsymbol{J}}) + \|\boldsymbol{D}_{\boldsymbol{v}}\boldsymbol{\mu}\|\sigma_{\max}(\boldsymbol{X}).$$

By [51, Corollary 5.35], $\sigma_{\max}(\boldsymbol{X}) \leq \sqrt{d}+2\sqrt{n}$, with probability at least $1-2e^{-n/2}$. The result follows be recalling that $n \geq d$.

C   Proof of Proposition 7.4

We first prove the claim under Assumption 2.3 (a). We start by bounding the entries of $\mathbf{\Gamma}$ and mean vector $\boldsymbol{\mu}$.

Recall that $\gamma_\phi(\sigma) = \mathbb{E}(\phi''(\sigma g))$, where $g \sim \mathcal{N}(0, 1)$. Also not that the function $|\gamma_\phi(\sigma)|$ is continuous and always positive by Assumption 2.3 (a). Therefore it attains its minimum over any compact set. Since $\|\boldsymbol{w}_\ell\|_{\ell_2} \in [w_{\min}, w_{\max}]$, for all $1 \leq \ell \leq k$, there exists a constant $\gamma_{\min} > 0$, such that $|\gamma_\phi(\|\boldsymbol{w}_\ell\|_{\ell_2})| \geq \gamma_{\min}$, for all $1 \leq \ell \leq k$. Furthermore, $|\phi''| \leq L$ implies $|\gamma_\phi(\sigma)| = |\mathbb{E}[\phi''(\sigma g)]| \leq L$. Hence, for $1 \leq \ell \leq k$, we have

$$0 < \gamma_{\min} < |\Gamma_{\ell\ell}| \leq L. \tag{C.1}$$

By a similar argument, we have $|\mu_\ell| > \mu_{\min} > 0$, for some constant $\mu_{\min}$ and for all $1 \leq \ell \leq k$. Furthermore,

$$
\begin{aligned}
\|\boldsymbol{\mu}\|_{\ell_2} &= \left\|\mathbb{E}[\phi'(\boldsymbol{W}\boldsymbol{x})]\right\|_{\ell_2}, \\
&\overset{(a)}{\leq} \mathbb{E}[\left\|\phi'(\boldsymbol{W}\boldsymbol{x})\right\|_{\ell_2}], \\
&\overset{(b)}{=} \mathbb{E}\left[\left\|\phi'(\boldsymbol{0}) + \mathrm{diag}\left(\int_0^1 \phi''(t\boldsymbol{W}\boldsymbol{x})\mathrm{d}t\right)\boldsymbol{W}\boldsymbol{x}\right\|_{\ell_2}\right], \\
&\overset{(c)}{\leq} \mathbb{E}\left[\left\|\phi'(0)\boldsymbol{1}_k\right\|_{\ell_2} + L\sigma_{\max}(\boldsymbol{W})\|x\|_{\ell_2}\right], \\
&\overset{(d)}{\leq} \phi'(0)\sqrt{k} + L\sigma_{\max}(\boldsymbol{W})\sqrt{d}.
\end{aligned}
$$

Here (a) follows from Jenson's inequality, (b) follows from Taylor's theorem, (c) from the triangular inequality together with the fact that $|\phi''| \leq L$ and the definition of the maximum eigenvalue, and (d) follows from Jenson's inequality which implies $(\mathbb{E}[\|\boldsymbol{x}\|_{\ell_2}])^2 \leq \mathbb{E}[\|\boldsymbol{x}\|_{\ell_2}^2] = d$. In conclusion, we have

$$\mu_{\min}\sqrt{k} \leq \|\boldsymbol{\mu}\|_{\ell_2} \leq \sqrt{k} + L\sigma_{\max}(\boldsymbol{W})\sqrt{d}. \tag{C.2}$$

We next verify the assumption of Proposition 7.1. By our assumptions in the current proposition,

$$\frac{n}{d^2} \leq \frac{c_0 \sigma_{\min}^4(\boldsymbol{W})}{\sigma_{\max}^8(\boldsymbol{W})} \leq \frac{c_0 w_{\max}^4}{w_{\min}^8}, \tag{C.3}$$

since for an arbitrary row $\ell \in [k]$, $\sigma_{\min}(\boldsymbol{W}) \leq \|\boldsymbol{w}_\ell\|_{\ell_2} \leq w_{\max}$. Also, $\sigma_{\max}(\boldsymbol{W}) \geq \|\boldsymbol{w}_\ell\|_{\ell_2} \geq w_{\min}$. For

$$c_0 < c\left(Lv_{\min}\frac{w_{\min}^3}{w_{\max}}\right)^4, \tag{C.4}$$

we have

$$\frac{n}{d^2} \leq \frac{c_0 w_{\max}^4}{w_{\min}^8} \leq c(Lv_{\min}w_{\min})^4 \leq c(L\|\boldsymbol{v}\|_{\ell_\infty}\sigma_{\max}(\boldsymbol{W}))^4.$$

Therefore, the assumption of Proposition 7.1 holds. Using this proposition, the following bounds hold with probability at least $1 - 2e^{-c'\sqrt{d}}$:

$$\sigma_{\min}^2(\boldsymbol{J}) \geq \sigma_{\min}^2(\tilde{\boldsymbol{J}}) - \left(4\sqrt{3}n + Cd^{\frac{3}{2}} + C\sqrt{dn}\right)L^2\|\boldsymbol{v}\|_{\ell_\infty}^2\|\boldsymbol{W}\|^2. \tag{C.5}$$

We next lower bound the right-hand side. By our assumption on n/d^2, we have

$$
\begin{aligned}
n\left\|\boldsymbol{W}\right\|^2 &\le \frac{c_0\sigma_{\min}^4(\boldsymbol{W})}{\sigma_{\max}^8(\boldsymbol{W})}\left\|\boldsymbol{W}\right\|^2 d^2 \\
&\le \frac{c_0\sigma_{\min}^2(\boldsymbol{W})}{\sigma_{\max}^6(\boldsymbol{W})}\sigma_{\min}^2(\boldsymbol{W})d^2 \\
&\le \frac{c_0 w_{\max}^2}{w_{\min}^6}\sigma_{\min}^2(\boldsymbol{W})d^2\,.
\end{aligned}
\tag{C.6}
$$

Likewise we have

$$
\begin{aligned}
\sqrt{nd}\left\|\boldsymbol{W}\right\|^2 &\le \frac{\sqrt{c_0}\sigma_{\min}^2(\boldsymbol{W})}{\sigma_{\max}^4(\boldsymbol{W})}\left\|\boldsymbol{W}\right\|^2 d^{3/2} \\
&\le \frac{\sqrt{c_0}}{\sigma_{\max}^2(\boldsymbol{W})}\sigma_{\min}^2(\boldsymbol{W})d^{3/2} \\
&\le \frac{\sqrt{c_0}}{w_{\min}^2}\sigma_{\min}^2(\boldsymbol{W})d^2\,.
\end{aligned}
\tag{C.7}
$$

Further, by invoking assumption (7.11), we have

$$
\begin{aligned}
\frac{\sigma_{\max}^8(\boldsymbol{W})}{\sigma_{\min}^4(\boldsymbol{W})c_0} \le d &\Rightarrow \frac{\left\|\boldsymbol{W}\right\|^2}{\sqrt{d}} \le \sqrt{c_0}\frac{\sigma_{\min}^2(\boldsymbol{W})}{\sigma_{\max}^2(\boldsymbol{W})} \\
&\Rightarrow \left\|\boldsymbol{W}\right\|^2 d^{3/2} \le \sqrt{c_0}\frac{\sigma_{\min}^2(\boldsymbol{W})}{\sigma_{\max}^2(\boldsymbol{W})}d^2 \le \frac{\sqrt{c_0}}{w_{\min}^2}\sigma_{\min}^2(\boldsymbol{W})d^2\,.
\end{aligned}
\tag{C.8}
$$

Putting (C.6), (C.7) and (C.8) together we arrive at

$$
\left(4\sqrt{3}n + Cd^{\frac{3}{2}} + C\sqrt{dn}\right)\left\|\boldsymbol{W}\right\|^2 \le \left(4\sqrt{3}\frac{c_0 w_{\max}^2}{w_{\min}^6} + 2C\frac{\sqrt{c_0}}{w_{\min}^2}\right)\sigma_{\min}^2(\boldsymbol{W})d^2\,.
$$

Using the above bound in (C.5) we get

$$
\sigma_{\min}^2\left(\boldsymbol{J}\right) \ge \sigma_{\min}^2\left(\tilde{\boldsymbol{J}}\right) - \left(4\sqrt{3}\frac{c_0 w_{\max}^2}{w_{\min}^6} + 2C\frac{\sqrt{c_0}}{w_{\min}^2}\right)L^2 v_{\max}^2\sigma_{\min}^2(\boldsymbol{W})d^2\,.
\tag{C.9}
$$

We then move to upper bound $\sigma_{\max}(\boldsymbol{J})$.

$$
\begin{aligned}
\sigma_{\max}(\boldsymbol{J}) &\le \sigma_{\max}(\tilde{\boldsymbol{J}}) + 3\sqrt{n}\left\|\boldsymbol{D_v}\boldsymbol{\mu}\right\|_{\ell_2}, \\
&\le \sigma_{\max}(\tilde{\boldsymbol{J}}) + 3\sqrt{n}v_{\max}(\sqrt{k} + L\sigma_{\max}(\boldsymbol{W})\sqrt{d}), \\
&\le \sigma_{\max}(\tilde{\boldsymbol{J}}) + 3\sqrt{nk}v_{\max}(L + 1/w_{\min})\sigma_{\max}(\boldsymbol{W})\,.
\end{aligned}
\tag{C.10}
$$

Here, in deriving the upper bound for $\sigma_{\max}(\boldsymbol{J})$, we used the bound on $\left\|\boldsymbol{\mu}\right\|_{\ell_2}$ from (C.2) in the first step. We then used the fact that $k \le d$ combined with $w_{\min} \le \sigma_{\max}(W)$ in the second step.

All that remains is to bound $\sigma_{\min}(\tilde{\boldsymbol{J}})$ and $\sigma_{\max}(\tilde{\boldsymbol{J}})$ by appealing to Proposition 7.2. Before applying this result however, we need to show that assumption (7.6) on the sample size holds. To

this aim note that

$$\frac{L^2\sigma_{\max}^4(\boldsymbol{W})+1}{\sigma_{\min}^2(\boldsymbol{\Gamma W})} \overset{(a)}{\leq} \frac{L^4\sigma_{\max}^4(\boldsymbol{W})+1}{\gamma_{\min}^2\sigma_{\min}^2(\boldsymbol{W})},$$

$$\overset{(b)}{\leq} \frac{\sigma_{\max}^4(\boldsymbol{W})}{\sigma_{\min}^2(\boldsymbol{W})} \cdot \frac{1}{\gamma_{\min}^2}\left(L^4 + \frac{1}{w_{\min}^4}\right),$$

$$\leq \frac{d}{\sqrt{n}} \cdot \frac{\sqrt{c_0}}{\gamma_{\min}^2}\left(L^4 + \frac{1}{w_{\min}^4}\right),$$

which immediately implies assumption (7.6). Here, (a) follows from $\sigma_{\min}(\boldsymbol{\Gamma W}) \geq \gamma_{\min}\sigma_{\min}(\boldsymbol{W})$, (b) from $\sigma_{\max}(\boldsymbol{W}) \geq w_{\min}$, and (c) from (7.11). With assumption (7.6) in place, we can now combine Proposition 7.2 with (C.9) to conclude that

$$\sigma_{\min}^2(\boldsymbol{J}) \geq \sigma_{\min}^2(\tilde{\boldsymbol{J}}) - \left(4\sqrt{3}\frac{c_0 w_{\max}^2}{w_{\min}^6} + 2C\frac{\sqrt{c_0}}{w_{\min}^2}\right)L^2 v_{\max}^2\sigma_{\min}^2(\boldsymbol{W})d^2$$

$$\geq \frac{v_{\min}^2\gamma_{\min}^2}{4}\sigma_{\min}^2(\boldsymbol{W})d^2 - \left(4\sqrt{3}\frac{c_0 w_{\max}^2}{w_{\min}^6} + 2C\frac{\sqrt{c_0}}{w_{\min}^2}\right)L^2 v_{\max}^2\sigma_{\min}^2(\boldsymbol{W})d^2$$

$$= \left\{\frac{v_{\min}^2\gamma_{\min}^2}{4} - \left(4\sqrt{3}\frac{c_0 w_{\max}^2}{w_{\min}^6} + 2C\frac{\sqrt{c_0}}{w_{\min}^2}\right)L^2 v_{\max}^2\right\}\sigma_{\min}^2(\boldsymbol{W})d^2$$

$$\geq c^2\sigma_{\min}^2(\boldsymbol{W})d^2,$$

for some constant $c > 0$ by choosing c_0 sufficiently small. In addition,

$$\sigma_{\max}(\boldsymbol{J}) \leq \sigma_{\max}(\tilde{\boldsymbol{J}}) + 3\sqrt{nk}v_{\max}(L + 1/w_{\min})\sigma_{\max}(\boldsymbol{W})$$

$$\leq C'(d + Lv_{\max}\sigma_{\max}(\boldsymbol{W})\sqrt{nk}) + 3\sqrt{nk}\,v_{\max}(L + 1/w_{\min})\sigma_{\max}(\boldsymbol{W})$$

$$\leq C'\,\sigma_{\max}(\boldsymbol{W})\sqrt{nk},$$

for a constants $C' > 0$ that depends on $v_{\max}$, $w_{\min}$, L. The claim under Assumption 2.3 (b) follows in a similar manner by using Remark 7.3 in lieu of Proposition 7.2.

D   Proof of Proposition 7.8

We prove Proposition 7.8 via an asymmetric version of Hanson-Wright inequality. We begin by the definition of the convex concentration property which is the most general condition under which it is known that Hanson-Wright inequality holds.

Definition D.1 (Convex concentration property) *Let $\boldsymbol{x}$ be a random vector in $\mathbb{R}^d$. We will say that $\boldsymbol{x}$ has the convex concentration property with constant K if for every 1-Lipschitz convex function $\psi : \mathbb{R}^n \to \mathbb{R}$, we have $\mathbb{E}[\psi(\boldsymbol{x})] < \infty$ and for every $t > 0$,*

$$\mathbb{P}\big\{|\psi(\boldsymbol{x}) - \mathbb{E}[\psi(\boldsymbol{x})]| \geq t\big\} \leq 2e^{-\frac{t^2}{2K^2}}.$$

We will prove the proposition by using a result by [1] on the Hanson-Wright inequality stated below.

Lemma D.2 *Let $\boldsymbol{u}$ be a mean zero random vector in $\mathbb{R}^d$. If $\boldsymbol{u}$ has the convex concentration property with constant K then for any matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ and every $t > 0$,*

$$\mathbb{P}\left\{ \left| \boldsymbol{u}^T \boldsymbol{A} \boldsymbol{u} - \mathbb{E}[\boldsymbol{u}^T \boldsymbol{A} \boldsymbol{u}] \right| \geq t \right\} \leq 2 \exp\left(-\frac{1}{C} \min\left(\frac{t^2}{2K^4 \|\boldsymbol{A}\|_F^2}, \frac{t}{K^2 \|\boldsymbol{A}\|} \right) \right).$$

We wish to apply this result to the vector $\boldsymbol{u} = (\boldsymbol{x}, \boldsymbol{z})^T$. Note that $\boldsymbol{u}$ has zero mean. To this aim, we first state a lemma about the convex concentration property for the random vector $\boldsymbol{u}$ whose proof appears in Section D.1 below.

Lemma D.3 *Let $\boldsymbol{x} \in \mathbb{R}^d$ be a random Gaussian vector distributed as $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$. Also assume that the nonlinear function $\phi : \mathbb{R} \to \mathbb{R}$ has bounded second derivative, i.e. $|\phi''| \leq L$. Then, the random vector $\boldsymbol{u} = (\boldsymbol{x}, \boldsymbol{z})^T$ obeys the convex concentration property with $K = (L^2 \sigma_{\max}^2(\boldsymbol{W}) + 1)^{1/2}$.*

Lemmas D.2 and D.3 combined allows us to conclude that

$$\mathbb{P}\left\{ \left| \boldsymbol{u}^T \boldsymbol{A} \boldsymbol{u} - \mathbb{E}[\boldsymbol{u}^T \boldsymbol{A} \boldsymbol{u}] \right| \geq t \right\} \leq 2 \exp\left(-\frac{1}{C} \min\left(\frac{t^2}{2K^4 \|\boldsymbol{A}\|_F^2}, \frac{t}{K^2 \|\boldsymbol{A}\|} \right) \right).$$

For $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{U} \\ \boldsymbol{U}^T & \boldsymbol{0} \end{bmatrix}$, this is equivalent to

$$\mathbb{P}\left\{ \left| \boldsymbol{x}^T \boldsymbol{U} \boldsymbol{z} - \mathbb{E}[\boldsymbol{x}^T \boldsymbol{U} \boldsymbol{z}] \right| \geq \frac{t}{2} \right\} \leq 2 \exp\left(-\frac{1}{C} \min\left(\frac{t^2}{2K^4 \|\boldsymbol{U}\|_F^2}, \frac{t}{K^2 \|\boldsymbol{U}\|} \right) \right).$$

The proof is complete by replacing t with $2t$.

D.1 Proof of Lemma D.3

Define $\boldsymbol{u}_1 = (\boldsymbol{x}_1, \boldsymbol{z}_1)^T$ and $\boldsymbol{u}_2 = (\boldsymbol{x}_2, \boldsymbol{z}_2)^T$. Note that for any convex function $\psi : \mathbb{R}^{2d} \to \mathbb{R}$ that is 1-Lipschitz we have

$$\begin{aligned}
|\psi(\boldsymbol{u}_2) - \psi(\boldsymbol{u}_1)|^2 &\leq \|\boldsymbol{u}_2 - \boldsymbol{u}_1\|_{\ell_2}^2 = \|\boldsymbol{x}_1 - \boldsymbol{z}_1\|_{\ell_2}^2 + \|\boldsymbol{x}_2 - \boldsymbol{z}_2\|_{\ell_2}^2 \\
&= \|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_{\ell_2}^2 + \|\phi'(\boldsymbol{W}\boldsymbol{x}_2) - \phi'(\boldsymbol{W}\boldsymbol{x}_1)\|_{\ell_2}^2 \\
&\leq \|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_{\ell_2}^2 + L^2 \|\boldsymbol{W}(\boldsymbol{x}_2 - \boldsymbol{x}_1)\|_{\ell_2}^2 \\
&\leq \left(1 + L^2 \sigma_{\max}^2(\boldsymbol{W})\right) \|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_{\ell_2}^2.
\end{aligned}$$

Thus $\psi(\boldsymbol{u})$ is a Lipschitz function of $\boldsymbol{x}$ with Lipschitz constant $K = \sqrt{1 + L^2 \sigma_{\max}^2(\boldsymbol{W})}$. The convex concentration property follows from Gaussian isoperimetry [29].

E Proof of Lemma 7.6

For $g \sim \mathcal{N}(0, 1)$ and $r \in \mathbb{R}$, we have

$$\mathbb{E}[\phi'(rg)g] = \frac{1}{r} \mathbb{E}[\phi'(rg)rg] = r\gamma_\phi(r).$$

Let us start by calculating $\mathbb{E}[\phi'(\boldsymbol{w}_i^T\boldsymbol{x})\boldsymbol{x}]$. Note that due to symmetry of the Gaussian distribution without loss of generality we can assume that $\boldsymbol{w}_i = \|\boldsymbol{w}_i\|_{\ell_2}\boldsymbol{e}_1$. In this case we have

$$\mathbb{E}[\phi'(\|\boldsymbol{w}_i\|_{\ell_2}\boldsymbol{e}_1^T\boldsymbol{x})\boldsymbol{x}] = \mathbb{E}[\phi'(\|\boldsymbol{w}_i\|_{\ell_2}x_1)x_1\boldsymbol{e}_1] + \mathbb{E}[\phi'(\|\boldsymbol{w}_i\|_{\ell_2}x_1)\left(\boldsymbol{I} - \boldsymbol{e}_1\boldsymbol{e}_1^T\right)\boldsymbol{x}]$$
$$= \|\boldsymbol{w}_i\|_{\ell_2}\gamma_\phi(\|\boldsymbol{w}_i\|_{\ell_2})\boldsymbol{e}_1\,,$$

where the last step holds because x_1 is independent of x_i, for $i \neq 1$. Replacing $\boldsymbol{e}_1$ with $\boldsymbol{w}_i/\|\boldsymbol{w}_i\|_{\ell_2}$ the latter calculation immediately implies that

$$\mathbb{E}[\phi'(\boldsymbol{w}_i^T\boldsymbol{x})\boldsymbol{x}] = \gamma_\phi(\|\boldsymbol{w}_i\|_{\ell_2})\boldsymbol{w}_i\,.$$

This also implies that

$$\mathbb{E}[(\phi'(\boldsymbol{w}_i^T\boldsymbol{x}) - \mathbb{E}[\phi'(\boldsymbol{w}_i^T\boldsymbol{x})])\boldsymbol{x}] = \gamma_\phi(\|\boldsymbol{w}_i\|_{\ell_2})\boldsymbol{w}_i\,.$$

Hence,

$$\mathbb{E}[\tilde{\boldsymbol{J}}_{\boldsymbol{x}}] = \mathbb{E}[\boldsymbol{x} \otimes \boldsymbol{D}_{\boldsymbol{v}}(\phi'(\boldsymbol{W}\boldsymbol{x}) - \mathbb{E}[\phi'(\boldsymbol{W}\boldsymbol{x})])] = \text{vec}(\boldsymbol{D}_{\boldsymbol{v}}\boldsymbol{\Gamma}\boldsymbol{W})\,, \tag{E.1}$$

with $\boldsymbol{\Gamma} = \text{diag}(\gamma_\phi(\|\boldsymbol{w}_1\|_{\ell_2}), \gamma_\phi(\|\boldsymbol{w}_2\|_{\ell_2}), \dots, \gamma_\phi(\|\boldsymbol{w}_k\|_{\ell_2}))$.

On the other hand, note that for any vector $\boldsymbol{a}$, $\mathbb{E}[\langle\boldsymbol{a},\boldsymbol{x}\rangle\boldsymbol{x}] = \boldsymbol{a}$.

Writing it in matrix form, for any matrix $\boldsymbol{A} \in \mathbb{R}^{k \times d}$ we have

$$\mathbb{E}[\boldsymbol{x} \otimes \boldsymbol{A}\boldsymbol{x}] = \text{vec}(\boldsymbol{A})\,. \tag{E.2}$$

By comparing (E.1) and (E.2), we get

$$\mathbb{E}[\tilde{\boldsymbol{J}}_{\boldsymbol{x}}] = \mathbb{E}[\boldsymbol{x} \otimes \boldsymbol{D}_{\boldsymbol{v}}\boldsymbol{\Gamma}\boldsymbol{W}\boldsymbol{x}]\,. \tag{E.3}$$

Recalling the definition $\boldsymbol{A} := \boldsymbol{D}_{\boldsymbol{v}}\boldsymbol{\Gamma}\boldsymbol{W}$ and using Lemma 5.1, we arrive at

$$\mathbb{E}[\tilde{\boldsymbol{J}}_{\boldsymbol{x}}] = \boldsymbol{D}_{\boldsymbol{A}}\,\mathbb{E}[\boldsymbol{x} \otimes \boldsymbol{x}] = \boldsymbol{D}_{\boldsymbol{A}}\boldsymbol{Q}\,,$$

concluding the proof.

F Proof of Proposition 7.9

Define

$$B_n = \sup_{z \in S^{n-1}} \left|\sum_{i \neq j}\langle z_i\boldsymbol{u_i}, z_j\boldsymbol{u_j}\rangle\right|^{1/2}$$

We write

$$\|\boldsymbol{U}\boldsymbol{z}\|^2 = \sum_{i=1}^n z_i^2\|\boldsymbol{u_i}\|^2 + \sum_{i \neq j}\langle z_i\boldsymbol{u_i}, z_j\boldsymbol{u_j}\rangle$$

Then, clearly for $\boldsymbol{z} \in \mathbb{S}^{n-1}$,

$$\|\boldsymbol{U}\boldsymbol{z}\|^2 \geq \eta_{\min}^2 - B_n^2\,, \quad \|\boldsymbol{U}\boldsymbol{z}\|^2 \leq \eta_{\max}^2 + B_n^2. \tag{F.1}$$

The following Lemma is similar to [2][Theorem 3.2].

Lemma F.1 *Let $u_1, \ldots, u_n$ be independent sub-exponential random vectors with $\psi = \max_{1 \leq i \leq n} \|u_i\|_{\psi_1}$. Let $\theta \in (0, 1/4)$, $K, K' \geq 1$ and assume that*

$$n \log^2(2/\theta) \leq \theta^2 \eta^2$$

Then setting $\xi = \psi K + K'$, the inequality

$$B_n^2 \leq C \xi^2 \theta \eta^2$$

holds with probability at least

$$1 - \exp\left(-cK\sqrt{n}\log\left(\frac{2}{\theta}\right)\right) - \mathbb{P}\left(\eta_{\max} \geq K'\eta\right)$$

We apply Lemma F.1 by setting $\eta = \eta_{\min}$ and letting θ be the solution of $n \log^2(2/\theta) = \theta^2 \eta^2$. Clearly, $\theta > \sqrt{n}/\eta_{\min}$. and therefore $\theta < (\sqrt{n}/\eta_{\min})\log(2\eta_{\min}/\sqrt{n})$. This gives

$$B_n^2 \leq C \xi^2 \eta_{\min} \sqrt{n} \log\left(\frac{2\eta_{\min}}{\sqrt{n}}\right). \tag{F.2}$$

with probability at least

$$1 - C \exp\left(-cK\sqrt{n}\log\left(\frac{2\eta_{\min}}{\sqrt{n}}\right)\right) - \mathbb{P}\left(\eta_{\max} \geq K'\eta_{\min}\right)$$

Using bound (F.2) in (F.1) we obtain the desired result.

Proof We first recall the following Lemma from Adamczak (this applies to vectors with non-trivial covariance).

Lemma F.2 *Let $u_1, \ldots, u_n$ be independent sub-exponential vectors with $\psi = \max_{1 \leq i \leq n} \|u_i\|_{\psi_1}$. For $\theta \in (0, 1/4)$ and $K \geq 1$, one has*

$$\mathbb{P}(B_n^2 \geq \max\{B^2, \eta_{\max}B, 24\theta\eta_{\max}^2\}) \leq (1 + 3\log n) \exp\left(-2K\sqrt{n}\log(2/\theta)\right)$$

and

$$B = C_0 \psi K \sqrt{n} \log(2/\theta)$$

Fix $K \geq 1$ and define

$$K_1 = \frac{K\theta\eta}{\sqrt{n}\log(2/\theta)} \geq K$$

By Lemma F.2, we have

$$\mathbb{P}(B_n^2 \geq \max\{B^2, \eta_{\max}B, 24\theta\eta_{\max}^2\}) \leq (1 + 3\log n) \exp\left(-2K_1\sqrt{n}\log(2/\theta)\right)$$

$$\leq \exp\left(-cK\sqrt{n}\log(2/\theta)\right)$$

where

$$B = C_0 \psi K_1 \sqrt{n} \log(2/\theta) = C_0 \psi K \theta \eta$$

Thus if $\eta_{\max} \leq K'\eta$ for some K', then

$$\max\{B^2, \eta_{\max}B, 24\theta\eta_{\max}^2\} \leq C_1 \theta\eta^2 \max\{\psi^2 K^2, \psi KK', K'^2\} \leq C_1 \theta\eta^2 (\psi K + K')^2,$$

where C_1 is an absolute constant. This completes the proof. ∎

G Proof of Lemma 7.10

Define the function $f(\boldsymbol{x}) = \|\boldsymbol{M}\boldsymbol{D_v}(\phi'(\boldsymbol{Wx}) - \mathbb{E}[\phi'(\boldsymbol{Wx})])\|_{\ell_2}$. We first compute the Lipschitz constant of f as follows:

$$
\begin{aligned}
|f(\boldsymbol{x}) - f(\tilde{\boldsymbol{x}})| &\le \left\|\boldsymbol{M}\boldsymbol{D_v}\left(\phi'(\boldsymbol{Wx}) - \phi'(\boldsymbol{W\tilde{x}})\right)\right\|_{\ell_2} \\
&\le \|\boldsymbol{M}\boldsymbol{D_v}\|\left\|\phi'(\boldsymbol{Wx}) - \phi'(\boldsymbol{W\tilde{x}})\right\|_{\ell_2} \\
&= \frac{1}{\sigma_{\min}(\boldsymbol{\Gamma W})}\left\|\phi'(\boldsymbol{Wx}) - \phi'(\boldsymbol{W\tilde{x}})\right\|_{\ell_2} \\
&\le \frac{L}{\sigma_{\min}(\boldsymbol{\Gamma W})}\left\|\boldsymbol{W}(\boldsymbol{x} - \tilde{\boldsymbol{x}})\right\|_{\ell_2} \\
&\le L\frac{\sigma_{\max}(\boldsymbol{W})}{\sigma_{\min}(\boldsymbol{\Gamma W})}\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|_{\ell_2} = \rho(\boldsymbol{W})\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|.
\end{aligned}
$$

Thus $\|\boldsymbol{z}\|_{\ell_2}$ is a Lipschitz function of $\boldsymbol{x}$ with Lipschitz constant $\rho(\boldsymbol{W})$. As a result,

$$
\mathbb{P}\left(\left|\|\boldsymbol{z}\|_{\ell_2} - \sqrt{\mathbb{E}[\|\boldsymbol{z}\|_{\ell_2}^2]}\right| \ge t\rho(\boldsymbol{W})\right) \le 2e^{-t^2/2}. \tag{G.1}
$$

We next bound $\mathbb{E}[\|\boldsymbol{z}\|_{\ell_2}^2]$. First note that

$$
\begin{aligned}
\mathbb{E}[\|\boldsymbol{z}\|_{\ell_2}^2] &= \mathbb{E}\left[\left\|\boldsymbol{M}\boldsymbol{D_v}\left(\phi'(\boldsymbol{Wx}) - \boldsymbol{\mu}\right)\right\|^2\right], \\
&= \mathbb{E}\left[\left\|\boldsymbol{M}\boldsymbol{D_v}\left(\phi'(\boldsymbol{Wx}) - \boldsymbol{\mu} - \boldsymbol{\Gamma Wx}\right) + \boldsymbol{x}\right\|^2\right], \\
&\ge \mathbb{E}[\|\boldsymbol{x}\|^2] = d, \tag{G.2}
\end{aligned}
$$

where in the last inequality, we used the observation that $\mathbb{E}[\langle\boldsymbol{x}, \boldsymbol{M}\boldsymbol{D_v}(\phi'(\boldsymbol{Wx}) - \boldsymbol{\mu} - \boldsymbol{\Gamma Wx})\rangle] = 0$ which holds based on Equation (E.3).

For upper bounding $\mathbb{E}[\|\boldsymbol{z}\|_{\ell_2}^2]$ note that

$$
\begin{aligned}
\mathbb{E}[\|\boldsymbol{z}\|_{\ell_2}^2] &= \mathbb{E}\left[\left\|\boldsymbol{M}\boldsymbol{D_v}(\phi'(\boldsymbol{Wx}) - \boldsymbol{\mu})\right\|_{\ell_2}^2\right] \\
&= \|\boldsymbol{M}\boldsymbol{D_v}\|^2\,\mathbb{E}\left[\left\|\phi'(\boldsymbol{Wx}) - \boldsymbol{\mu}\right\|_{\ell_2}^2\right] \\
&= \|\boldsymbol{M}\boldsymbol{D_v}\|^2\,\mathbb{E}\left[\left\|\phi'(\boldsymbol{Wx}) - \phi'(0)\boldsymbol{1}_k + \phi'(0)\boldsymbol{1}_k - \boldsymbol{\mu}\right\|_{\ell_2}^2\right] \\
&= \|\boldsymbol{M}\boldsymbol{D_v}\|^2\left\{\mathbb{E}\left[\left\|\phi'(\boldsymbol{Wx}) - \phi'(0)\boldsymbol{1}_k\right\|_{\ell_2}^2\right] - \left\|\boldsymbol{\mu} - \phi'(0)\boldsymbol{1}_k\right\|_{\ell_2}^2\right\} \\
&\le \|\boldsymbol{M}\boldsymbol{D_v}\|^2\,\mathbb{E}\left[\left\|\phi'(\boldsymbol{Wx}) - \phi'(0)\boldsymbol{1}_k\right\|_{\ell_2}^2\right] \\
&\le \frac{L^2\sigma_{\max}^2(\boldsymbol{W})}{\sigma_{\min}^2(\boldsymbol{\Gamma W})}\,\mathbb{E}[\|\boldsymbol{x}\|^2] = \rho^2(\boldsymbol{W})d. \tag{G.3}
\end{aligned}
$$

Here, $\boldsymbol{1}_k \in \mathbb{R}^k$ is the all-one vector. The result follows by putting together bounds (G.2) and (G.3) into (G.1).

H   Proof of Lemma 7.11

Note that

$$\left(J(\widetilde{\boldsymbol{v}}, \widetilde{\boldsymbol{W}}) - J(\boldsymbol{v}, \boldsymbol{W})\right)_i = \frac{1}{n}\boldsymbol{x}_i \otimes \left(\boldsymbol{D}_{\widetilde{\boldsymbol{v}}}\phi'(\widetilde{\boldsymbol{W}}\boldsymbol{x}_i) - \boldsymbol{D}_{\boldsymbol{v}}\phi'(\boldsymbol{W}\boldsymbol{x}_i)\right).$$

Further,

$$\begin{aligned}
&\widetilde{v}_\ell\phi'(\widetilde{\boldsymbol{w}}_\ell^T\boldsymbol{x}_i) - v_\ell\phi'(\boldsymbol{w}_\ell^T\boldsymbol{x}_i) \\
&= (\widetilde{v}_\ell - v_\ell)\phi'(\widetilde{\boldsymbol{w}}_\ell^T\boldsymbol{x}_i) + v_\ell\Big(\phi'(\widetilde{\boldsymbol{w}}_\ell^T\boldsymbol{x}_i) - \phi'(\boldsymbol{w}_\ell^T\boldsymbol{x}_i)\Big) \\
&= (\widetilde{v}_\ell - v_\ell)\phi'(\widetilde{\boldsymbol{w}}_\ell^T\boldsymbol{x}_i) + v_\ell\Big(\int_0^1 \phi''\big((t\widetilde{\boldsymbol{w}}_\ell + (1-t)\boldsymbol{w}_\ell)^T\boldsymbol{x}_i\big)\,\mathrm{d}t\Big)(\widetilde{\boldsymbol{w}}_\ell - \boldsymbol{w}_\ell)^T\boldsymbol{x}_i.
\end{aligned} \tag{H.1}$$

To simplify our exposition, define

$$\lambda_{i\ell} := \int_0^1 \phi''\big((t\widetilde{\boldsymbol{w}}_\ell + (1-t)\boldsymbol{w}_\ell)^T\boldsymbol{x}_i\big)\,\mathrm{d}t \qquad \eta_{i\ell} := \phi'(\widetilde{\boldsymbol{w}}_\ell^T\boldsymbol{x}_i).$$

Writing (H.1) in terms of $\lambda_{i\ell}$ and $\eta_{i\ell}$, we have

$$\widetilde{v}_\ell\phi'(\widetilde{\boldsymbol{w}}_\ell^T\boldsymbol{x}_i) - v_\ell\phi'(\boldsymbol{w}_\ell^T\boldsymbol{x}_i) = \eta_{i\ell}(\widetilde{v}_\ell - v_\ell) + \lambda_{i\ell}v_\ell(\widetilde{\boldsymbol{w}}_\ell - \boldsymbol{w}_\ell)^T\boldsymbol{x}_i.$$

Writing the above identity in matrix form, we get

$$\boldsymbol{D}_{\widetilde{\boldsymbol{v}}}\phi'(\widetilde{\boldsymbol{W}}\boldsymbol{x}_i) - \boldsymbol{D}_{\boldsymbol{v}}\phi'(\boldsymbol{W}\boldsymbol{x}_i) = \mathrm{diag}(\boldsymbol{\eta}_i)(\widetilde{\boldsymbol{v}} - \boldsymbol{v}) + \mathrm{diag}(\boldsymbol{\lambda}_i)\boldsymbol{D}_{\boldsymbol{v}}(\widetilde{\boldsymbol{W}} - \boldsymbol{W})\boldsymbol{x}_i. \tag{H.2}$$

By triangle inequality, we have

$$\left\|J(\widetilde{\boldsymbol{v}}, \widetilde{\boldsymbol{W}}) - J(\boldsymbol{v}, \boldsymbol{W})\right\| \le \left\|J(\boldsymbol{v}, \widetilde{\boldsymbol{W}}) - J(\boldsymbol{v}, \boldsymbol{W})\right\| + \left\|J(\widetilde{\boldsymbol{v}}, \widetilde{\boldsymbol{W}}) - J(\boldsymbol{v}, \widetilde{\boldsymbol{W}})\right\|. \tag{H.3}$$

We proceed by bounding the two terms above. For the first term note that by applying (H.2), we arrive at

$$\left\|J(\boldsymbol{v}, \widetilde{\boldsymbol{W}}) - J(\boldsymbol{v}, \boldsymbol{W})\right\| = \sup_{\boldsymbol{u}\in\mathbb{R}^n, \|\boldsymbol{u}\|_{\ell_2}=1} \left\|\frac{1}{n}\sum_{i=1}^n u_i\mathrm{diag}(\boldsymbol{\lambda}_i)\boldsymbol{D}_{\boldsymbol{v}}(\widetilde{\boldsymbol{W}} - \boldsymbol{W})\boldsymbol{x}_i\boldsymbol{x}_i^T\right\|_F.$$

Define a matrix $\boldsymbol{M} \in \mathbb{R}^{d^2 \times n}$ with columns given by

$$\boldsymbol{M}_{\boldsymbol{x}} = \boldsymbol{x} \otimes \boldsymbol{x}.$$

Invoking Corollary 7.5,

$$\|\boldsymbol{M}\| \le C\sqrt{nd}. \tag{H.4}$$

holds with probability at least $1 - ne^{-b_1\sqrt{n}} - n^{-1} - 2ne^{-b_2 d}$ for some constants $b_1, b_2 > 0$.

Now using the above we can also write

$$\left\| J(\boldsymbol{v}, \widehat{\boldsymbol{W}}) - J(\boldsymbol{v}, \boldsymbol{W}) \right\|$$

$$= \left\| \text{blockdiag}\left(\text{diag}(\boldsymbol{\lambda_1})\boldsymbol{D_v}(\widehat{\boldsymbol{W}} - \boldsymbol{W}), \text{diag}(\boldsymbol{\lambda_2})\boldsymbol{D_v}(\widehat{\boldsymbol{W}} - \boldsymbol{W}), \ldots, \text{diag}(\boldsymbol{\lambda_n})\boldsymbol{D_v}(\widehat{\boldsymbol{W}} - \boldsymbol{W}) \right)(\boldsymbol{x} \otimes \boldsymbol{x}) \right\|$$

$$\leq \left\| \text{blockdiag}\left(\text{diag}(\boldsymbol{\lambda_1})\boldsymbol{D_v}(\widehat{\boldsymbol{W}} - \boldsymbol{W}), \text{diag}(\boldsymbol{\lambda_2})\boldsymbol{D_v}(\widehat{\boldsymbol{W}} - \boldsymbol{W}), \ldots, \text{diag}(\boldsymbol{\lambda_n})\boldsymbol{D_v}(\widehat{\boldsymbol{W}} - \boldsymbol{W}) \right) \right\| \left\| \boldsymbol{x} \otimes \boldsymbol{x} \right\|$$

$$\leq \left\| \text{blockdiag}\left(\text{diag}(\boldsymbol{\lambda_1}), \text{diag}(\boldsymbol{\lambda_2}), \ldots, \text{diag}(\boldsymbol{\lambda_n}) \right) \right\| \left\| \boldsymbol{D_v}(\widehat{\boldsymbol{W}} - \boldsymbol{W}) \right\| \left\| \boldsymbol{x} \otimes \boldsymbol{x} \right\|$$

$$\leq L \|\boldsymbol{v}\|_{\ell_\infty} \left\| \widehat{\boldsymbol{W}} - \boldsymbol{W} \right\| \left\| \boldsymbol{x} \otimes \boldsymbol{x} \right\|$$

$$\leq C L \|\boldsymbol{v}\|_{\ell_\infty} \sqrt{nd} \left\| \widehat{\boldsymbol{W}} - \boldsymbol{W} \right\|, \tag{H.5}$$

where in the penultimate inequality, we use the fact that $|\lambda_{i\ell}| < L$. This concludes our bound on the first term of (H.3). To bound the second term in (H.3) note that,

$$\left\| J(\widetilde{\boldsymbol{v}}, \widehat{\boldsymbol{W}}) - J(\boldsymbol{v}, \widehat{\boldsymbol{W}}) \right\| = \sup_{\boldsymbol{u} \in \mathbb{R}^n, \|\boldsymbol{u}\|_{\ell_2}=1} \left\| \frac{1}{n} \sum_{i=1}^{n} u_i \text{diag}(\boldsymbol{\eta_i})(\widetilde{\boldsymbol{v}} - \boldsymbol{v})\boldsymbol{x}_i \boldsymbol{x}_i^T \right\|_F.$$

Using the fact that $|\eta_{i\ell}| < B$, with an analogous argument to the one we used for bounding the first we arrive at

$$\left\| J(\widetilde{\boldsymbol{v}}, \widehat{\boldsymbol{W}}) - J(\boldsymbol{v}, \widehat{\boldsymbol{W}}) \right\| \leq C B \sqrt{nd} \left\| \widetilde{\boldsymbol{v}} - \boldsymbol{v} \right\|_{\ell_\infty}. \tag{H.6}$$

Combining inequalities (H.5) and (H.6), we have

$$\left\| J(\widetilde{\boldsymbol{v}}, \widehat{\boldsymbol{W}}) - J(\boldsymbol{v}, \boldsymbol{W}) \right\| \leq C \sqrt{nd} \left(\|\boldsymbol{v}\|_{\ell_\infty} \left\| \widehat{\boldsymbol{W}} - \boldsymbol{W} \right\| + \left\| \widetilde{\boldsymbol{v}} - \boldsymbol{v} \right\|_{\ell_\infty} \right),$$

where C depends on constants L and B.

I   Proof of Lemma 7.12

Recall that

$$\nabla_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{v}, \boldsymbol{W}) = \frac{1}{n} \boldsymbol{J}(\boldsymbol{v}, \boldsymbol{W})\boldsymbol{r}, \quad \mathcal{L}(\boldsymbol{v}, \boldsymbol{W}) = \frac{1}{2n} \|\boldsymbol{r}\|_{\ell_2}^2. \tag{I.1}$$

Given that $(\boldsymbol{v}, \boldsymbol{W}) \in \Omega$, we have

$$\sigma_{\min}(\boldsymbol{J}(\boldsymbol{v}, \boldsymbol{W})) \geq \sigma_{\min}(\boldsymbol{J}(\boldsymbol{v}^*, \boldsymbol{W}^*)) - \|\boldsymbol{J}(\boldsymbol{v}, \boldsymbol{W}) - \boldsymbol{J}(\boldsymbol{v}^*, \boldsymbol{W}^*)\|$$

$$\geq \sigma_{\min}(\boldsymbol{J}(\boldsymbol{v}^*, \boldsymbol{W}^*)) - C'\sqrt{nd}\left(\|\boldsymbol{v}^*\|_{\ell_\infty} \|\boldsymbol{W} - \boldsymbol{W}^*\| + \|\boldsymbol{v} - \boldsymbol{v}^*\|_{\ell_\infty} \right)$$

$$\geq \frac{c}{2} \sigma_{\min}(\boldsymbol{W}^*)d,$$

where the first inequality follows from Lemma 7.11 and the second one follows readily from definition of set Ω and Equation (7.42).

Likewise,

$$\sigma_{\max}(\boldsymbol{J}(\boldsymbol{v}, \boldsymbol{W})) \leq \sigma_{\max}(\boldsymbol{J}(\boldsymbol{v}^*, \boldsymbol{W}^*)) + \|\boldsymbol{J}(\boldsymbol{v}, \boldsymbol{W}) - \boldsymbol{J}(\boldsymbol{v}^*, \boldsymbol{W}^*)\|$$

$$\leq \sigma_{\max}(\boldsymbol{J}(\boldsymbol{v}^*, \boldsymbol{W}^*)) + C'\sqrt{nd}\left(\|\boldsymbol{v}^*\|_{\ell_\infty} \|\boldsymbol{W} - \boldsymbol{W}^*\| + \|\boldsymbol{v} - \boldsymbol{v}^*\|_{\ell_\infty} \right)$$

$$\leq C\sigma_{\max}(\boldsymbol{W}^*)\sqrt{nk} + \frac{c}{2}\sigma_{\min}(\boldsymbol{W}^*)d,$$

$$\leq \frac{3}{2}C\sigma_{\max}(\boldsymbol{W}^*)\sqrt{nk}. \tag{I.2}$$

Therefore,

$$
\begin{aligned}
\|\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{v},\boldsymbol{W})\|_F^2 &= \frac{1}{n^2}\|\boldsymbol{J}(\boldsymbol{v},\boldsymbol{W})\boldsymbol{r}\|_{\ell_2}^2 \\
&\geq \frac{2}{n}\sigma_{\min}^2(\boldsymbol{J}(\boldsymbol{v},\boldsymbol{W}))\mathcal{L}(\boldsymbol{v},\boldsymbol{W}) \\
&\geq \frac{c^2}{2}\sigma_{\min}^2(\boldsymbol{W}^*)\frac{d^2}{n}\,\mathcal{L}(\boldsymbol{v},\boldsymbol{W})\,,
\end{aligned}
$$

proving Equation (7.44).

Similarly, we have

$$
\begin{aligned}
\|\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{v},\boldsymbol{W})\|_F^2 &= \frac{1}{n^2}\|\boldsymbol{J}(\boldsymbol{v},\boldsymbol{W})\boldsymbol{r}\|_{\ell_2}^2 \\
&\leq \frac{2}{n}\sigma_{\max}^2(\boldsymbol{J}(\boldsymbol{v},\boldsymbol{W}))\mathcal{L}(\boldsymbol{v},\boldsymbol{W}) \\
&\leq \frac{9}{2}C^2\sigma_{\max}^2(\boldsymbol{W}^*)k\,\mathcal{L}(\boldsymbol{v},\boldsymbol{W})\,,
\end{aligned}
$$

proving Equation (7.45).

To prove the last bound (i.e. (7.46)), we recall Equation (5.6)

$$
\nabla_{\boldsymbol{v}}\mathcal{L}(\boldsymbol{v},\boldsymbol{W}) = \frac{1}{n}\phi(\boldsymbol{W}\boldsymbol{X})\boldsymbol{r}\,.
$$

We write

$$
\|\nabla_{\boldsymbol{v}}\mathcal{L}(\boldsymbol{v},\boldsymbol{W})\|_{\ell_\infty} \leq \left\|\sqrt{\frac{2}{n}}\phi(\boldsymbol{W}\boldsymbol{X})\frac{1}{\sqrt{2n}}\boldsymbol{r}\right\|_{\ell_\infty} \tag{I.3}
$$

$$
\leq \sqrt{\frac{2}{n}}\|\phi(\boldsymbol{W}\boldsymbol{X})\|_{2,\infty}\sqrt{\mathcal{L}(\boldsymbol{v},\boldsymbol{W})}\,, \tag{I.4}
$$

where for a matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|_{2,\infty}$ denotes the maximum ℓ_2 norm of its rows. Hence,

$$
\begin{aligned}
\|\phi(\boldsymbol{W}\boldsymbol{X})\|_{2,\infty} &= \max_{\ell\in[k]}\Big(\sum_{i=1}^n\phi(\boldsymbol{w}_\ell^T\boldsymbol{x}_i)^2\Big)^{1/2} \\
&\leq \max_{\ell\in[k]}\Big(\sum_{i=1}^n\big(|\phi(0)|+B|\boldsymbol{w}_\ell^T\boldsymbol{x}_i|\big)^2\Big)^{1/2} \\
&\leq \max_{\ell\in[k]}\Big(\sum_{i=1}^n\big(2\phi^2(0)+2B^2|\boldsymbol{w}_\ell^T\boldsymbol{x}_i|^2\big)\Big)^{1/2} \\
&= \max_{\ell\in[k]}\Big(2\phi^2(0)n+2B^2\big\|\boldsymbol{w}_\ell^T\boldsymbol{X}\big\|_{\ell_2}^2\Big)^{1/2} \\
&\leq \max_{\ell\in[k]}\Big(2\phi^2(0)n+2B^2\|\boldsymbol{w}_\ell\|_{\ell_2}^2\|\boldsymbol{X}\|^2\Big)^{1/2} \tag{I.5}
\end{aligned}
$$

Here, we used the assumption that $\phi(z)$ is B-Lipschitz.

Next, by the Bai-Yin law [5], we have $\|\boldsymbol{X}\| \le 4\sqrt{n}$, with probability at least $1-2e^{-2n}$. Continuing with Equation (I.5), we have

$$
\begin{aligned}
\left\|\phi'(\boldsymbol{W}\boldsymbol{X})\right\|_{2,\infty} &\le \max_{\ell\in[k]}\left(2\phi^2(0)n + 32B^2 n \|\boldsymbol{w}_\ell\|_{\ell_2}^2\right)^{1/2} \\
&\le \max_{\ell\in[k]}\left(2\phi^2(0)n + 32B^2 n \left(\|\boldsymbol{w}_\ell^*\|_{\ell_2} + 1\right)^2\right)^{1/2} \\
&\le \max_{\ell\in[k]}\left(2\phi^2(0)n + 32B^2 n \left(2\|\boldsymbol{w}_\ell^*\|_{\ell_2}^2 + 2\right)\right)^{1/2} \\
&\le \max_{\ell\in[k]}\left((2\phi^2(0) + 64B^2)n + 64B^2 n \|\boldsymbol{w}_\ell^*\|_{\ell_2}^2\right)^{1/2} \\
&\le \left(2\phi^2(0) + 64B^2 + 64B^2 w_{\max}^2\right)^{1/2}\sqrt{n}\,,
\end{aligned}
\tag{I.6}
$$

where in the second inequality, we use the fact that $(\boldsymbol{v},\boldsymbol{W}) \in \Omega$ and hence $\|\boldsymbol{W} - \boldsymbol{W}^*\|_F \le 1$ (since $R < v_{\max}$). Using bound (I.6) in (I.3), we obtain

$$
\|\nabla_{\boldsymbol{v}}\mathcal{L}(\boldsymbol{v},\boldsymbol{W})\|_{\ell_\infty}^2 \le \left(4\phi^2(0) + 128B^2 + 128B^2 w_{\max}^2\right)\mathcal{L}(\boldsymbol{v},\boldsymbol{W})\,.
\tag{I.7}
$$

concluding the proof of Equation (7.46).

J   Proof of Lemma 7.13

We have

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{v},\boldsymbol{W}) &= \frac{1}{2n}\sum_{i=1}^n \left(\boldsymbol{v}^T\phi(\boldsymbol{W}\boldsymbol{x}_i) - \boldsymbol{v}^{*T}\phi(\boldsymbol{W}^*\boldsymbol{x}_i)\right)^2, \\
&\le \underbrace{\frac{1}{n}\|\boldsymbol{v}\|_{\ell_2}^2\sum_{i=1}^n\|\phi(\boldsymbol{W}\boldsymbol{x}_i) - \phi(\boldsymbol{W}^*\boldsymbol{x}_i)\|_{\ell_2}^2}_{T_1} + \underbrace{\frac{1}{n}\sum_{i=1}^n\left((\boldsymbol{v}-\boldsymbol{v}^*)^T\phi(\boldsymbol{W}^*\boldsymbol{x}_i)\right)^2}_{T_2}.
\end{aligned}
\tag{J.1}
$$

By standard concentration of sample covariance matrices we conclude that for any $\delta > 0$, there exist constants $c, b_0 > 0$, such that for $n \ge \frac{c^2}{\delta^2}d$, we have

$$
\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T - \boldsymbol{I}\right\| \le \delta\,,
\tag{J.2}
$$

with probability at least $1-2e^{-b_0 d}$. Using this bound, with $\delta = 1$, along with the 1-Lipschitz property of the activation $\phi(z)$, we arrive at

$$
\begin{aligned}
T_1 &\le \frac{\|\boldsymbol{v}\|^2}{2n}\sum_{i=1}^n\|\phi(\boldsymbol{W}\boldsymbol{x}_i) - \phi(\boldsymbol{W}^*\boldsymbol{x}_i)\|_{\ell_2}^2, \\
&\le \frac{B\|\boldsymbol{v}\|^2}{2n}\sum_{i=1}^n\|(\boldsymbol{W} - \boldsymbol{W}^*)\boldsymbol{x}_i\|_{\ell_2}^2, \\
&= \frac{B\|\boldsymbol{v}\|^2}{2}\langle(\boldsymbol{W} - \boldsymbol{W}^*)^T(\boldsymbol{W} - \boldsymbol{W}^*), \frac{1}{n}\sum_{i=1}^n\boldsymbol{x}_i\boldsymbol{x}_i^T\rangle, \\
&\le B\|\boldsymbol{v}\|^2\|\boldsymbol{W} - \boldsymbol{W}^*\|_F^2\,.
\end{aligned}
\tag{J.3}
$$

57

To bound the second term note that we have

$$T_2 \le \frac{1}{n} \sum_{i=1}^n \left((\boldsymbol{v} - \boldsymbol{v}^*)^T \phi(\boldsymbol{W}^* \boldsymbol{x_i}) \right)^2,$$

$$\le \frac{1}{n} \|\boldsymbol{v} - \boldsymbol{v}^*\|_{\ell_\infty}^2 \sum_{i=1}^n \|\phi(\boldsymbol{W}^* \boldsymbol{x}_i)\|_{\ell_1}^2,$$

$$\le \frac{1}{n} \|\boldsymbol{v} - \boldsymbol{v}^*\|_{\ell_\infty}^2 \sum_{i=1}^n \left(\sum_{\ell=1}^k \cdot \phi(\boldsymbol{w}_\ell^{*T} \boldsymbol{x}_i) \right)^2$$

Note that for any $z \in \mathbb{R}$, we have $|\phi(z)| < |\phi(0)| + B|z|$. Continuing with the above inequality, we write

$$T_2 \le \frac{1}{n} \|\boldsymbol{v} - \boldsymbol{v}^*\|_{\ell_\infty}^2 \sum_{i=1}^n \left(\sum_{\ell=1}^k \left(|\phi(0)| + B|\boldsymbol{w}_\ell^{*T} \boldsymbol{x}_i| \right) \right)^2,$$

$$\le \frac{k}{n} \|\boldsymbol{v} - \boldsymbol{v}^*\|_{\ell_\infty}^2 \sum_{i=1}^n \sum_{\ell=1}^k \left(|\phi(0)| + B|\boldsymbol{w}_\ell^{*T} \boldsymbol{x}_i| \right)^2,$$

$$\le \frac{k}{n} \|\boldsymbol{v} - \boldsymbol{v}^*\|_{\ell_\infty}^2 \sum_{i=1}^n \left(2|\phi(0)|k + 2B \sum_{\ell=1}^k (\boldsymbol{w}_\ell^{*T} \boldsymbol{x}_i)^2 \right),$$

$$= \frac{2k}{n} \|\boldsymbol{v} - \boldsymbol{v}^*\|_{\ell_\infty}^2 \left(\phi^2(0)kn + B^2 \|\boldsymbol{W}^* \boldsymbol{X}\|_F^2 \right),$$

$$\le \frac{2k}{n} \|\boldsymbol{v} - \boldsymbol{v}^*\|_{\ell_\infty}^2 \left(\phi^2(0)kn + B^2 \|\boldsymbol{W}^*\|_F^2 \|\boldsymbol{X}\|^2 \right),$$

$$\le 2k^2 \|\boldsymbol{v} - \boldsymbol{v}^*\|_{\ell_\infty}^2 \left(\phi^2(0) + 2B^2 w_{\max}^2 \right),$$

where in the last step, we used (J.2) and the fact that $\|\boldsymbol{W}^*\|_F^2 \le k w_{\max}^2$. Combining the bounds on T_1 and T_2, completes the proof.

K   Proof of Lemma 7.14

Note that the spectral norm of a matrix is bounded above by the sum of the spectral norms of the diagonal blocks of that matrix. Hence,

$$\left\| \nabla^2 \mathcal{L}(\boldsymbol{v}, \boldsymbol{W}) \right\| \le \left\| \nabla_{\boldsymbol{W}}^2 \mathcal{L}(\boldsymbol{v}, \boldsymbol{W}) \right\| + \left\| \nabla_{\boldsymbol{v}}^2 \mathcal{L}(\boldsymbol{v}, \boldsymbol{W}) \right\|. \tag{K.1}$$

We proceed by bounding each of these two terms. Using Identity (5.8), we have

$$\nabla_{\boldsymbol{W}}^2 \mathcal{L}(\boldsymbol{v}, \boldsymbol{W}) =$$
$$\frac{1}{n} \boldsymbol{J} \boldsymbol{J}^T + \frac{1}{n} \text{blockdiag} \left(\sum_{i=1}^n v_1 \phi''(\boldsymbol{w}_1^T \boldsymbol{x}_i) r_i \boldsymbol{x}_i \boldsymbol{x}_i^T, \sum_{i=1}^n v_2 \phi''(\boldsymbol{w}_2^T \boldsymbol{x}_i) r_i \boldsymbol{x}_i \boldsymbol{x}_i^T, \dots, \sum_{i=1}^n v_k \phi''(\boldsymbol{w}_k^T \boldsymbol{x}_i) r_i \boldsymbol{x}_i \boldsymbol{x}_i^T \right).$$

Since $(\boldsymbol{v}, \boldsymbol{W}) \in \Omega$, we have

$$\|\boldsymbol{v}\|_{\ell_\infty} \le \|\boldsymbol{v}^*\|_{\ell_\infty} + \|\boldsymbol{v} - \boldsymbol{v}^*\|_{\ell_\infty} \le R + v_{\max} < 2v_{\max}.$$

Furthermore using (J.2) together with $|\phi''| < L$ we conclude that

$$\left\| \nabla_{\boldsymbol{W}}^2 \mathcal{L}(\boldsymbol{v}, \boldsymbol{W}) - \frac{1}{n} \boldsymbol{J} \boldsymbol{J}^T \right\| \leq 2L v_{\max} \left\| \frac{1}{n} \sum_{i=1}^n r_i \boldsymbol{x}_i \boldsymbol{x}_i^T \right\|,$$

$$\leq 2L v_{\max} \| \boldsymbol{r} \|_{\ell_\infty} \left\| \frac{1}{n} \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^T \right\|,$$

$$\leq 4L v_{\max} \| \boldsymbol{r} \|_{\ell_\infty}, \tag{K.2}$$

holds with probability at least $1 - 2e^{-b_0 d}$. Moreover, using the B-Lipschitz property of the activation, we have

$$|r_i| = |\boldsymbol{v}^T \phi(\boldsymbol{W} \boldsymbol{x}_i) - \boldsymbol{v}^T \phi(\boldsymbol{W}^* \boldsymbol{x}_i)|$$

$$\leq \| \boldsymbol{v} \|_{\ell_2} \left\| \phi(\boldsymbol{W} \boldsymbol{x}_i) - \boldsymbol{v}^T \phi(\boldsymbol{W}^* \boldsymbol{x}_i) \right\|_{\ell_2}$$

$$\leq B v_{\max} \sqrt{k} \| (\boldsymbol{W} - \boldsymbol{W}^*) \boldsymbol{x}_i \|_{\ell_2}$$

$$\leq 2B v_{\max} \sqrt{kd} \| \boldsymbol{W} - \boldsymbol{W}^* \|, \tag{K.3}$$

where in the last step we used the fact that $\| \boldsymbol{x}_i \|_{\ell_2}^2 \leq d$, which holds with probability at least $1 - e^{-d/8}$ by standard concentration bound for χ^2- random variables. Using (K.3) in (K.2), we obtain

$$\left\| \nabla_{\boldsymbol{W}}^2 \mathcal{L}(\boldsymbol{v}, \boldsymbol{W}) - \frac{1}{n} \boldsymbol{J} \boldsymbol{J}^T \right\| \leq 8BL v_{\max}^2 \sqrt{kd} \| \boldsymbol{W} - \boldsymbol{W}^* \|,$$

$$\overset{(a)}{\leq} 8BL v_{\max} BL \sqrt{kd} R,$$

$$\overset{(b)}{\leq} 8BL v_{\max}^2 BL k, \tag{K.4}$$

where (a) follows from the definition of the set Ω and (b) follows from $R \leq v_{\max}$ and $k \geq d$. Note that while we can plug in for R to obtain a tighter bound, we use a looser bound on R as this term will be dominated by the other term (i.e. $\| \boldsymbol{J} \boldsymbol{J}^T / n \|$) and thus a more accurate bound is unnecessary.

Restating the bound on $\sigma_{\max}(\boldsymbol{J}(\boldsymbol{v}, \boldsymbol{W}))$ for $(\boldsymbol{v}, \boldsymbol{W}) \in \Omega$, given by (I.2), we have

$$\left\| \frac{1}{n} \boldsymbol{J} \boldsymbol{J}^T \right\| \leq \frac{9}{4} C^2 \sigma_{\max}^2(\boldsymbol{W}^*) k. \tag{K.5}$$

Combining (K.4) and (K.5), we arrive at

$$\left\| \nabla_{\boldsymbol{W}}^2 \mathcal{L}(\boldsymbol{v}, \boldsymbol{W}) \right\| \leq \left(3C^2 \sigma_{\max}^2(\boldsymbol{W}^*) + 8 v_{\max}^2 BL \right) k. \tag{K.6}$$

We next bound $\left\|\nabla_{\boldsymbol{v}}^2 \mathcal{L}(\boldsymbol{v}, \boldsymbol{W})\right\|$. Starting with Equation (5.9), we have

$$
\begin{aligned}
\left\|\nabla_{\boldsymbol{v}}^2 \mathcal{L}(\boldsymbol{v}, \boldsymbol{W})\right\| &= \frac{1}{n}\left\|\phi(\boldsymbol{W}\boldsymbol{X})\phi(\boldsymbol{W}\boldsymbol{X})^T\right\| \\
&\leq \frac{1}{n}\left\|\sum_{i=1}^n \phi(\boldsymbol{W}\boldsymbol{x}_i)\phi(\boldsymbol{W}\boldsymbol{x}_i)^T\right\| \\
&\leq \frac{1}{n}\sum_{i=1}^n \left\|\phi(\boldsymbol{W}\boldsymbol{x}_i)\phi(\boldsymbol{W}\boldsymbol{x}_i)^T\right\| \\
&= \frac{1}{n}\sum_{i=1}^n \left\|\phi(\boldsymbol{W}\boldsymbol{x}_i)\right\|_{\ell_2}^2 \\
&\overset{(a)}{\leq} \frac{1}{n}\sum_{i=1}^n \sum_{\ell=1}^k \left(|\phi(0)| + B|\boldsymbol{w}_\ell^{*T}\boldsymbol{x}_i|\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^n \left(2\phi(0)^2 k + 2B^2 \sum_{\ell=1}^k (\boldsymbol{w}_\ell^{*T}\boldsymbol{x}_i)^2\right) \\
&= \frac{2}{n}\left(\phi^2(0)kn + B^2\left\|\boldsymbol{W}^*\boldsymbol{X}\right\|_F^2\right) \\
&\leq \frac{2}{n}\left(\phi^2(0)kn + B^2\left\|\boldsymbol{W}^*\right\|_F^2\left\|\boldsymbol{X}\right\|^2\right) \\
&\overset{(b)}{\leq} 2k\left(\phi^2(0) + 2B^2 w_{\max}^2\right).
\end{aligned} \tag{K.7}
$$

Here, (a) holds because ϕ is B-Lipschitz; (b) holds because $\left\|\boldsymbol{W}\right\|_F^2 \leq k w_{\max}^2$ and $\left\|\boldsymbol{X}\right\| \leq \sqrt{2n}$, with probability at least $1 - 2e^{-2n}$.

Using bounds (K.6) and (K.7), we obtain

$$
\left\|\nabla_{\boldsymbol{W}}^2 \mathcal{L}(\boldsymbol{v}, \boldsymbol{W})\right\| + \left\|\nabla_{\boldsymbol{v}}^2 \mathcal{L}(\boldsymbol{v}, \boldsymbol{W})\right\| \leq \left(3C^2\sigma_{\max}^2(\boldsymbol{W}^*) + 8v_{\max}^2 BL + 4B^2 w_{\max}^2 + 2\phi^2(0)\right)k. \tag{K.8}
$$