RESTRICTED STRONG CONVEXITY OF DEEP LEARNING MODELS WITH SMOOTH ACTIVATIONS

Arindam Banerjee  
Department of Computer Science  
University of Illinois at Urbana-Champaign  
arindamb@illinois.edu

Pedro Cisneros-Velarde  
Department of Computer Science  
University of Illinois at Urbana-Champaign  
pacisne@gmail.com

Libin Zhu  
Department of Computer Science  
University of California, San Diego  
l5zhu@ucsd.edu

Mikhail Belkin  
Haliciouglu Data Science Institute  
University of California, San Diego  
mbelkin@ucsd.edu

ABSTRACT

We consider the problem of optimization of deep learning models with smooth activation functions. While there exist influential results on the problem from the “near initialization” perspective, we shed considerable new light on the problem. In particular, we make two key technical contributions for such models with $L$ layers, $m$ width, and $\sigma_0^2$ initialization variance. First, for suitable $\sigma_0^2$, we establish a $O\left(\frac{\text{poly}(L)}{\sqrt{m}}\right)$ upper bound on the spectral norm of the Hessian of such models, considerably sharpening prior results. Second, we introduce a new analysis of optimization based on Restricted Strong Convexity (RSC) which holds as long as the squared norm of the average gradient of predictors is $\Omega\left(\frac{\text{poly}(L)}{\sqrt{m}}\right)$ for the square loss. We also present results for more general losses. The RSC based analysis does not need the “near initialization” perspective and guarantees geometric convergence for gradient descent (GD). To the best of our knowledge, ours is the first result on establishing geometric convergence of GD based on RSC for deep learning models, thus becoming an alternative sufficient condition for convergence that does not depend on the widely-used Neural Tangent Kernel (NTK). We share preliminary experimental results supporting our theoretical advances.

1 INTRODUCTION

Recent years have seen advances in understanding convergence of gradient descent (GD) and variants for deep learning models (Du et al., 2019; Allen-Zhu et al., 2019; Zou & Gu, 2019; Zou et al., 2020; Liu et al., 2022; Ji & Telgarsky, 2019; Oymak & Soltanolkotabi, 2020; Nguyen, 2021). Despite the fact that such optimization problems are non-convex, a series of recent results have shown that GD has geometric convergence and finds near global solution “near initialization” for wide networks. Such analysis is typically done based on the Neural Tangent Kernel (NTK) (Jacot et al., 2018), in particular by showing that the NTK is positive definite “near initialization,” in turn implying the optimization problem satisfies a condition closely related to the Polyak-Łojasiewicz (PL) condition, which in turn implies geometric convergence to the global minima (Liu et al., 2022; Nguyen, 2021). Such results have been generalized to more flexible forms of “lazy learning” where similar guarantees hold (Chizat et al., 2019). However, there are concerns regarding whether such “near initialization” or “lazy learning” truly explains the optimization behavior in realistic deep learning models (Geiger et al., 2020; Yang & Hu, 2020; Fort et al., 2020; Chizat et al., 2019).

Our work focuses on optimization of deep models with smooth activation functions, which have become increasingly popular in recent years (Du et al., 2019; Liu et al., 2022; Huang & Yau, 2020). Much of the theoretical convergence analysis of GD has focused on ReLU networks (Allen-Zhu et al., 2019; Nguyen, 2021). Some progress has also been made for deep models with smooth activations, but existing results are based on a variant of the NTK analysis, and the requirements on the width of
such models are high (Du et al., 2019; Liu et al., 2022). Based on such background and context, the
motivating question behind our work is: Are there other (meaningful) sufficient conditions beyond
NTK which lead to (geometric) convergence of GD for deep learning optimization?

Based on such motivation, we make two technical contributions in this paper which shed light on
optimization of deep learning models with smooth activations and with $L$ layers, $m$ width, and $\sigma_0^2$
initialization variance. First, for suitable $\sigma_0^2$, we establish a $O\left(\frac{\sigma_0^2}{\sqrt{m}}\right)$ upper bound on the spectral
norm of the Hessian of such models (Section 4). The bound holds over a large layerwise spectral norm
(instead of Frobenius norm) ball $B_{\rho,\rho_1}^{\text{Spec}}(\theta_0)$ around the random initialization $\theta_0$, where the
radius $\rho < \sqrt{m}$, arguably much bigger than what real world deep models need. Our analysis builds
on and sharpens recent prior work on the topic (Liu et al., 2020). While our analysis holds for
Gaussian random initialization of weights with any variance $\sigma_0^2$, the poly$(L)$ dependence happens
when $\sigma_0^2 \leq \frac{1}{4+o(1)} \frac{1}{m}$ (we handle the $\frac{1}{m}$ scaling explicitly).

Second, based on our Hessian spectral norm bound, we introduce a new approach to the analysis
of optimization of deep models with smooth activations based on the concept of Restricted Strong Conv-
exity (RSC) (Section 5) (Wainwright, 2019; Negahban et al., 2012; Negahban & Wainwright, 2012;
Banerjee et al., 2014; Chen & Banerjee, 2015). While RSC has been a core theme in high-dimensional
statistics especially for linear models and convex losses (Wainwright, 2019), to the best of our knowl-
edge, RSC has not been considered in the context of non-convex optimization of overparameterized
deep models. For a normalized total loss function $L(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \hat{y}_i), \hat{y}_i = f(\theta; x_i)$ with
predictor or neural network model $f$ parameterized by vector $\theta$ and data points $\{(x_i, y_i)\}_{i=1}^n$, when $\ell$

corresponds to the square loss we show that the total loss function satisfies RSC on a suitable restricted
set $Q^* \subset \mathbb{R}^p$ (Definition 5.2 in Section 5) at step $t$ as long as $\|\frac{1}{n} \sum_{i=1}^n \nabla_\theta f(\theta; x_i)\|_2^2 = \Omega\left(\frac{1}{\sqrt{m}}\right)$.

We also present similar results for general losses for which additional assumptions are needed. We
show that the RSC property implies a Restricted Polyak-Łojasiewicz (RPL) condition on $Q^*_t$, in turn
implying a geometric one-step decrease of the loss towards the minimum in $Q^*_t$, and subsequently
implying geometric decrease of the loss towards the minimum in the large (layerwise spectral norm)
ball $B_{\rho,\rho_1}^{\text{Spec}}(\theta_0)$. The geometric convergence due to RSC is a novel approach in the context of deep
learning optimization which does not depend on properties of the NTK. Thus, the RSC condition
provides an alternative sufficient condition for geometric convergence for deep learning optimization
to the widely-used NTK condition.

The rest of the paper is organized as follows. We briefly present related work in Section 2 and discuss
the problem setup in Section 3. We establish the Hessian spectral norm bound in Section 4 and introduce
the RSC based optimization analysis in Section 5. We experimental results corresponding
to the RSC condition in Section 6 and conclude in Section 7. All technical proofs are in the Appendix.

2 RELATED WORK

The literature on gradient descent and variants for deep learning is increasingly large, and we refer
the readers to the following surveys for an overview of the field (Fan et al., 2021; Bartlett et al., 2021).
Among the theoretical works, we consider (Du et al., 2019; Allen-Zhu et al., 2019; Zou & Gu, 2019;
Zou et al., 2020; Liu et al., 2022) as the closest to our work in terms of their study of convergence on
multi-layer neural networks. For a literature review on shallow and/or linear networks, we refer to
the recent survey (Fan et al., 2021). Due to the rapidly growing related work, we only refer to the most
related or recent work for most parts.

(Du et al., 2019; Zou & Gu, 2019; Allen-Zhu et al., 2019; Liu et al., 2022) considered optimization
of square loss, which we also consider for our main results, and we also present extensions to more
general class of loss functions. (Zou & Gu, 2019; Zou et al., 2020; Allen-Zhu et al., 2019; Nguyen
& Mondelli, 2020; Nguyen, 2021; Nguyen et al., 2021) analyzed deep ReLU networks. Instead, we
consider smooth activation functions, similar to (Du et al., 2019; Liu et al., 2022). The convergence
analysis of the gradient descent in (Du et al., 2019; Allen-Zhu et al., 2019; Zou & Gu, 2019; Zou
et al., 2020; Liu et al., 2022) relied on the near constancy of NTK for wide neural networks (Jacot
et al., 2018; Lee et al., 2019; Arora et al., 2019; Liu et al., 2020), which yield certain desirable
properties for their training using gradient descent based methods. One such property is related to the
PL condition (Karimi et al., 2016; Nguyen, 2021), formulated as PL* condition in (Liu et al., 2022).
Our work uses a different optimization analysis based on RSC (Wainwright, 2019; Negahban et al.)
We make the following assumption regarding the activation function. We define two types of balls over parameters that will be used throughout our analysis.

**Assumption 1**

\[ |W| \leq \rho \]

where \( W \) is the smooth (pointwise) activation function, and the total set of parameters \( \theta := (\text{vec}(W^{(1)})^\top, \ldots, \text{vec}(W^{(L)})^\top, v^\top) \in \mathbb{R}^{\sum_{l=1}^{L} m_l m_{l-1} + m_L} \),

with \( m_0 = d \). For simplicity, we will assume that the width of all the layers is the same, i.e., \( m_l = m \), \( l \in [L] \), and so that \( \theta \in \mathbb{R}^{Lm^2 + m} \). For simplicity, we also consider deep models with only one output, i.e., \( f(\theta; \mathbf{x}) \in \mathbb{R} \) as in [Du et al., 2019], but our results can be extended to multi-dimension outputs as in [Zou & Gu, 2019], using \( V \in \mathbb{R}^{mL \times k} \) for \( k \) outputs at the last layer; see Appendix C.

We make the following assumption regarding the activation function \( \phi \):

**Assumption 1 (Activation function).** The activation \( \phi \) is \( l \)-Lipschitz, i.e., \( |\phi'| \leq 1 \), and \( \beta_{\phi} \)-smooth, i.e., \( |\phi'| \leq \beta_{\phi} \).

**Remark 3.1.** Our analysis holds for any \( \zeta_{\phi} \)-Lipschitz smooth activations, with a dependence on \( \zeta_{\phi} \) on most key results. The main (qualitative) conclusions stay true if \( \zeta_{\phi} \leq 1 + o(1) \) or \( \zeta_{\phi} = \text{poly}(L) \), which is typically satisfied for commonly used smooth activations and moderate values of \( L \).

We define two types of balls over parameters that will be used throughout our analysis.

**Definition 3.1 (Norm balls).** Given \( \mathbf{v} \in \mathbb{R}^p \) of the form (\( 3 \)) with parameters \( \theta \), \( \phi \), and \( \psi \), we define

\[ B_{\rho, \phi1}^{\text{Spec}}(\bar{\theta}) := \left\{ \theta \in \mathbb{R}^p \text{ as in (3)} \mid ||W^{(l)} - \bar{W}^{(l)}||_2 \leq \rho, l \in [L], ||v - \bar{v)||_2 \leq \rho_1 \right\} , \]

\[ B_{\rho}^{\text{Spec}}(\bar{\theta}) := \left\{ \theta \in \mathbb{R}^p \text{ as in (3)} \mid ||\theta - \bar{\theta}||_2 \leq \rho \right\} . \]

**Remark 3.2.** The layerwise spectral norm ball \( B_{\rho, \phi1}^{\text{Spec}} \) plays a key role in our analysis. The last layer radius of \( \rho_1 \) gives more flexibility, and we will usually assume \( \rho_1 \leq \rho \); e.g., we could choose the desirable operating regime of \( \rho < \sqrt{m} \) and \( \rho_1 = O(1) \). Our analysis in fact goes through for any choice of \( \rho, \rho_1 \) and the detailed results will indicate specific dependencies on both \( \rho \) and \( \rho_1 \).
4 SPECTRAL NORM OF THE HESSIAN OF THE MODEL

We start with the following assumption regarding the random initialization of the weights.

**Assumption 2 (Initialization weights and data normalization).** The initialization weights \( w_{0,l,i} \) \( \sim \mathcal{N}(0, \sigma_0^2) \) for \( l \leq [L] \) where \( \sigma_0 = \frac{\sigma_1}{\sqrt{1 + \frac{\sigma_1^2}{\sqrt{m}}} \theta} \), \( \sigma_1 > 0 \), and \( \mathbf{v}_0 \) is a random unit vector with \( \|\mathbf{v}_0\|_2 = 1 \). Further, we assume the input data satisfies: \( \|\mathbf{x}_i\|_2 = \sqrt{d}, i \in [n] \).

We focus on bounding the spectral norm of the Hessian \( \|\nabla^2 f(\theta; \mathbf{x})\|_2 \) for \( \theta \in B_{\rho,\rho_1}^{\Spec}(\theta_0) \) and any input \( \mathbf{x} \in \mathbb{R}^d \) with \( \|\mathbf{x}\|_2 = \sqrt{d} \). The assumption \( \|\mathbf{x}\|_2 = \sqrt{d} \) is for convenient scaling, such assumptions are common in the literature (Allen-Zhu et al., 2019; Oymak & Soltanolkotabi, 2020; Nguyen et al., 2021). Prior work (Liu et al., 2020) has considered a similar analysis for \( \theta \in B_{\rho}^{\Euc}(\theta_0) \), effectively the layerwise Frobenius norm ball, which is much smaller than \( B_{\rho,\rho_1}^{\Spec}(\theta_0) \), the layerwise spectral norm ball. We choose a unit value for the last layer’s weight norm for convenience, since our results hold under appropriate scaling for any other constant in \( O(1) \). All missing proofs are in Appendix A.

**Theorem 4.1 (Hessian Spectral Norm Bound).** Under Assumptions 1 and 2 for \( \theta \in B_{\rho,\rho_1}^{\Spec}(\theta_0) \), with probability at least \( 1 - \frac{2(L+1)}{m} \), for any \( \mathbf{x}_i, i \in [n] \), we have

\[
\|\nabla^2 f(\theta; \mathbf{x}_i)\|_2 \leq \frac{c_H}{\sqrt{m}},
\]

with \( c_H = O(L^5(1 + \gamma^6L)(1 + \rho_1)) \) where \( \gamma := \sigma_1 + \frac{\rho}{\sqrt{m}} \).

**Remark 4.1 (Desirable operating regimes).** The constant \( \gamma \) needs careful scrutiny as \( c_H \) depends on \( \gamma^6L \). Let us choose \( \rho_1 = O(\poly(L)) \). For any choice of the spectral norm radius \( \rho < \sqrt{m} \), we can choose \( \sigma_1 < 1 - \frac{\rho}{\sqrt{m}} \) ensuring \( \gamma \leq 1 \) and hence \( c_H = O(\poly(L)) \). If \( \rho = O(1) \), we can keep \( \sigma_1 = 1 \) so that \( \gamma = 1 + \frac{O(1)}{\sqrt{m}} \), and \( c_H = O(\poly(L)) \) as long as \( L < \sqrt{m} \), which is common. Both of these give good choices for \( \sigma_1 \) and desirable operating regime for the result. If we choose \( \sigma_1 > 1 \), an undesirable operating regime, then \( c_H = O(\theta^{6L}) \), \( c > 1 \), and we will need \( m = \Omega(\theta^{6L}) \) for the result to be of interest.

**Remark 4.2 (Recent Related Work).** In recent work, Liu et al. (2020) analyzed the Hessian spectral norm bound and showed that \( c_H = O(\rho^{3L}) \) for \( \theta \in B_{\rho,\rho_1}^{\Euc}(\theta_0) \) (logarithmic terms hidden in \( O(\cdot) \)). Our analysis builds on and sharpens the result in (Liu et al., 2020) in three respects: (a) we have \( c_H = O(\poly(L)) \) for \( \rho_1 = O(\poly(L)) \) where we can choose \( \sigma_1 \) to make \( \gamma \leq 1 \) and thus obtain \( c_H = O(\poly(L)) \), instead of the worse \( c_H = O(\rho^L) \) in Liu et al. (2020); (b) even for the same \( \rho, \) our results hold for a much larger spectral norm ball \( B_{\rho,\rho_1}^{\Spec}(\theta_0) \) compared to their Euclidean norm ball \( B_{\rho,\rho_1}^{\Euc}(\theta_0) \) in Liu et al. (2020); and (c) to avoid an exponential term, the bound in Liu et al. (2020) needs \( \rho \leq 1 \) whereas our result can use radius \( \rho < \sqrt{m} \) for all intermediate layer matrices and \( \rho_1 = O(\poly(L)) \) for the last layer vector. Moreover, as a consequence of (b) and (c), our results hold for a larger (spectral norm) ball whose radius can increase with \( m \), unlike the results in Liu et al. (2020) which hold for a smaller (Euclidean) ball with constant radius, i.e., “near initialization.”

**Remark 4.3 (Exact constant \( c_H \)).** For completeness, we show the exact expression of the constant \( c_H \) in Theorem 4.1 so the dependencies on different factors is clear. Let \( h(l) := \gamma^{l+1} + |\phi(0)| \sum_{i=1}^{l-1} \gamma^{i-1} \).

Then,

\[
c_H = 2L(L^2 \gamma^2 L^2 + L^2 \gamma^2 L^2 + 1 + \rho_1) \cdot \psi_H \cdot \max_{l \in [L]} \gamma^{L-l} + 2L \gamma^L \max_{l \in [L]} h(l),
\]

where

\[
\psi_H = \max_{1 \leq i \leq L} \left\{ \sigma_0^2 h(l_1)^2, h(l_1) \left( \frac{\sigma_0^2}{2} (\gamma^2 + (h(l_2))^2) + 1 \right), \beta_0 \gamma^2 h(l_1) h(l_2) \right\}.
\]

The source of the terms will be discussed shortly. Note the dependence on \( \rho_1 \), the radius for the last layer in \( B_{\rho,\rho_1}^{\Spec}(\theta_0) \), and why \( \rho_1 = O(\poly(L)) \) is a desirable operating regime.

---

1See the end of Appendix A for a quick note about the network architecture in our work and the one in Liu et al. (2020).
Next, we give a high level outline of the proof of Theorem 4.1.

**Proof sketch.** Our analysis follows the structure developed in [Liu et al. (2020), Theorem 3.1], but is considerably sharper as discussed in Remark 4.2. We start by defining the following quantities: \( Q_\infty(f) := \max_{1 \leq l \leq L} \| \frac{\partial f}{\partial \alpha_{l}} \| \infty \), \( Q_2(f) := \max_{1 \leq l \leq L} \left\{ \left\| \frac{\partial f}{\partial \alpha_{l}} \right\|_2 \right\} \), \( w : = \text{vec}(W(l)) \)， \( \partial_{\alpha_{l}}(T) \in \mathbb{R}^{m \times m} \)， and \( Q_{2.2.1}(f) \) is the maximum over \( 1 \leq l_1 < l_2 < l_3 \leq L \) among the three quantities \( \frac{\partial^2 f(l_1)}{\partial \alpha_{l_1} \partial \alpha_{l_3}} \|_{2,2,1}, \frac{\partial^2 f(l_2)}{\partial \alpha_{l_2} \partial \alpha_{l_3}} \|_{2,2,1}, \text{ and } \frac{\partial^2 f(l_3)}{\partial \alpha_{l_1} \partial \alpha_{l_2}} \|_{2,2,1} \). Under the assumption of square losses, further bounds can be obtained.

Remark 4.4. Our analysis in Lemma 4.1 provides a bound on the Lipschitz constant of the predictor, \( \rho \). Under Assumptions 1 and 2, for any \( \rho, \rho \), the probability at least \( \| \nabla f(\theta) \| \leq C_1 Q_{2.2.1}(f) Q_\infty(f) + \frac{2}{\sqrt{m}} C_2 Q_2(f) \), (8)

In order to prove Theorem 4.1, we prove that Theorem 4.2 holds with high-probability where \( \delta = \gamma \). \( Q_2(f) = O(L(1 + \gamma L)) \), \( Q_{2.2.1}(f) = O(L^3(1 + \gamma^3 L)) \), and \( Q_\infty(f) = O \left( \frac{(1+\gamma)(1+\rho_1)}{\sqrt{m}} \right) \). Thus we obtain that the upper bound (4.2) becomes \( O \left( \frac{\text{poly}(L)(1+\gamma^3 L)}{\sqrt{m}} \right) \), providing a benign polynomial dependence on \( L \) when \( \gamma \leq 1 \), rather than an exponential dependence on the radius \( \rho \) as in [Liu et al. (2020)].

The analysis for bounding the spectral norm of the Hessian can be used to establish additional bounds, which we believe are of independent interest, some of which will be used later in Section 5. First, we bound the norms of gradient of the predictor and the loss w.r.t. the weight vector \( \theta \) and the input data \( x \).

**Lemma 4.1 (Predictor gradient bounds).** Under Assumptions 1 and 2 for \( \theta \in B^{\text{spec}}_{\rho_1, \rho_3}(\theta_0) \), with probability at least \( \left( 1 - \frac{2(L+1)}{m} \right) \), we have \( \| \nabla_{\theta} f(\theta; x) \|_2 \leq \varrho \) and \( \| \nabla_{x} f(\theta; x) \|_2 \leq \frac{\gamma^L}{\sqrt{m}} (1 + \rho_1) \), (9)

with \( \varrho^2 = (h(L + 1))^2 + \frac{1}{m} (1 + \rho_1)^2 \sum_{l=1}^{L} (h(l))^2 \gamma^{2(L-l)}, \gamma = \sigma_1 + \frac{\rho}{\sqrt{m}}, h(l) = \gamma^{l-1} + |\phi(0)| \sum_{l=1}^{L} \gamma^{l-1} \).

Remark 4.4. Our analysis in Lemma 4.1 provides a bound on the Lipschitz constant of the predictor, a quantity which has generated interest in recent work on robust training [Salman et al., 2019; Cohen et al., 2020; Bubeck & Sellke 2021].

Under the assumption of square losses, further bounds can be obtained.

**Lemma 4.2 (Loss bounds).** Consider the square loss. Under Assumptions 1 and 2 for \( \gamma = \sigma_1 + \frac{\rho}{\sqrt{m}} \), each of the following inequalities hold with probability at least \( \left( 1 - \frac{2(L+1)}{m} \right) \): \( \mathcal{L}(\theta_0) \leq c_{a,b} \), and \( \mathcal{L}(\theta) \leq c_{\rho_1, \gamma} \) for \( \theta \in B^{\text{spec}}_{\rho_1, \rho_3}(\theta_0) \), where \( c_{a,b} = \frac{2}{n} \sum_{i=1}^{n} y_i^2 + 2(1 + a)^2 |g(b)|^2 \) and \( g(a) = a^L + |\phi(0)| \sum_{i=1}^{L} a^i \) for any \( a, b \in \mathbb{R} \).

**Corollary 4.1 (Loss gradient bound).** Consider the square loss. Under Assumptions 1 and 2 for \( \theta \in B^{\text{spec}}_{\rho_1, \rho_3}(\theta_0) \), with probability at least \( \left( 1 - \frac{2(L+1)}{m} \right) \), we have \( \| \nabla_{\theta} \mathcal{L}(\theta) \|_2 \leq 2 \sqrt{\mathcal{L}(\theta)} \varrho \leq 2 \sqrt{c_{\rho_1, \gamma} \varrho} \), with \( \varrho \) as in Lemma 4.1 and \( c_{\rho_1, \gamma} \) as in Lemma 4.2.
5 Optimization Guarantees with Restricted Strong Convexity

We focus on minimizing the empirical loss \( \mathcal{L}(\theta) \) over \( \theta \in B_{p;\rho_1}^{\text{Spec}}(\theta_0) \), the layerwise spectral norm ball in \([3]\). Our analysis is based on Restricted Strong Convexity (RSC) \( \text{(Negahban et al., 2012; Banerjee et al., 2014; [4],[5]} \), which relaxes the definition of strong convexity by only needing strong convexity in certain directions or a subset of the ambient space. We introduce the following specific definition of RSC with respect to a tuple \((S, \theta)\).

**Definition 5.1 (Restricted Strong Convexity (RSC)).** A function \( \mathcal{L} \) is said to satisfy \( \alpha \)-restricted strong convexity \((\alpha \text{-RSC})\) with respect to the tuple \((S, \theta)\) if for any \( \theta' \in S \subseteq \mathbb{R}^p \) and some fixed \( \theta \in \mathbb{R}^p \), we have \( \mathcal{L}(\theta') \geq \mathcal{L}(\theta) + \langle \theta' - \theta, \nabla_{\theta} \mathcal{L}(\theta) \rangle + \frac{\alpha}{2} \| \theta' - \theta \|^2_2 \), with \( \alpha > 0 \).

Note that \( \mathcal{L} \) being \( \alpha \)-RSC w.r.t. \((S, \theta)\) does not need \( \mathcal{L} \) to be convex on \( \mathbb{R}^p \). Let us consider a sequence of iterates \( \{\theta_t\}_{t \geq 0} \subseteq \mathbb{R}^p \). Our RSC analysis will rely on the following \( Q_t^\kappa \)-sets at step \( t \), which avoid directions almost orthogonal to the average gradient of the predictor. We define the following notation: for two vectors \( \pi \) and \( \bar{\pi} \), \( \cos(\theta - \pi, \bar{\pi}) \) denotes the cosine of the angle between \( \pi \) and \( \bar{\pi} \).

**Definition 5.2 (\( Q_t^\kappa \)-sets).** For iterate \( \theta_t \in \mathbb{R}^p \), let \( g_t = \frac{1}{n} \sum_{i=1}^{n} \nabla f(\theta_t; x_i) \). For any \( \kappa \in (0, 1] \), define \( Q_t^\kappa := \{ \theta \in \mathbb{R}^p | \| \cos(\theta - \theta_t, g_t) \| \geq \kappa \} \).

We define the set \( B_t := Q_t^\kappa \cap B_{p;\rho_1}^{\text{Spec}}(\theta_0) \cap B_{p;\rho_2}^{\text{Euc}}(\theta_t) \). We focus on establishing RSC w.r.t. the tuple \((B_t, \theta_t)\), where \( B_{p;\rho_1}^{\text{Spec}}(\theta_0) \) becomes the feasible set for the optimization and \( B_{p;\rho_2}^{\text{Euc}}(\theta_t) \) is a Euclidean ball around the current iterate.

**Assumption 3 (Loss function).** The loss \( \ell_{i}, i \in [n] \), is (i) strongly convex, i.e., \( \ell''_i \geq a > 0 \) and (ii) smooth, i.e., \( \ell''_i \leq b \).

Assumption 3 is satisfied by commonly used loss functions such as square loss, where \( a = b = 2 \). We state the RSC result for square loss; the result for other losses and proofs of all technical results in this section are in Appendix 3.

**Theorem 5.1 (RSC for Square Loss).** For square loss, under Assumptions 3 and 2 with probability at least \( 1 - \frac{2(L+1)}{m} \), \( \forall \theta' \in Q_t^\kappa \cap B_{p;\rho_1}^{\text{Spec}}(\theta_0) \cap B_{p;\rho_2}^{\text{Euc}}(\theta_t) \) with \( \theta_t \in B_{p;\rho_1}^{\text{Spec}}(\theta_0) \),

\[
\mathcal{L}(\theta') \geq \mathcal{L}(\theta_t) + \langle \theta' - \theta_t, \nabla_{\theta} \mathcal{L}(\theta_t) \rangle + \frac{\alpha_t}{2} \| \theta' - \theta_t \|^2_2 , \quad \text{with} \quad \alpha_t = c_1 \| g_t \|^2 - \frac{c_2}{\sqrt{m}},
\]

where \( g_t = \frac{1}{n} \sum_{i=1}^{n} \nabla f(\theta_t; x_i) \), \( c_1 = 2k^2 \) and \( c_2 = 2c_H(2\rho_2 + \sqrt{c_{p,1}\gamma}) \), with \( c_H \) as in Theorem 4.1, \( \rho \) as in Lemma 4.1, and \( c_{p,1,\gamma} \) as in Lemma 4.2. Consequently, \( \mathcal{L} \) satisfies RSC w.r.t. \((Q_t^\kappa \cap B_{p;\rho_1}^{\text{Spec}}(\theta_0) \cap B_{p;\rho_2}^{\text{Euc}}(\theta_t)) \) whenever \( \alpha_t > 0 \).

**Remark 5.1.** The RSC condition \( \alpha_t > 0 \) is satisfied at iteration \( t \) as long as \( \| g_t \|^2 > \frac{c_2}{c_1 \sqrt{m}} \) where \( c_1, c_2 \) are exactly specified in Theorem 5.1. Indeed, if \( \gamma \) (and so \( \sigma_1 \) and \( \rho \)) is chosen according to the desirable operating regimes (see Remark 4.1), \( \rho_1 = O(\text{poly}(\Gamma)) \) and \( \rho_2 = O(\text{poly}(\Gamma)) \), then we can use the bounds from Lemma 4.2 and obtain that the RSC condition is satisfied when \( \| g_t \|^2 > \frac{O(\text{poly}(\Gamma))}{\sqrt{m}} \). The condition is arguably mild; does not need the NTK condition \( \lambda_{\text{min}}(K_{\text{ntk}}(\cdot; \theta_t)) > 0 \), and is expected to hold till convergence (see Remark 5.3). Moreover, it is a local condition at step \( t \) and has no dependence on being "near initialization" in the sense of \( \theta_t \in B_{p;\rho_1}^{\text{Euc}}(\theta_0) \) for \( \rho = O(1) \) as in [6].

For the convergence analysis, we also need to establish a smoothness property of the total loss.

**Theorem 5.2 (Local Smoothness for Square Loss).** For square loss, under Assumptions 3 and 2 with probability at least \( 1 - \frac{2(L+1)}{m} \), \( \forall \theta, \theta' \in B_{p;\rho_1}^{\text{Spec}}(\theta_0) \),

\[
\mathcal{L}(\theta') \leq \mathcal{L}(\theta) + \langle \theta' - \theta, \nabla_{\theta} \mathcal{L}(\theta) \rangle + \frac{\beta}{2} \| \theta' - \theta \|^2_2 , \quad \text{with} \quad \beta = 2 \rho_2^2 + \frac{2c_H \sqrt{c_{p,1}\gamma}}{\sqrt{m}},
\]

with \( c_H \) as in Theorem 4.1, \( \rho \) as in Lemma 4.1, and \( c_{p,1,\gamma} \) as in Lemma 4.2. Consequently, \( \mathcal{L} \) is locally \( \beta \)-smooth. Moreover, \( \sigma_1 \) (and so \( \sigma_1 \) and \( \rho \)) is chosen according to the desirable operating regimes (see Remark 4.1) and \( \rho_1 = O(\text{poly}(\Gamma)) \), then \( \beta = O(\text{poly}(\Gamma)) \).
We now consider the particular case of gradient descent (GD) for the iterates:

\[ L(\theta_t) - \inf_{\theta \in B_t} L(\theta) \leq \frac{1}{2\alpha_t} \| \nabla_{\theta} L(\theta_t) \|_2^2, \tag{12} \]

with probability at least \(1 - \frac{2(L+1)}{m}\).

For the rest of the convergence analysis, we make the following assumption where \(T\) can be viewed as the stopping time so the convergence analysis holds given the assumptions are satisfied.

**Assumption 4 (Iterates’ conditions).** For iterates \(\{\theta_t\}_{t=0,1,...,T}\): (A4.1) \(\alpha_t > 0\); (A4.2) \(\theta_t \in B_{p_{p_1}}^{\text{Spec}}(\theta_0)\).

**Remark 5.3 (Assumption (A4.1)).** From Remark 5.1 (A4.1) is satisfied as long as \(\|\dot{g}_t\|_2^2 > \frac{c_2}{c_1 \omega_t}\) where \(c_1, c_2\) are as in Theorem 5.1 which is arguably a mild condition. In Section 6 we will present some empirical findings that show that this condition on \(\|\dot{g}_t\|_2^2\) behaves well empirically.

We now consider the particular case of gradient descent (GD) for the iterates: \(\theta_{t+1} = \theta_t - \eta_t \nabla L(\theta_t)\), where \(\eta_t\) is chosen so that \(\theta_{t+1} \in B_{p_{p_1}}^{\text{Spec}}(\theta_t)\) and \(p_2\) is chosen so that \(\theta_{t+1} \in B_{p_{p_2}}^{\text{Euc}}(\theta_t)\), which are sufficient for the analysis of Theorem 5.1 — we specify suitable choices in the sequel (see Remark 5.4).

Given RPL w.r.t. \((B_t, \theta_t)\), gradient descent leads to a strict decrease of loss in \(B_t\).

**Lemma 5.2 (Local Loss Reduction in \(B_t\)).** Let \(\alpha_t, \beta\) be as in Theorems 5.1 and 5.2 respectively, and \(B_t := Q^*_t \cap B_{p_{p_1}}^{\text{Spec}}(\theta_0) \cap B_{p_{p_2}}^{\text{Euc}}(\theta_t)\). Consider Assumptions 4.1 and 4.2 and gradient descent with step size \(\eta_t = \frac{\omega_t}{\beta}, \omega_t \in (0, 2)\). Then, for any \(\bar{\theta}_{t+1} \in \arg\inf_{\theta \in B_t} L(\theta)\), we have with probability at least \(1 - \frac{2(L+1)}{m}\),

\[ L(\theta_{t+1}) - L(\bar{\theta}_{t+1}) \leq \left(1 - \frac{\alpha_t \omega_t}{\beta}(2 - \omega_t)\right) (L(\theta_t) - L(\bar{\theta}_{t+1})). \tag{13} \]

Building on Lemma 5.2, we show that GD in fact leads to a geometric decrease in the loss relative to the minimum value of \(L(\cdot)\) in the set \(B_{p_{p_1}}^{\text{Spec}}(\theta_0)\).

**Theorem 5.3 (Global Loss Reduction in \(B_{p_{p_1}}^{\text{Spec}}(\theta_0)\)).** Let \(\alpha_t, \beta\) be as in Theorems 5.1 and 5.2 respectively, and \(B_t := Q^*_t \cap B_{p_{p_1}}^{\text{Spec}}(\theta_0) \cap B_{p_{p_2}}^{\text{Euc}}(\theta_t)\). Let \(\bar{\theta}^* \in \arg\inf_{\theta \in B_{p_{p_1}}^{\text{Spec}}(\theta_0)} L(\theta)\), \(\bar{\theta}_{t+1} \in \arg\inf_{\theta \in B_t} L(\theta)\), and \(\gamma_t := \frac{L(\bar{\theta}_{t+1}) - L(\bar{\theta}^*)}{L(\theta_t) - L(\bar{\theta}^*)}\). Then, with probability at least \(1 - \frac{2(L+1)}{m}\), we have we have \(\gamma_t \in [0, 1)\) and

\[ L(\theta_{t+1}) - L(\bar{\theta}^*) \leq \left(1 - \frac{\alpha_t \omega_t}{\beta}(1 - \gamma_t)(2 - \omega_t)\right) (L(\theta_t) - L(\bar{\theta}^*)). \tag{14} \]

As long as the conditions in Theorem 5.3 are kept across iterations, there will be a geometric decrease in loss. For Assumption 4 we have discussed (A4.1) in Remark 5.1 and we discuss (A4.2) next.
Remark 5.4 (Assumption (A4.2)). Consider we run gradient descent iterations until some stopping time $T > 0$. Given radius $\rho < \sqrt{m}$. Assumption (A4.2) $\theta_t \in B_{\rho,\rho}^{\text{spec}}(\theta_0)$, $t = 0, \ldots, T$, can be verified empirically. Alternatively, we can choose suitable step sizes $\eta_t$ to ensure the property using the geometric convergence from Theorem 5.3. Assume that our goal is to get $\mathcal{L}(\theta_T) - \mathcal{L}(\theta^*) \leq \epsilon$. Then, with $\chi_T := \min_{t \in [T]} \frac{2}{\rho^2} (1 - \gamma_t)(2 - \omega_t)$, Assumption (A4.1) along with Remark 5.2 ensures $\chi_T < 1$. Then, it suffices to have $T = \lceil \log \left( \frac{\mathcal{L}(\theta_0) - \mathcal{L}(\theta^*)}{\epsilon} \right) / \log \left( \frac{1}{1 - \chi_T} \right) \rceil = \Theta \left( \log \frac{1}{\epsilon} \right)$. Then, to ensure $\theta_t \in B_{\rho,\rho}^{\text{spec}}(\theta_0), t \in [T]$, in the case of the square loss, since $\|\nabla \mathcal{L}(\theta_t)\|_2 \leq c$ for some constant $c$ (see Corollary 4.1), it suffices to have $\eta_t \leq \min \left( \frac{\rho,\rho}{c(\log \frac{1}{\epsilon})} \right)$. Moreover, we point out that having $\rho_2 \geq \rho c$ ensures $\|\theta_{t+1} - \bar{\theta}_t\|_2 \leq \rho_2 \Rightarrow \theta_{t+1} \in B_{\rho,\rho}^{\text{spec}}(\theta_0)$, which in this case can be guaranteed if $\rho_2 \geq \min \left( \frac{\rho,\rho_2}{c(\log \frac{1}{\epsilon})} \right)$. The argument above is informal, but illustrates that Assumption (A4.1) along with suitable constant step sizes $\eta_t$ would ensure (A4.2). Thus, Assumption (A4.1), which ensures the RSC condition, is the main assumption behind the analysis.

The conditions in Assumption 4 (see Remarks 5.1 and 5.4) along with Theorem 5.3 imply that the RSC based convergence analysis holds for a much larger layerwise spectral radius norm ball $B_{\rho,\rho}^{\text{spec}}(\theta_0)$ with any radius $\rho < \sqrt{m}$ and $\rho_1 = O(\text{poly}(L))$.

Remark 5.5 (RSC and NTK). In the context of square loss, the NTK condition for geometric convergence needs $\lambda_{\min}(K_{\text{ntk}}(\cdot; \theta_t)) \geq c_0 > 0$ for every $t$, i.e., uniformly bounded away from 0 by a constant $c_0 > 0$. The NTK condition can also be written as

$$\inf_{\|v\|_2 = 1} \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla \theta f(\theta_t; x_i) \right\|_2^2 \geq c_0 > 0. \quad (15)$$

In contrast, the proposed RSC condition (Theorem 5.1) needs

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \nabla \theta f(\theta_t; x_i) \right\|_2^2 \geq \frac{c_0}{\sqrt{m}}. \quad (16)$$

where $m$ is the width and $c_0 = \frac{c_1}{c_2}$ where $c_1, c_2$ are constants defined in Theorem 5.1. As a quadratic form on the NTK, the RSC condition can be viewed as using a specific $v$ in (15), i.e., $v_i = \frac{1}{\sqrt{n}}$ for $i \in [n]$, since the RSC condition is $\left\| \sum_{i=1}^{n} \frac{1}{\sqrt{n}} \nabla \theta f(\theta_t; x_i) \right\|_2^2 \geq \frac{c_0}{\sqrt{m}}$. For $m = \Omega(n^2)$, the RSC condition is more general since NTK $\Rightarrow$ RSC, but the converse is not necessarily true.

Remark 5.6 (RSC covers different settings than NTK). The NTK condition may be violated in certain settings, e.g., $\nabla \theta f(\theta_t; x_i), i = 1, \ldots, n$ are linearly dependent, $x_i \approx x_j$ for some $i \neq j$, layer widths are small $m_t < n$, etc., but the optimization may work in practice. The RSC condition provides a way to analyze convergence in such settings. The RSC condition gets violated when $\frac{1}{n} \sum_{i=1}^{n} \nabla \theta f(\theta_t; x_i) \approx 0$, which does not seem to happen in practice (see Section 6), and future work will focus on understanding the phenomena. Finally, note that it is possible to construct a set of gradient vectors which satisfy the NTK condition but violates the RSC condition. Our perspective is to view the NTK and the RSC as two different sufficient conditions and geometric convergence of gradient descent (GD) is guaranteed as long as one of them is satisfied in any step.

6 RSC CONDITION: EXPERIMENTAL RESULTS

In this section, we present experimental results verifying the RSC condition $\left\| \frac{1}{n} \sum_{i=1}^{n} \nabla \theta f(\theta_t; x_i) \right\|_2^2 = \Omega \left( \frac{\text{poly}(L)}{\sqrt{m}} \right), t = 1, \ldots, T$, on standard benchmarks: CIFAR-10, MNIST, and Fashion-MNIST. For simplicity, as before, we use $\bar{\theta}_t = \frac{1}{n} \sum_{i=1}^{n} \nabla \theta f(\theta_t; x_i)$.

In Figure 1(a), we consider CIFAR-10 and show the trajectory of $\|\bar{\theta}_t\|_2$ over iterations $t$, for different values of the network width $m$. For any width, the value of $\|\bar{\theta}_t\|_2$ stabilizes to a constant value over iterations, empirically validating the RSC condition $\|\bar{\theta}_t\|_2^2 = \Omega(\text{poly}(L)\sqrt{m})$. Interestingly, the smallest value of $\|\bar{\theta}_t\|_2$ seems to increase with the width. To study the width dependence further, in Figure 1(b), we plot $\min_{t \in [T]} \|\bar{\theta}_t\|_2$ as a function of width $m$ for several values of the width.
The plot shows that $\min_{t \in [T]} \| \bar{g}_t \|_2$ increases steadily with $m$ illustrating that the RSC condition is empirically satisfied more comfortably for wider networks. In Figure 1(c) and (d), we show similar plots for MNIST and Fashion-MNIST illustrating the same phenomena of $\min_{t \in [T]} \| \bar{g}_t \|_2$ increasing with $m$.

For the experiments, the network architecture we used had 3-layer fully connected neural network with tanh activation function. The training algorithm is gradient descent (GD) width constant learning rate, chosen appropriately to keep the training in NTK regime. Since we are using GD, we use 512 randomly chosen training points for the experiments. The stopping criteria is either training loss $< 10^{-3}$ or number of iterations larger than 3000.

7 CONCLUSIONS

In this paper, we revisit deep learning optimization for feedforward models with smooth activations, and make two technical contributions. First, we bound the spectral norm of the Hessian over a large layerwise spectral norm radius ball, highlighting the role of initialization in such analysis. Second, we introduce a new approach to showing geometric convergence in deep learning optimization using restricted strong convexity (RSC). Our analysis sheds considerably new light on deep learning optimization problems, underscores the importance of initialization variance, and introduces a RSC based alternative to the prevailing NTK based analysis, which may fuel future work.
ACKNOWLEDGMENTS

AB is grateful for support from the National Science Foundation (NSF) through awards IIS 21-31335, OAC 21-30835, DBI 20-21898, as well as a C3.ai research award. MB and LZ are grateful for support from the National Science Foundation (NSF) and the Simons Foundation for the Collaboration on the Theoretical Foundations of Deep Learning through awards DMS-2031883 and 814639 as well as NSF IIS-1815697 and the TILOS institute (NSF CCF-2112665).

REFERENCES

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pp. 242–252. PMLR, 2019.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc., 2019.

A. Banerjee, S. Chen, F. Fazayeli, and V. Sivakumar. Estimation with norm regularization. In Advances in Neural Information Processing Systems (NIPS), 2014.

Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. Acta Numerica, 30:87–201, 2021.

Sebastien Bubeck and Mark Sellke. A universal law of robustness via isoperimetry. In Neural Information Processing Systems (NeurIPS), 2021.

S. Chen and A. Banerjee. Structured estimation with atomic norms: General bounds and applications. In Advances in Neural Information Processing Systems (NIPS), 2015.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. Advances in Neural Information Processing Systems, 32, 2019.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In International Conference on Machine Learning (ICML), 2020.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pp. 1675–1685. PMLR, 09–15 Jun 2019.

Jianqing Fan, Cong Ma, and Yiqiao Zhong. A selective overview of deep learning. Statistical Science, 36(2):264 – 290, 2021.

Cong Fang, Hanze Dong, and Tong Zhang. Mathematical models of overparameterized neural networks. Proceedings of the IEEE, 109(5):683–703, 2021.

Stanislav Fort, Gintare Karolina Dziugaite, Mansheej Paul, Sepideh Kharaghani, Daniel M Roy, and Surya Ganguli. Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. Advances in Neural Information Processing Systems, 33:5850–5861, 2020.

Mario Geiger, Stefano Spigler, Arthur Jacot, and Matthieu Wyart. Disentangling feature and lazy training in deep neural networks. Journal of Statistical Mechanics: Theory and Experiment, 2020 (11):113301, 2020.

Jiaoyang Huang and Horng-Tzer Yau. Dynamics of deep neural networks and neural tangent hierarchy. In International conference on machine learning, pp. 4542–4551. PMLR, 2020.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In Advances in Neural Information Processing Systems, volume 31. Curran Associates, Inc., 2018.
Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. In *International Conference on Learning Representations*, 2019.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-Lojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 795–811. Springer, 2016.

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

C. Liu, L. Zhu, and M. Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022.

Chaoyue Liu, Libin Zhu, and Misha Belkin. On the linearity of large non-linear models: when and why the tangent kernel is constant. In *Advances in Neural Information Processing Systems*, 2020.

Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13(1): 1665–1697, 2012.

Sahand Negahban, Pradeep Ravikumar, Martin Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4), 2012.

Q. Nguyen and M. Mondelli. Global convergence of deep networks with one wide layer followed by pyramidal topology. In *Neural Information Processing Systems (NeurIPS)*, 2020.

Quynh Nguyen. On the proof of global convergence of gradient descent for deep relu networks with linear widths. In *International Conference on Machine Learning (ICML)*, 2021.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing: Theory and Applications*, pp. 210–268. Cambridge University Press, 2012.

S. Oymak and M. Soltanolkotabi. Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 2020.

Hadi Salman, Greg Yang, Jerry Li, Pengchuan Zhang, Huan Zhang, Ilya Razenshteyn, and Sébastien Bubeck. Provably robust deep learning via adversarially trained smoothed classifiers. In *Neural Information Processing Systems (NeurIPS)*, 2019.

Greg Yang and Edward J Hu. Feature learning in infinite-width neural networks. *arXiv preprint arXiv:2011.14522*, 2020.

Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine Learning*, 109:467–492, 2020.
A Spectral Norm of the Hessian

We establish the main theorem from Section 4 in this Appendix.

**Theorem 4.1 (Hessian Spectral Norm Bound).** Under Assumptions 1 and 2 for \( \theta \in B_{\mathrm{Spec}}(\theta_0) \), with probability at least \((1 - \frac{2(1 + L)}{m})\), for any \( x_i, i \in [n] \), we have

\[
\left\| \nabla^2 f(\theta; x_i) \right\|_2 \leq \frac{c_H}{\sqrt{m}},
\]

with \( c_H = O(L^5(1 + \gamma^6L)(1 + \rho_1)) \) where \( \gamma := \sigma_1 + \frac{\rho}{\sqrt{m}} \).

### A.1 Analysis Outline

Our analysis follows that of Liu et al. (2020) and sharpens the analysis to get better dependence on the depth \( L \) of the neural network.

We start by defining the following quantities:

\[
Q_\infty(f) := \max_{1 \leq i \leq L} \left\{ \frac{\partial f}{\partial \alpha^{(i)}} \right\}_\infty, \quad \frac{\partial f}{\partial \alpha^{(i)}} \in \mathbb{R}^m,
\]

\[
Q_2(f) := \max_{1 \leq i \leq L} \left\{ \frac{\partial \alpha^{(i)}}{\partial w^{(i)}} \right\}_2, \quad w^{(i)} := \text{vec}(W^{(i)}), \quad \frac{\partial \alpha^{(i)}}{\partial w^{(i)}} \in \mathbb{R}^{m \times m^2},
\]

\[
Q_{2,2,1}(f) := \max_{1 \leq i < j < k \leq L} \left\{ \frac{\partial^2 \alpha^{(i)}}{\partial w^{(i)} \partial w^{(j)}}, \frac{\partial \alpha^{(i)}}{\partial w^{(j)}}, \frac{\partial \alpha^{(i)}}{\partial w^{(k)}}, \frac{\partial^2 \alpha^{(i)}}{\partial w^{(j)} \partial w^{(k)}} \right\}_{2,2,1},
\]

where for an order-3 tensor \( T \in \mathbb{R}^{d_1 \times d_2 \times d_3} \) we define the \((2, 2, 1)-\)norm as follows,

\[
\left\| T \right\|_{2,2,1} := \sup_{\| x \|_2 = \| z \|_2 = 1} \sum_{k=1}^{d_3} \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} T_{ijk} x_i z_j, \quad x \in \mathbb{R}^{d_1}, z \in \mathbb{R}^{d_2}.
\]

We will also use the notation \( W^{(L+1)} := v \).

A key result established in Liu et al. (2020) provides an upper bound to the spectral norm of the Hessian:

**Theorem 4.2 (Liu et al. (2020), Theorem 3.1).** Under Assumptions 1 assuming there is \( \delta \) such that \( \left\| \frac{\partial \alpha^{(i)}}{\partial w^{(i)} \alpha^{(i)}} \right\|_2 \leq \delta \), with \( C_1 \leq L^2 \delta^{2L} + L \delta^L + L \) and \( C_2 \leq L \delta^L \), we have

\[
\left\| \nabla^2 f(\theta; x) \right\|_2 \leq 2C_1 Q_{2,2,1}(f) Q_\infty(f) + \frac{2}{\sqrt{m}} C_2 Q_2(f),
\]

In order to prove Theorem 4.1, we prove that Theorem 4.2 holds with high-probability where

- \( \delta = \gamma \) follows from Lemma A.3
- \( Q_2(f) = O(L(1 + \gamma^L)) \) follows from Lemma A.4
- \( Q_{2,2,1}(f) = O(L^3(1 + \gamma^3L)) \) follows from Lemma A.4 and Lemma A.5 and
- \( Q_\infty(f) = O\left( \frac{L^5(1 + \gamma^6L)(1 + \rho_1)}{\sqrt{m}} \right) \) follows from Lemma A.7.

while also establishing precise constants to get a proper form for the constant \( c_H \) in Theorem 4.1. As a result, \( c_H \leq O\left( \frac{L^5(1 + \gamma^6L)(1 + \rho_1)}{\sqrt{m}} \right) \).
A.2 Spectral norms of $W^{(l)}$ and $L_2$ norms of $\alpha^{(l)}$

We start by bounding the spectral norm of the layer-wise matrices at initialization.

**Lemma A.1.** Consider any $l \in [L]$. If the parameters are initialized as $w_{0,i,j}^{(l)} \sim \mathcal{N}(0, \sigma_0^2)$ where $\sigma_0 = \frac{\sigma_1}{2(1+\sqrt{\frac{m_0}{m}})}$ as in Assumption 2 then with probability at least $(1 - \frac{2}{m})$, we have

\[
\|W_0^{(l)}\|_2 \leq \sigma_1 \sqrt{m}. \tag{22}
\]

**Proof.** For a $(m_l \times m_{l-1})$ random matrix $W_0^{(l)}$ with i.i.d. entries $w_{0,i,j}^{(l)} \sim \mathcal{N}(0, \sigma_0^2)$, with probability at least $(1 - 2 \exp(-t^2/2\sigma_0^2))$, the largest singular value of $W_0$ is bounded by

\[
\sigma_{\text{max}}(W_0^{(l)}) \leq \sigma_0(\sqrt{m_l} + \sqrt{m_{l-1}}) + t.
\]

This concentration result can be easily derived as follows: notice that $W_0 = \sigma_0 W_0^{(l)}$, where $\tilde{w}_{0,i,j}^{(l)} \sim \mathcal{N}(0, 1)$, thus we can use the expectation $E[\|W_0^{(l)}\|_2^2] = \sigma_0^2 E[\|W_0\|_2^2] = \sigma_0^2 (\sqrt{m_l} + \sqrt{m_{l-1}})$ from Gordon’s Theorem for Gaussian matrices [Vershynin, 2012, Theorem 5.32] in the Gaussian concentration result for Lipschitz functions [Vershynin, 2012, Proposition 3.4] considering that $B \mapsto \|\sigma_0 B\|_2$ is a $\sigma_0$-Lipschitz function when the matrix $B$ is treated as a vector. Let us choose $t = \sigma_0 \sqrt{2\log m}$ so that (23) holds with probability at least $(1 - \frac{2}{m})$. Then, to obtain (22).

**Case 1:** $l = 1$. With $m_0 = d$ and $m_1 = m$,

\[
\|W_0^{(1)}\|_2 \leq \sigma_0(\sqrt{d} + \sqrt{m} + \sqrt{2\log m}) \leq \sigma_0(2\sqrt{m} + \sqrt{2\log m})
\]

since we are in the over-parameterized regime $m \geq d$.

**Case 2:** $2 \leq l \leq L$. With $m_l = m_{l-1} = m$,

\[
\|W_0^{(l)}\|_2 \leq \sigma_0(2\sqrt{m} + \sqrt{2\log m})
\]

Now, using $\sigma_0 = \frac{\sigma_1}{2(1+\sqrt{\frac{m_0}{m}})}$ in both cases completes the proof. \hfill \square

Next we bound the spectral norm of layerwise matrices.

**Proposition A.1.** Under Assumptions 2 for $\theta \in B^{\text{Spec}}_{\rho,p,l_1}(\theta_0)$, with probability at least $(1 - \frac{2}{m})$, 

\[
\|W^{(l)}\|_2 \leq \left( \sigma_1 + \frac{\rho}{\sqrt{m}} \right) \sqrt{m}, \quad l \in [L].
\]

**Proof.** By triangle inequality, for $l \in [L]$,

\[
\|W^{(l)}\|_2 \leq \|W_0^{(l)}\|_2 + \|W^{(l)} - W_0^{(l)}\|_2 \overset{(a)}{=} \sigma_1 \sqrt{m} + \rho,
\]

where (a) follows from Lemma A.1. This completes the proof. \hfill \square

Next, we show that the output $\alpha^{(l)}$ of layer $l$ has an $L_2$ norm bounded by $O(\sqrt{m})$.

**Lemma A.2.** Consider any $l \in [L]$. Under Assumptions 7 and 2 for $\theta \in B^{\text{Spec}}_{\rho,p,l_1}(\theta_0)$, with probability at least $(1 - \frac{2}{m})$, we have

\[
\|\alpha^{(l)}\|_2 \leq \sqrt{m} \left( \sigma_1 + \frac{\rho}{\sqrt{m}} \right)^l + \sqrt{m} \sum_{i=1}^l \left( \sigma_1 + \frac{\rho}{\sqrt{m}} \right)^{i-1} |\phi(0)| \left( \gamma^l + |\phi(0)| \sum_{i=1}^l \gamma^{i-1} \right) \sqrt{m}.
\]
Proof. Following [Allen-Zhu et al. (2019); Liu et al. (2020)], we prove the result by recursion. First, recall that since \( \|x\|_2^2 = d \), we have \( \|\alpha^{(0)}\|_2 = \sqrt{d} \). Then, since \( m_0 = d \) and \( \phi \) is 1-Lipschitz,

\[
\| \phi \left( \frac{1}{\sqrt{d}} W^{(1)} \alpha^{(0)} \right) \|_2 - \| \phi(0) \|_2 \leq \| \phi \left( \frac{1}{\sqrt{d}} W^{(1)} \alpha^{(0)} \right) - \phi(0) \|_2 \leq \frac{1}{\sqrt{d}} \| W^{(1)} \alpha^{(0)} \|_2 ,
\]

so that

\[
\|\alpha^{(1)}\|_2 \leq \| \phi \left( \frac{1}{\sqrt{d}} W^{(1)} \alpha^{(0)} \right) \|_2 \leq \frac{1}{\sqrt{d}} \| W^{(1)} \alpha^{(0)} \|_2 + \| \phi(0) \|_2 \leq \frac{1}{\sqrt{d}} \| W^{(1)} \alpha^{(0)} \|_2 + \frac{\rho}{\sqrt{m}} \| \phi(0) \|_2 ,
\]

\[
\leq \left( \sigma_1 + \frac{\rho}{\sqrt{m}} \right) \sqrt{m} + \frac{\rho}{\sqrt{m}} \| \phi(0) \|_2 ,
\]

where we used Proposition [A.1] in the last inequality, which holds with probability at least \( 1 - \frac{2}{m} \). For the inductive step, we assume that for some \( l - 1 \), we have

\[
\|\alpha^{(l-1)}\|_2 \leq \sqrt{m} \left( \sigma_1 + \frac{\rho}{\sqrt{m}} \right) \sum_{i=1}^{l-1} \left( \sigma_1 + \frac{\rho}{\sqrt{m}} \right) |\phi(0)| ,
\]

which holds with the probability at least \( 1 - \frac{2(l-1)}{m} \). Since \( \phi \) is 1-Lipschitz, for layer \( l \), we have

\[
\| \phi \left( \frac{1}{\sqrt{m}} W^{(l)} \alpha^{(l-1)} \right) \|_2 - \| \phi(0) \|_2 \leq \| \phi \left( \frac{1}{\sqrt{m}} W^{(l)} \alpha^{(l-1)} \right) - \phi(0) \|_2 \leq \frac{1}{\sqrt{m}} \| W^{(l)} \alpha^{(l-1)} \|_2 ,
\]

so that

\[
\|\alpha^{(l)}\|_2 \leq \| \phi \left( \frac{1}{\sqrt{m}} W^{(l)} \alpha^{(l-1)} \right) \|_2 \leq \frac{1}{\sqrt{m}} \| W^{(l)} \alpha^{(l-1)} \|_2 + \| \phi(0) \|_2 \leq \frac{1}{\sqrt{m}} \| W^{(l)} \alpha^{(l-1)} \|_2 + \sqrt{m} |\phi(0)|
\]

\[
\leq \left( \sigma_1 + \frac{\rho}{\sqrt{m}} \right) \|\alpha^{(l-1)}\|_2 + \sqrt{m} |\phi(0)|
\]

\[
\leq \left( \sigma_1 + \frac{\rho}{\sqrt{m}} \right) \sum_{i=1}^{l-1} \left( \sigma_1 + \frac{\rho}{\sqrt{m}} \right)^{i-1} |\phi(0)| ,
\]

where (a) follows from Proposition [A.1] and (b) from the inductive step. Since we have used Proposition [A.1] \( \ell \) times, after a union bound, our result would hold with probability at least \( 1 - \frac{2}{m} \). This completes the proof. \( \square \)

### A.3 Spectral norms of \( \frac{\partial \alpha^{(l)}}{\partial w^{(l)}} \) and \( \frac{\partial \alpha^{(l)}}{\partial \alpha^{(l-1)}} \)

Recall that in our setup, the layerwise outputs and pre-activations are respectively given by:

\[
\alpha^{(l)} = \phi \left( \tilde{\alpha}^{(l)} \right) , \quad \tilde{\alpha}^{(l)} := \frac{1}{\sqrt{m_l-1}} W^{(l)} \alpha^{(l-1)} . \tag{24}
\]

**Lemma A.3.** Consider any \( l \in \{2, \ldots, L\} \). Under Assumptions \( I \) and \( 2 \) for \( \theta \in B_{Spec}^{\theta_0} (\theta_0) \), with probability at least \( (1 - \frac{2}{m}) \),

\[
\left\| \frac{\partial \alpha^{(l)}}{\partial \alpha^{(l-1)}} \right\|_2^2 \leq \left( \sigma_1 + \frac{\rho}{\sqrt{m}} \right)^2 = \gamma^2 . \tag{25}
\]

Published as a conference paper at ICLR 2023
\textbf{Proof.} By definition, we have
\begin{equation}
\frac{\partial \alpha^{(l)}}{\partial \alpha^{(l-1)}} \bigg|_{i,j} = \frac{1}{\sqrt{m}} \phi'(\hat{\alpha}_i^{(l)}) W_{ij}^{(l)} . \tag{26}
\end{equation}
Since \(\|A\|_2 = \sup_{\|v\|_2=1} \|Av\|_2\), so that \(\|A\|_2^2 = \sup_{\|v\|_2=1} \sum_i (a_i, v)^2\), we have that for \(2 \leq l \leq L\),
\begin{equation}
\left\| \frac{\partial \alpha^{(l)}}{\partial \alpha^{(l-1)}} \right\|_2^2 = \sup_{\|v\|_2=1} \frac{1}{m} \sum_{i=1}^{m} \left( \phi'(\hat{\alpha}_i^{(l)}) \sum_{j=1}^{m} W_{ij}^{(l)} v_j \right)^2
\leq (a) \sup_{\|v\|_2=1} \frac{1}{m} \|W^{(l)}v\|_2^2
= \frac{1}{m} \|W^{(l)}\|_2^2
(b) \leq \gamma^2 ,
\end{equation}
where (a) follows from \(\phi\) being 1-Lipschitz by Assumption 1 and (b) from Proposition A.1. This completes the proof.

\textbf{Lemma A.4.} Consider any \(l \in [L]\). Under Assumptions 1 and 2 for \(\theta \in B_{\rho_0,\rho_1}^{\text{Spec}}(\theta_0)\), with probability at least \(1 - \frac{\eta}{m}\),
\begin{equation}
\left\| \frac{\partial \alpha^{(l)}}{\partial w(0)} \right\|_2^2 \leq \frac{1}{m} \left[ \sqrt{m} \left( \sigma_1 + \frac{\rho}{\sqrt{m}} \right)^{l-1} + \sqrt{m} \sum_{i=1}^{l-1} \left( \sigma_1 + \frac{\rho}{\sqrt{m}} \right)^{i-1} |\phi(0)| \right]^2
= \left( \gamma^{l-1} + |\phi(0)| \sum_{i=1}^{l-1} \gamma^{i-1} \right)^2 . \tag{27}
\end{equation}

\textbf{Proof.} Note that the parameter vector \(w^{(l)} = \text{vec}(W^{(l)})\) and can be indexed with \(j \in [d]\) when \(l = 1\) and \(j' \in [m]\) when \(l \geq 2\). Then, we have
\begin{equation}
\left[ \frac{\partial \alpha^{(l)}}{\partial w(l)} \right]_{i,j,j'} = \left[ \frac{\partial \alpha^{(l)}}{\partial W(l)} \right]_{i,j,j'} = \frac{1}{\sqrt{m}} \phi'(\hat{\alpha}_i^{(l)}) a_{j,j'}^{(l-1)} \mathbb{1}_{i=j} . \tag{28}
\end{equation}
For \(l \in \{2, \ldots, L\}\), noting that \(\frac{\partial \alpha^{(l)}}{\partial w^{(l)}} \in \mathbb{R}^{m \times m^2}\) and \(\|V\|_F = \|\text{vec}(V)\|_2\) for any matrix \(V\), we have
\begin{equation}
\left\| \frac{\partial \alpha^{(l)}}{\partial w(l)} \right\|_2^2 = \sup_{\|V\|_F=1} \frac{1}{m} \sum_{i=1}^{m} \left( \phi'(\hat{\alpha}_i^{(l)}) \sum_{j,j'=1}^{m} a_{j,j'}^{(l-1)} \mathbb{1}_{i=j} V_{jj'} \right)^2
\leq \frac{1}{m} \sup_{\|V\|_F=1} \left\| V \alpha^{(l-1)} \right\|_2^2
\leq \frac{1}{m} \sup_{\|V\|_F=1} \|V\|_2^2 \|\alpha^{(l-1)}\|_2^2
\leq (a) \frac{1}{m} \|\alpha^{(l-1)}\|_2^2
\leq (b) \frac{1}{m} \left[ \sqrt{m} \left( \sigma_1 + \frac{\rho}{\sqrt{m}} \right)^{l-1} + \sqrt{m} \sum_{i=1}^{l-1} \left( \sigma_1 + \frac{\rho}{\sqrt{m}} \right)^{i-1} |\phi(0)| \right]^2
= \left( \gamma^{l-1} + |\phi(0)| \sum_{i=1}^{l-1} \gamma^{i-1} \right)^2 .
\end{equation}
where (a) follows from \( \|V\|^2_2 \leq \|V\|_{F}^2 \) for any matrix \( V \), and (b) from Lemma A.2.

The \( l = 1 \) case follows in a similar manner:

\[
\left\| \frac{\partial \alpha^{(l)}}{\partial W^{(l)}} \right\|_{2,2,1}^2 \leq \frac{1}{d} \|\alpha^{(0)}\|_2^2 = \frac{1}{d} \|x\|_2^2 = 1
\]

which satisfies the form for \( l = 1 \). That completes the proof. \( \square \)

### A.4 (2, 2, 1)-NORMS OF ORDER 3 TENSORS

**Lemma A.5.** Under Assumptions [1] and [2] for \( \theta \in B_{\text{spec}}^{2}\rho (\theta_0) \), each of the following inequalities hold with probability at least \( (1 - \frac{2}{m}) \),

\[
\left\| \frac{\partial^2 \alpha^{(l)}}{(\partial \alpha^{(l-1)})^2} \right\|_{2,2,1} \leq \beta_\phi \gamma^2, \tag{29}
\]

\[
\left\| \frac{\partial^2 \alpha^{(l)}}{(\partial \alpha^{(l-1)} \partial W^{(l)})^2} \right\|_{2,2,1} \leq \beta_\phi \left( \gamma^{l-1} + |\phi(0)| \sum_{i=1}^{l-1} \gamma^{i-1} \right)^2 + 1, \tag{30}
\]

for \( l = 2, \ldots, L \); and

\[
\left\| \frac{\partial^2 \alpha^{(l)}}{(\partial W^{(l)})^2} \right\|_{2,2,1} \leq \beta_\phi \left( \gamma^{l-1} + |\phi(0)| \sum_{i=1}^{l-1} \gamma^{i-1} \right)^2, \tag{31}
\]

for \( l \in [L] \).

**Proof.** For the inequality (29), note that from (26) we obtain

\[
\frac{1}{m} \phi''(\hat{\alpha}_i^{(l)}) W_{ik}^{(l)} W_{lj}^{(l)},
\]

and so

\[
\left\| \frac{\partial^2 \alpha^{(l)}}{(\partial \alpha^{(l-1)})^2} \right\|_{2,2,1} = \sup_{\|v_1\|_2 = \|v_2\|_2 = 1} \frac{1}{m} \sum_{i=1}^{m} \phi''(\hat{\alpha}_i^{(l)}) (W^{(l)} v_1)_i (W^{(l)} v_2)_i
\]

\[
\leq \sup_{\|v_1\|_2 = \|v_2\|_2 = 1} \frac{1}{m} \beta_\phi \sum_{i=1}^{m} (W^{(l)} v_1)_i (W^{(l)} v_2)_i
\]

\[
\leq (a) \sup_{\|v_1\|_2 = \|v_2\|_2 = 1} \frac{1}{2m} \beta_\phi \sum_{i=1}^{m} (W^{(l)} v_1)_i^2 + (W^{(l)} v_2)_i^2
\]

\[
\leq \frac{1}{2m} \beta_\phi (\|W^{(l)} v_1\|_2^2 + \|W^{(l)} v_2\|_2^2)
\]

\[
\leq \frac{1}{2m} \beta_\phi (\|W^{(l)}\|_2^2 + \|W^{(l)}\|_2^2)
\]

\[
\leq \beta_\phi (\|W^{(l)}\|_2^2 + \|W^{(l)}\|_2^2)
\]

\[
\leq \beta_\phi (\sigma_1 + \rho / \sqrt{m})^2 = \beta_\phi \gamma^2,
\]

where (a) follows from \( 2ab \leq a^2 + b^2 \) for \( a, b \in \mathbb{R} \), and (b) from Proposition A.1 with probability at least \( 1 - \frac{2}{m} \).

For the inequality (30), carefully following the chain rule in (28) we obtain

\[
\left( \frac{\partial^2 \alpha^{(l)}}{(\partial \alpha^{(l-1)} \partial W^{(l)})} \right)_{i,j',k} = \frac{1}{m} \phi''(\hat{\alpha}_i^{(l)}) W_{ik}^{(l)} \alpha_{j'-1}^{(l)} \mathbb{1}_{[j'=i]} + \frac{1}{\sqrt{m}} \phi'(\hat{\alpha}_i^{(l)}) \mathbb{1}_{[i=j]} \mathbb{1}_{[j'=k]}.
\]

16
Then, we have

\[ \left\| \frac{\partial^{2} \alpha^{(l)}}{\partial \alpha^{(l-1)} \partial W^{(l)}} \right\|_{2,1,1} \]

\[ = \sup_{\|v_1\|_2 = \|v_2\|_2} \left( \sum_{i=1}^{m} \left| \sum_{k=1}^{m} \sum_{j=1}^{m} \sum_{j'=1}^{m} \left( \frac{1}{m} \phi''(\tilde{\alpha}_i^{(l)}) W_{ik}^{(l)} \alpha_{j'}^{(l-1)} I[j=j'] \right) + \frac{1}{\sqrt{m}} \phi'(\tilde{\alpha}_i^{(l)}) I[i=j] I[j'=k] \right) v_{1,k} v_{2,j'} \right| \]

\[ = \sup_{\|v_1\|_2 = \|v_2\|_2} \left( \sum_{k=1}^{m} \left| \sum_{j'=1}^{m} \phi''(\tilde{\alpha}_i^{(l)}) \alpha_{j'}^{(l-1)} V_{2,j'} \left( \sum_{k=1}^{m} W_{ik}^{(l)} v_{1,k} \right) + \frac{1}{\sqrt{m}} \sum_{k=1}^{m} \phi'(\tilde{\alpha}_i^{(l)}) v_{1,k} v_{2,i,k} \right| \right) \]

\[ \leq \sup_{\|v_1\|_2 = \|v_2\|_2} \frac{1}{m} \beta_{\phi} \left( \sum_{i=1}^{m} \left| (W^{(l)} v_1)_{i} (V_2 \alpha^{(l-1)})_{i} \right| \right) + \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \sum_{k=1}^{m} \|v_{1,k} v_{2,i,k}\| \]

\[ \leq \sup_{\|v_1\|_2 = \|v_2\|_2} \frac{1}{m} \beta_{\phi} \left( \sum_{i=1}^{m} (W^{(l)} v_1)^{2} + (V_2 \alpha^{(l-1)})^{2} \right) + \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \|v_{2,i}\|_2 \]

\[ \leq \frac{1}{2m} \beta_{\phi} \left( \|W^{(l)}\|_2^{2} + \|\alpha^{(l-1)}\|_2^{2} \right) + \|V_2\|_F \]

\[ \leq \frac{\beta_{\phi}}{2} \left( \gamma^{2} + \left( \gamma^{l-1} + |\phi(0)| \sum_{i=1}^{l-1} \gamma^{i-1} \right)^{2} \right) + 1 \]

where (a) follows from \( \|V_2 \alpha^{(l-1)}\|_2 \leq \|V_2\|_2 \|\alpha^{l-1}\| \leq \|V_2\|_F \|\alpha^{l-1}\|_2 = \|\alpha^{l-1}\|_2 \) and \( \sum_{i=1}^{m} \|V_{2,i}\|_2 \leq \sqrt{m} \sqrt{\sum_{i=1}^{m} \|V_{2,i}\|_2^2} \), and (b) follows from Proposition A.1 and Lemma A.2, with altogether holds with probability at least \( 1 - \frac{2l}{m} \).

For the last inequality [31], we start with the analysis for \( l \geq 2 \). Carefully following the chain rule in [28] we obtain

\[ \left( \frac{\partial^{2} \alpha^{(l)}}{\partial W^{(l)}^{2}} \right)_{i,j,j',k,k'} = \frac{1}{m} \phi''(\tilde{\alpha}_i^{(l)}) \alpha_{j'}^{(l-1)} \alpha_{j'}^{(l-1)} I[j=j] I[j'=k]. \]
Then, we have
\[
\left\| \frac{\partial^2 \alpha}{(\partial W)^2} \right\|_{2,1,1} \leq \frac{2m}{m} \sum_{i=1}^{m} \left| \sum_{j',k'} \left( \frac{1}{m} \phi''(\tilde{\alpha}_i^{(l)}) \alpha_{k'}^{(l-1)} \alpha_{j'}^{(l-1)} \mathbb{1}_{j'=j} \mathbb{1}_{k'=k} V_{1,j',k} V_{2,k'} \right) \right|
\]
\[
= \frac{1}{m} \beta \sum_{i=1}^{m} \left( V_1 \alpha^{(l-1)}_1 (V_2 \alpha^{(l-1)}) \right)
\]
\[
\leq \frac{1}{2m} \beta \left( \left\| V_2 \alpha^{(l-1)} \right\|_2^2 + \left\| V_2 \alpha^{(l-1)} \right\|_2^2 \right)
\]
\[
= \frac{1}{2m} \beta \left( \left( \gamma^l - 1 + |\phi(0)| \sum_{i=0}^{l-1} \gamma^i \right)^2 \right)
\]
which holds with probability at least \(1 - \frac{2^{l-1}}{m}\). For the case \(l = 1\), it is easy to show that \(\left( \frac{\partial^2 \alpha}{(\partial W)^2} \right)_{i,j',k,k'} = \frac{1}{d} \phi''(\tilde{\alpha}_i^{(1)}) x_{i,j} x_{j'} x_{k} x_{k'} \mathbb{1}_{j'=j} \mathbb{1}_{k'=k} \mathbb{1}_{i=0} = 1\) and so \(\left\| \frac{\partial^2 \alpha}{(\partial W)^2} \right\|_{2,1,1} \leq \beta \). This completes the proof. \(\square\)

A.5 \(L_\infty\) Norm of \(\frac{\partial f}{\partial \alpha^{(l)}}\)

Let \(b^{(l)} := \frac{\partial f}{\partial \alpha^{(l)}} \in \mathbb{R}^m\) for any \(l \in [L]\). Let \(b^{(l)}_0\) denote \(b^{(l)}\) at initialization. By a direct calculation, we have
\[
b^{(l)} = \frac{\partial f}{\partial \alpha^{(l)}} = \left( \prod_{l'=l+1}^L \frac{\partial \alpha^{(l)}}{\partial \alpha^{(l-1)}} \right) \frac{\partial f}{\partial \alpha^{(L)}}
\]
\[
= \left( \prod_{l'=l+1}^L \frac{1}{\sqrt{m}} (W^{(l')})^\top D^{(l')} \right) \frac{1}{\sqrt{m}} v,
\]
where \(D^{(l')}\) is a diagonal matrix of the gradient of activations, i.e., \(D^{(l')}_{ii} = \phi'(\tilde{\alpha}_i^{(l)})\). Note that we also have the following recursion:
\[
b^{(l)} = \frac{\partial f}{\partial \alpha^{(l)}} = \frac{\partial \alpha^{(l+1)}}{\partial \alpha^{(l)}} \frac{\partial f}{\partial \alpha^{(l+1)}}
\]
\[
= \frac{1}{\sqrt{m}} (W^{(l+1)})^\top D^{(l+1)} b^{(l+1)}.
\]

Lemma A.6. Consider any \(l \in [L]\). Under Assumptions 1 and 2 for \(\theta \in B^\mathrm{Spec}_{\rho_0,\rho_1}(\theta_0)\), with probability at least \(1 - \frac{2^{L+1}}{m}\),
\[
\left\| b^{(l)} \right\|_2 \leq \frac{1}{\sqrt{m}} \left( \sigma_1 + \frac{\rho}{\sqrt{m}} \right)^{L-l} (1 + \rho_1)
\]
and
\[
\left\| b^{(l)}_0 \right\|_2 \leq \frac{\sigma_1^{L-l}}{\sqrt{m}} \leq \frac{\gamma^{L-l}}{\sqrt{m}}.
\]
Proof. First, note that \( \|b^{(L)}\|_2 = \frac{1}{\sqrt{m}} \|v\|_2 \leq \frac{1}{\sqrt{m}}(\|v_0\|_2 + \|v - v_0\|_2) \leq \frac{1}{\sqrt{m}}(1 + \rho_1) \), where the inequality follows from from Proposition A.1. Now, for the inductive step, assume \( \|b^{(l)}\|_2 \leq \left( \sigma_1 + \frac{\rho}{\sqrt{m}} \right)^{L-l} \frac{1}{\sqrt{m}}(1 + \rho_1) \) with probability at least \( 1 - \frac{2}{m} \). Then, for any \( l \in [L] \),

\[
\|b^{(l)}\|_2 = \left\| \frac{\partial \alpha^{(l)}}{\partial \alpha^{(l-1)}} b^{(l)} \right\|_2 \leq \left\| \frac{\partial \alpha^{(l)}}{\partial \alpha^{(l-1)}} \right\|_2 \|b^{(l)}\|_2 \leq \left( \sigma_1 + \frac{\rho}{\sqrt{m}} \right)^{L-l} \frac{1}{\sqrt{m}}(1 + \rho_1)
\]

where the last inequality follows from Lemma A.3, with probability at least \( 1 - \frac{2}{m} \). Since we use Proposition A.1 once at layer \( L \) and then Lemma A.3 \((L - l)\) times at layer \( l \), then we have that everything holds altogether with probability at least \( 1 - \frac{2}{m} \). We have finished the proof by induction.

**Lemma A.7.** Consider any \( l \in [L] \). Under Assumptions 1 and 2 for \( \theta \in B^{\text{Spec}}(\theta_0) \), with probability at least \( 1 - \frac{2(L-l)}{m} \),

\[
\|b^{(l)}\|_\infty \leq \frac{\gamma^{L-l}}{\sqrt{m}}(1 + \rho_1).
\]

**Proof.** For any \( l \in [L] \), by definition \( i \)-th component of \( b^{(l)} \), i.e., \( b^{(l)}_i \), takes the form

\[
b^{(l)}_i = \frac{\partial \alpha^{(L)}}{\partial \alpha^{(l-1)}} f = \frac{\partial \alpha^{(L)}}{\partial \alpha^{(i)}} = \frac{\partial \alpha^{(L)}}{\partial \alpha^{(i)}} \frac{1}{\sqrt{m}}v.
\]

Then, with \( W^{(l)}_{i;\cdot} \) denoting the \( i \)-th column of the matrix \( W^{(l)} \),

\[
\left\| \frac{\partial \alpha^{(L)}}{\partial \alpha^{(l)}} \right\|_2 \leq \frac{\|\phi'(\hat{\alpha}^{(l)}_i)(W^{(l)}_{i;\cdot})^T\|_2}{\sqrt{m}} \left( \prod_{l'=l+2}^L \left\| \frac{\partial \alpha^{(l')}}{\partial \alpha^{(l'-1)}} \right\|_2 \right) \leq \frac{1}{\sqrt{m}} \|W^{(l)}_{i;\cdot}\|_2 \prod_{l'=l+2}^L \left\| \frac{\partial \alpha^{(l')}}{\partial \alpha^{(l'-1)}} \right\|_2 \leq \frac{1}{\sqrt{m}} \|W^{(l)}_{i;\cdot}\|_2 \gamma^{L-l-1} \leq \gamma \gamma^{L-l-1} = \gamma^{L-l}
\]

where (a) follows from \( \frac{\partial \alpha^{(l+1)}_i}{\partial \alpha^{(l)}_i} = \frac{1}{\sqrt{m}} \phi'(\hat{\alpha}^{(l)}_i)(W^{(l)}_{i;\cdot})^T \), (b) from \( \phi \) being 1-Lipschitz, (c) from Lemma A.3, and (d) from \( \|W^{(l)}_{i;\cdot}\|_2 \leq \|W^{(l)}\|_2 \) and Proposition A.1, which altogether holds with probability \( 1 - \frac{2}{m} \).

Therefore, for every \( i \in [m] \),

\[
\|b^{(l)}_i\| \leq \frac{1}{\sqrt{m}} \left( \sigma_1 + \frac{\rho}{\sqrt{m}} \right)^{L-l} \frac{1}{\sqrt{m}} \|v\|_2 \leq \frac{1}{\sqrt{m}} \left( \sigma_1 + \frac{\rho}{\sqrt{m}} \right)^{L-l} \frac{1}{\sqrt{m}}(1 + \rho_1) \]

where the last inequality follows from (36) and \( \|v\|_2 \leq \|v_0\|_2 + \|v - v_0\|_2 \leq 1 + \rho_1 \). This completes the proof.
### A.6 Useful Bounds

**Lemma 4.1 (Predictor gradient bounds).** Under Assumptions 7 and 8 for $\theta \in B_{\rho_0}^{\rho_1}(\theta_0)$, with probability at least $1 - \frac{2(L+1)}{m}$, we have

$$
\|\nabla_{\theta} f(\theta; x)\|_2 \leq \varrho \quad \text{and} \quad \|\nabla_x f(\theta; x)\|_2 \leq \frac{\gamma L}{\sqrt{m}} (1 + \rho_1),
$$

where $\varrho^2 = (h(L + 1))^2 + \frac{1}{m} (1 + \rho_1)^2 \sum_{l=1}^{L} (h(l))^2 \gamma^{2(L-l)}$, $\gamma = \sigma_1 + \frac{\rho_1}{\sqrt{m}}$, $h(l) = \gamma^{l-1} + |\phi(0)| \sum_{i=1}^{l-1} \gamma^{-i}$.

**Proof.** We first prove the bound on the gradient with respect to the weights. Using the chain rule,

$$
\frac{\partial f}{\partial w(l)} = \frac{\partial \alpha(i)}{\partial w(l)} \prod_{l' = l+1}^{L} \frac{\partial \alpha(l')}{\partial \alpha(l-1)} \frac{\partial f}{\partial \alpha(l)}
$$

and so

$$
\left\| \frac{\partial f}{\partial w(l)} \right\|_2^2 \leq \left\| \frac{\partial \alpha(l)}{\partial w(l)} \right\|_2^2 \left\| \prod_{l' = l+1}^{L} \frac{\partial \alpha(l')}{\partial \alpha(l-1)} \frac{\partial f}{\partial \alpha(l)} \right\|_2^2 \leq \left( \gamma^{l-1} + |\phi(0)| \sum_{i=1}^{l-1} \gamma^{-i} \right) \gamma^{2(L-l)} \cdot \frac{1}{m} (1 + \rho_1)^2
$$

for $l \in [L]$, where (a) follows from Lemma A.6, (b) follows from Lemma A.4. Similarly,

$$
\left\| \frac{\partial f}{\partial w(L+1)} \right\|_2^2 = \frac{1}{m} \left\| \alpha(L) \right\|_2^2 \leq \left( \gamma^{L} + |\phi(0)| \sum_{i=1}^{L} \gamma^{-i} \right)^2
$$

where we used Lemma A.2 for the inequality. Now,

$$
\|\nabla_{\theta} f\|^2 = \sum_{l=1}^{L+1} \left\| \frac{\partial f}{\partial w(l)} \right\|_2^2
$$

$$
\leq \left( \gamma^{L} + |\phi(0)| \sum_{i=1}^{L} \gamma^{-i} \right)^2 + \frac{1}{m} (1 + \rho_1)^2 \sum_{l=1}^{L} \left( \gamma^{l-1} + |\phi(0)| \sum_{i=1}^{l-1} \gamma^{-i} \right)^2 \gamma^{2(L-l)}
$$

$$
= \varrho^2,
$$

where (a) follows with probability $1 - \frac{2}{m}(L + 1)$ using a union bound from all the previously used results.

We now prove the bound on the gradient with respect to the input data. Using the chain rule,

$$
\frac{\partial f}{\partial x} = \frac{\partial f}{\partial \alpha(0)} \prod_{l=2}^{L} \frac{\partial \alpha(l')}{\partial \alpha(l-1)} \frac{\partial f}{\partial \alpha(L)}
$$

and so

$$
\left\| \frac{\partial f}{\partial x} \right\|_2 \leq \left\| \frac{\partial \alpha(1)}{\partial \alpha(0)} \right\|_2 \left\| \prod_{l' = 2}^{L} \frac{\partial \alpha(l')}{\partial \alpha(l-1)} \frac{\partial f}{\partial \alpha(L)} \right\|_2 \leq \left( \gamma \cdot \gamma^{L-1} \right) \cdot \frac{1}{\sqrt{m}} (1 + \rho_1)
$$

$$
= \gamma L \sqrt{\frac{L}{m}} (1 + \rho_1)
$$

where (a) follows from Lemma A.3 and Lemma A.6 with probability at least $1 - \frac{2L}{m}$ due to union bound. This completes the proof.
Lemma 4.2 (Loss bounds). Consider the square loss. Under Assumptions 1 and 2 for $\gamma = \sigma_1 + \frac{\rho}{\sqrt{m}}$, each of the following inequalities hold with probability at least $\left(1 - \frac{2(L+1)}{m}\right)$: $L(\theta_0) \leq c_{0,\sigma_1}$ and $L(\theta) \leq c_{p_1,\gamma}$ for $\theta \in B_{\rho,p_1}^{\text{Spec}}(\theta_0)$, where $c_{a,b} = \frac{2}{n} \sum_{i=1}^{n} y_i^2 + 2(1+a)^2 |g(b)|^2$ and $g(a) = a^L + |\phi(0)| \sum_{i=1}^{L} a^i$ for any $a, b \in \mathbb{R}$.

Proof. We start by noticing that for $\theta \in B_{\rho,p_1}^{\text{Spec}}(\theta_0)$,

$$L(\theta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(\theta; x_i))^2 \leq \frac{1}{n} \sum_{i=1}^{n} (2y_i^2 + 2|f(\theta; x_i)|^2).$$

(37)

Now, let us consider the particular case $\theta = \theta_0$ and a generic $\|x\|_2 = \sqrt{d}$. Let $\alpha_o^{(L)}$ be the layerwise output of layer $l$ at initialization. Then,

$$|f(\theta_0; x)| = \frac{1}{\sqrt{m}} v_0^T \alpha_o^{(L)}(x)$$

$$\leq \frac{1}{\sqrt{m}} \|v_0\|_2 \left\|[\alpha_o^{(L)}(x)]_2 \right\|$$

$$\leq \frac{1}{\sqrt{m}} \cdot 1 \cdot \left\|[\alpha_o^{(L)}(x)]_2 \right\|$$

$$\leq \frac{1}{\sqrt{m}} \left(\sigma_i^L + |\phi(0)| \sum_{i=1}^{L} \sigma_i^{i-1}\right) \sqrt{m}$$

$$= g(\sigma_1),$$

where (a) follows by assumption and (b) follows from following the same proof as in Lemma A.2 with the difference that we consider the weights at initialization. Now, replacing this result back in (37) we obtain $L(\theta_0) \leq c_{0,\sigma_1}$.

Now, let us consider the general case of $\theta \in B_{\rho,p_1}^{\text{Spec}}(\theta_0)$,

$$|f(\theta; x)| = \frac{1}{\sqrt{m}} v^T \alpha_o^{(L)}(x)$$

$$\leq \frac{1}{\sqrt{m}} \|v\|_2 \left\|[\alpha_o^{(L)}(x)]_2 \right\|$$

$$\leq \frac{1}{\sqrt{m}} \left(1 + \rho_1\right) \left\|[\alpha_o^{(L)}(x)]_2 \right\|$$

$$\leq \frac{1}{\sqrt{m}} \left(1 + \rho_1\right) \left(\sigma_i^L + |\phi(0)| \sum_{i=1}^{L} \sigma_i^{i-1}\right) \sqrt{m}$$

$$= (1 + \rho_1)g(\gamma),$$

where (a) follows from $\|v\|_2 \leq \|v_o\|_2 + \|v - v_0\|_2 \leq 1 + \rho_1$, and (b) follows from Lemma A.2. Now, replacing this result back in (37) we obtain $L(\theta_0) \leq c_{p_1,\gamma}$.

In either case, a union bound let us obtain the probability with which the results hold. This finishes the proof. \(\square\)

Corollary 4.1 (Loss gradient bound). Consider the square loss. Under Assumptions 1 and 2 for $\theta \in B_{\rho,p_1}^{\text{Spec}}(\theta_0)$, with probability at least $\left(1 - \frac{2(L+1)}{m}\right)$, we have $\|\nabla_\theta L(\theta)\|_2 \leq 2 \sqrt{L(\theta)} \varrho \leq 2 \sqrt{c_{p_1,\gamma}} \varrho$, with $\varrho$ as in Lemma 4.1 and $c_{p_1,\gamma}$ as in Lemma 4.2.

Proof. We have that $\|\nabla_\theta L(\theta)\|_2 = \|\frac{1}{n} \sum_{i=1}^{n} \ell'_i \nabla_\theta f\|_2 \leq \frac{1}{n} \sum_{i=1}^{n} \|\ell'_i\| \|\nabla_\theta f\|_2 \leq \frac{2\rho}{n} \sum_{i=1}^{n} y_i - \hat{y}_i \leq 2\rho \sqrt{L(\theta)} \leq 2 \sqrt{c_{p_1,\gamma}} \varrho$ where (a) follows from Lemma 4.1 and (b) from Lemma 4.2. \(\square\)
We establish the results from Section 5 in this appendix.

A difference between the neural network used in our work and the one in Liu et al. (2020) is that we normalize the norm of the last layer’s weight at initialization, whereas Liu et al. (2020) does not. However, if we did not normalize our last layer, then our result on the Hessian spectral norm bound would still hold with $O$ instead of $O$; consequently, our comparison with Liu et al. (2020) on the dependence on the network’s depth $L$ (our polynomial dependence against their exponential dependence) would still hold as stated in Remark 4.2.

B Restricted Strong Convexity

We establish the results from Section 5 in this appendix.

B.1 Restricted Strong Convexity and Smoothness

Theorem 5.1 (RSC for Square Loss). For square loss, under Assumptions 2 and 3 with probability at least $1 - 2 (L+1) / m$, $\forall \theta' \in Q^n \cap B^\text{Spec}_{\rho,1}(\theta_0) \cap B^\text{Euc}_{\rho_2}(\theta_1)$ with $\theta_1 \in B^\text{Spec}_{\rho,1}(\theta_0)$,

$$\mathcal{L}(\theta') \geq \mathcal{L}(\theta_t) + \langle \theta' - \theta_t, \nabla_\theta \mathcal{L}(\theta_t) \rangle + \frac{\alpha_t}{2} \| \theta' - \theta_t \|^2,$$

where $\bar{g}_t = \frac{1}{n} \sum_{i=1}^n \nabla \theta f(\theta_i; x_i)$, $c_1 = 2\kappa^2$ and $c_2 = 2c_H(2\rho_2 + \sqrt{2\rho_2 \gamma})$, with $c_H$ as in Theorem 4.2, and $c_{\rho_2 \gamma}$ as in Lemma 4.2. Consequently, $\mathcal{L}$ satisfies RSC w.r.t. $Q^n \cap B^\text{Spec}_{\rho,1}(\theta_0) \cap B^\text{Euc}_{\rho_2}(\theta_1)$ whenever $\alpha_t > 0$.

Proof. For any $\theta' \in Q^n \cap B^\text{Euc}_{\rho_2}(\theta_0)$, by the second order Taylor expansion around $\theta_t$, we have

$$\mathcal{L}(\theta') = \mathcal{L}(\theta_t) + \langle \theta' - \theta_t, \nabla_\theta \mathcal{L}(\theta_t) \rangle + \frac{1}{2} (\theta' - \theta_t)^\top \frac{\partial^2 \mathcal{L}(\theta_t)}{\partial \theta^2} (\theta' - \theta_t),$$

where $\tilde{\theta}_t = \xi \theta' + (1 - \xi) \theta_t$ for some $\xi \in [0, 1]$. We note that $\tilde{\theta}_t \in B^\text{Spec}_{\rho,1}(\theta_0)$ since,

\begin{itemize}
  \item $\| \tilde{W}_l^{(i)} - W_0^{(i)} \|_2 = \| \xi W'^{(i)} - \xi W^{(i)} + (1 - \xi) W(t) - (1 - \xi) W_0^{(i)} \|_2 \leq \xi \| W'^{(i)} - W_0^{(i)} \|^2 + (1 - \xi) \| W_l^{(i)} - W_0^{(i)} \|_2 \leq \rho$, for any $l \in [L]$, where the last inequality follows from our assumption $\theta', \theta_t \in B^\text{Spec}_{\rho,1}(\theta_0)$;
  \item $\| \tilde{W}_l^{(L+1)} - W_0^{(L+1)} \|_2 = \| \tilde{v} - v_0 \|_2 \leq \rho_1$, by following a similar derivation as in the previous point.
\end{itemize}

Focusing on the quadratic form in the Taylor expansion and recalling the form of the Hessian, we get

$$(\theta' - \theta_t)^\top \frac{\partial^2 \mathcal{L}(\tilde{\theta}_t)}{\partial \theta^2} (\theta' - \theta_t) = (\theta' - \theta_t)^\top \frac{1}{n} \sum_{i=1}^n \left[ \ell_i' \frac{\partial f(\tilde{\theta}_t; x_i)}{\partial \theta} \frac{\partial f(\tilde{\theta}_t; x_i)}{\partial \theta} + \ell_i' \frac{\partial^2 f(\tilde{\theta}_t; x_i)}{\partial \theta^2} \right] (\theta' - \theta_t)$$

$$= \frac{1}{n} \sum_{i=1}^n \left( \theta' - \theta_t, \frac{\partial f(\tilde{\theta}_t; x_i)}{\partial \theta} \right)^2 + \frac{1}{n} \sum_{i=1}^n \left( \ell_i' \theta' - \theta_t \right)^2 \frac{\partial^2 f(\tilde{\theta}_t; x_i)}{\partial \theta^2} \left( \theta' - \theta_t \right),$$
where $\ell_i = \ell(y_i, f(\tilde{\theta}_i, x_i))$, $\ell'_i = \frac{\partial \ell(y_i, z)}{\partial z} \bigg|_{z=f(\theta, x_i)}$, and $\ell''_i = \frac{\partial^2 \ell(y_i, z)}{\partial z^2} \bigg|_{z=f(\theta, x_i)}$. Now, note that

$$I_1 = \frac{1}{n} \sum_{i=1}^{n} \ell''_i \left( \left\langle \theta' - \theta, \frac{\partial f(\tilde{\theta}; x_i)}{\partial \theta} \right\rangle \right)^2$$

$$\geq \frac{2}{n} \sum_{i=1}^{n} \left( \theta' - \theta, \frac{\partial f(\tilde{\theta}; x_i)}{\partial \theta} \right)^2 \left( \frac{\partial f(\tilde{\theta}; x_i)}{\partial \theta} - \frac{\partial f(\tilde{\theta}; x_i)}{\partial \theta} \right)^2$$

$$= \frac{2}{n} \sum_{i=1}^{n} \left( \theta' - \theta, \frac{\partial f(\tilde{\theta}; x_i)}{\partial \theta} \right)^2 + \frac{2}{n} \sum_{i=1}^{n} \left( \theta' - \theta, \frac{\partial f(\tilde{\theta}; x_i)}{\partial \theta} - \frac{\partial f(\tilde{\theta}; x_i)}{\partial \theta} \right)^2$$

$$\quad + \frac{4}{n} \sum_{i=1}^{n} \left( \theta' - \theta, \frac{\partial f(\tilde{\theta}; x_i)}{\partial \theta} \right) \left( \theta' - \theta, \frac{\partial f(\tilde{\theta}; x_i)}{\partial \theta} - \frac{\partial f(\tilde{\theta}; x_i)}{\partial \theta} \right)$$

$$\geq \frac{2}{n} \sum_{i=1}^{n} \left( \theta' - \theta, \frac{\partial f(\tilde{\theta}; x_i)}{\partial \theta} \right)^2 - \frac{4}{n} \sum_{i=1}^{n} \left\| \frac{\partial f(\tilde{\theta}; x_i)}{\partial \theta} \right\|_2 \left\| \frac{\partial f(\tilde{\theta}; x_i)}{\partial \theta} - \frac{\partial f(\tilde{\theta}; x_i)}{\partial \theta} \right\|_2$$

$$\times \left\| \theta' - \theta \right\|_2^2$$

$$\geq \frac{2}{n} \left( \theta' - \theta, \frac{1}{n} \sum_{i=1}^{n} \frac{\partial f(\tilde{\theta}; x_i)}{\partial \theta} \right)^2 - \frac{4}{n} \sum_{i=1}^{n} \frac{c_H}{\sqrt{m}} \left\| \tilde{\theta} - \theta \right\|_2 \left\| \theta' - \theta \right\|_2^2$$

$$\geq \frac{2}{n} \left( \theta' - \theta, \frac{1}{n} \sum_{i=1}^{n} \frac{\partial f(\tilde{\theta}; x_i)}{\partial \theta} \right)^2 - \frac{4}{n} \sum_{i=1}^{n} \frac{c_H}{\sqrt{m}} \left\| \theta' - \theta \right\|_2^3$$

$$= \frac{2\kappa^2}{n} \left[ \frac{1}{n} \sum_{i=1}^{n} \left( \frac{\partial f(\tilde{\theta}; x_i)}{\partial \theta} \right)^2 \right] \left\| \theta' - \theta \right\|_2^2 - \frac{4c_H}{\sqrt{m}} \left\| \theta' - \theta \right\|_2^3$$

where (a) follows by Cauchy-Schwartz inequality; (b) follows by Jensen’s inequality (first term) and the use of Theorem 4.1 and Lemma 4.1 due to $\tilde{\theta}_i \in B_{p,p_1}(\theta_0)$; (c) follows from $\left\| \tilde{\theta}_i - \theta \right\|_2 = \left\| \xi \theta' + (1 - \xi) \theta_i - \theta \right\|_2 \leq \left( \left\| \theta' - \theta \right\|_2 \right) \left\| \theta_i - \theta \right\|_2$; (d) follows since $\theta' \in Q^*$ and from the fact that $p^t q = \cos(p, q) \|p\| \|q\|$ for any vectors $p, q$.

For analyzing $I_2$, first note that for square loss, $\ell'_i = 2(y_i, t - y_i) + (y_i, t - y_i)$, so that for the vector $[\ell'_i]$, we have $\frac{1}{n} \left\| [\ell'_i] \right\|_2^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i, t - y_i)^2 = 4L(\theta_i)$. Further, with $Q_{t,i} = (\theta' - \theta_i)^T \frac{\partial^2 f(\tilde{\theta}; x_i)}{\partial \theta^2}(\theta' - \theta_i)$, we have

$$|Q_{t,i}| = \left( (\theta' - \theta_i)^T \frac{\partial^2 f(\tilde{\theta}; x_i)}{\partial \theta^2}(\theta' - \theta_i) \right) \leq \left\| \theta' - \theta \right\|_2^2 \left\| \frac{\partial^2 f(\tilde{\theta}; x_i)}{\partial \theta^2} \right\|_2 \leq \frac{c_H}{\sqrt{m}} \left\| \theta' - \theta_i \right\|_2^2.$$

Now, note that

$$I_2 = \frac{1}{n} \sum_{i=1}^{n} \ell'_i Q_{t,i}$$

$$\geq - \left( \frac{1}{n} \sum_{i=1}^{n} \ell'_i \right) \left( \frac{1}{n} \sum_{i=1}^{n} Q_{t,i} \right)^{1/2}$$

$$\geq - \frac{1}{n} \left\| [\ell'_i] \right\|_2^{1/2} \left( \frac{1}{n} \sum_{i=1}^{n} Q_{t,i}^2 \right)^{1/2}$$

$$\geq - 2 \sqrt{L(\theta_i) c_H \left\| \theta' - \theta_i \right\|_2^2 / \sqrt{m}},$$
where (a) follows by Cauchy-Schwartz inequality. Putting the lower bounds on $I_1$ and $I_2$ back, with $\tilde{g}_t = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial f(\theta_i, \bar{x})}{\partial \theta}$, we have

$$
(\theta' - \theta_t) \top \frac{\partial^2 \bar{L}(\tilde{\theta}_t)}{\partial \theta^2} (\theta' - \theta_t) \geq 2\kappa^2 \| \tilde{g}_t \|_2^2 - \frac{4\varphi c_H \| \theta' - \theta_t \|_2 + 2c_H \sqrt{\bar{L}(\tilde{\theta}_t)}}{\sqrt{m}} \| \theta' - \theta_t \|_2^2.
$$

Now, since $\theta' \in B^E_{p_2}(\theta_t)$, $\| \theta' - \theta_t \|_2 \leq \rho_2$, so we have

$$
(\theta' - \theta_t) \top \frac{\partial^2 \bar{L}(\tilde{\theta}_t)}{\partial \theta^2} (\theta' - \theta_t) \geq 2\kappa^2 \| \tilde{g}_t \|_2^2 - \frac{4\varphi c_H \rho_2 + 2c_H \sqrt{\bar{L}(\tilde{\theta}_t)}}{\sqrt{m}} \| \theta' - \theta_t \|_2^2
$$

where the last inequality follows from Lemma 4.2. That completes the proof.

Next, we state and prove the RSC result for general losses.

**Theorem B.1 (RSC of Loss).** Under Assumptions 1, 2 and 3, with probability at least $(1 - \frac{2(L+1)}{m})$, $\forall \theta' \in Q^t_n \cap B^E_{p_1} \cap B^E_{p_2}(\theta_t)$, $\mathbb{L}(\theta') \geq \mathbb{L}(\theta_t) + \langle \theta' - \theta_t, \nabla_{\theta} \mathbb{L}(\theta_t) \rangle + \frac{\alpha_t^2}{2} \| \theta' - \theta_t \|_2^2$, with $\alpha_t = c_1 \| \tilde{g}_t \|_2 - \frac{c_4 + c_{4,t}}{\sqrt{m}}$, (38)

where $\tilde{g}_t = \frac{1}{n} \sum_{i=1}^{n} \nabla_{\theta} f(\theta_t; \bar{x}_i)$, $c_1 = \alpha \kappa$, $c_4 = 2\alpha \varphi c_H \rho_2$, $c_H$ is as in Theorem 4.1 and $c_{4,t} = c_H \sqrt{\lambda_t}$ where $\lambda_t = \frac{1}{n} \sum_{i=1}^{n} (\ell'_i)^2$ with $\ell'_i = \frac{\partial f(\theta_t; \bar{x}_i)}{\partial \theta}$ and $\tilde{\theta} \in B^E_{p_1}$ being some point in the segment that joins $\theta'$ and $\theta_t$. Consequently, $\mathbb{L}$ satisfies RSC w.r.t. $(Q^t_n \cap B^E_{p_1} \cap B^E_{p_2}(\theta_t), \theta_t)$ whenever $\alpha_t > 0$.

**Proof.** For any $\theta' \in Q^t_n \cap B^E_{p_1}(\theta_0)$, by the second order Taylor expansion around $\theta_t$, we have

$$
\mathbb{L}(\theta') = \mathbb{L}(\theta_t) + \langle \theta' - \theta_t, \nabla_{\theta} \mathbb{L}(\theta_t) \rangle + \frac{1}{2} \langle \theta' - \theta_t, \frac{\partial^2 \bar{L}(\tilde{\theta}_t)}{\partial \theta^2} (\theta' - \theta_t) \rangle,
$$

where $\tilde{\theta}_t = \xi \theta' + (1 - \xi) \theta_t$ for some $\xi \in [0, 1]$. We note that $\tilde{\theta}_t \in B^E_{p_1}(\theta_0)$ since,

- $\| \tilde{W}'(t) - W'_{0}(t) \|_2 \geq \| \xi W'_{0}(t) - \xi W'_{0}(t) + (1 - \xi) W'_{t} - (1 - \xi) W'_{0}(t) \|_2 \leq \rho$, for any $t \in [L]$, where the last inequality follows from our assumption $\theta'_t, \theta_t \in B^E_{p_1}(\theta_0)$; and
- $\| \tilde{W}'(t) - W'_{0}(t) \|_2 \geq \| \bar{y} - v_{0} \|_2 \leq \rho_1$, by following a similar derivation as in the previous point.

Focusing on the quadratic form in the Taylor expansion and recalling the form of the Hessian, we get

$$
(\theta' - \theta_t) \top \frac{\partial^2 \bar{L}(\tilde{\theta}_t)}{\partial \theta^2} (\theta' - \theta_t)
$$

$$
= (\theta' - \theta_t) \top \frac{1}{n} \sum_{i=1}^{n} \left[ \ell'_i \top \frac{\partial f(\tilde{\theta}_t; \bar{x}_i)}{\partial \theta} \right] + \frac{1}{n} \sum_{i=1}^{n} \left[ \ell'_i (\theta' - \theta_t) \top (\theta' - \theta_t) \right],
$$
where \( \ell_i = \ell(y_i, f(\tilde{\theta}_i, x_i)), \ell'_i = \frac{\partial \ell(y_i, z)}{\partial z}_{z=f(\tilde{\theta}, x_i)}, \) and \( \ell''_i = \frac{\partial^2 \ell(y_i, z)}{\partial^2 z}_{z=f(\tilde{\theta}, x_i)} \). Now, note that

\[
I_1 = \frac{1}{n} \sum_{i=1}^{n} \ell''_i \left( \theta' - \theta_i, \frac{\partial f(\tilde{\theta}; x_i)}{\partial \theta} \right)^2
\]

\[
\geq \frac{2}{n} \sum_{i=1}^{n} \left( \theta' - \theta_i, \frac{\partial f(\tilde{\theta}; x_i)}{\partial \theta} + \left( \frac{\partial f(\tilde{\theta}; x_i)}{\partial \theta} - \frac{\partial f(\tilde{\theta}; x_i)}{\partial \theta} \right) \right)^2
\]

\[
\geq \frac{2}{n} \sum_{i=1}^{n} \left( \theta' - \theta_i, \frac{\partial f(\tilde{\theta}; x_i)}{\partial \theta} \right)^2 + \frac{2}{n} \sum_{i=1}^{n} \left( \theta' - \theta_i, \frac{\partial f(\tilde{\theta}; x_i)}{\partial \theta} - \frac{\partial f(\tilde{\theta}; x_i)}{\partial \theta} \right)^2
\]

\[
+ \frac{4}{n} \sum_{i=1}^{n} \left( \theta' - \theta_i, \frac{\partial f(\tilde{\theta}; x_i)}{\partial \theta} \right) \left( \theta' - \theta_i, \frac{\partial f(\tilde{\theta}; x_i)}{\partial \theta} - \frac{\partial f(\tilde{\theta}; x_i)}{\partial \theta} \right)
\]

\[
(\text{a}) \quad \geq \frac{2}{n} \sum_{i=1}^{n} \left( \theta' - \theta_i, \frac{\partial f(\tilde{\theta}; x_i)}{\partial \theta} \right)^2
\]

\[
- \frac{4}{n} \sum_{i=1}^{n} \left\| \frac{\partial f(\tilde{\theta}; x_i)}{\partial \theta} \right\|_2^2 \left\| \frac{\partial f(\tilde{\theta}; x_i)}{\partial \theta} - \frac{\partial f(\tilde{\theta}; x_i)}{\partial \theta} \right\|_2 \left\| \theta' - \theta_i \right\|_2^2
\]

\[
(\text{b}) \quad \geq a \left( \theta' - \theta_i, \frac{1}{n} \sum_{i=1}^{n} \frac{\partial f(\tilde{\theta}; x_i)}{\partial \theta} \right)^2 - \frac{4}{n} \sum_{i=1}^{n} \frac{c_H}{\sqrt{m}} \left\| \theta_i - \theta_i \right\|_2 \left\| \theta' - \theta_i \right\|_2^2
\]

\[
(\text{c}) \quad \geq a \left( \theta' - \theta_i, \frac{1}{n} \sum_{i=1}^{n} \frac{\partial f(\tilde{\theta}; x_i)}{\partial \theta} \right)^2 - \frac{2 a c_H}{\sqrt{m}} \left\| \theta' - \theta_i \right\|_2^3
\]

\[
(\text{d}) \quad \geq a \kappa^2 \left( \frac{1}{n} \sum_{i=1}^{n} \frac{\partial f(\tilde{\theta}; x_i)}{\partial \theta} \right)^2 \left\| \theta' - \theta_i \right\|_2^2 - \frac{2 a c_H}{\sqrt{m}} \left\| \theta' - \theta_i \right\|_2 \left\| \theta' - \theta_i \right\|_2^2
\]

\[
= \left( a \kappa^2 \frac{1}{n} \sum_{i=1}^{n} \frac{\partial f(\tilde{\theta}; x_i)}{\partial \theta} \right)^2 \left\| \theta' - \theta_i \right\|_2^2 - \frac{2 a c_H \left\| \theta' - \theta_i \right\|_2 \left\| \theta' - \theta_i \right\|_2^2}{\sqrt{m}}
\]

where (a) follows by Cauchy-Schwartz inequality; (b) follows by Jensen’s inequality (first term) and the use of Theorem 4.1 and Lemma 4.1 due to \( \hat{\theta}_t \in B_{p, p_0}(\theta_0) \); (c) follows from \( \left\| \hat{\theta}_t - \theta_i \right\|_2 = \left\| \xi \theta' + (1 - \xi) \theta_t - \theta_i \right\|_2 \leq \left\| \theta' - \theta_i \right\|_2 \); (d) follows since \( \theta' \in Q^t \) and from the fact that \( p^\top q = \cos(p, q) \|p\| \|q\| \) for any vectors \( p, q \).

For analyzing \( I_2 \), let \( \lambda_t := \frac{1}{n} \sum_{i=1}^{n} (\ell'_i)^2 \). As before, with \( Q_{t,i} = (\theta' - \theta_i)^\top \frac{\partial^2 f(\tilde{\theta}; x_i)}{\partial \theta^2} (\theta' - \theta_i) \), we have

\[
\left| Q_{t,i} \right| = \left( \theta' - \theta_i \right)^\top \frac{\partial^2 f(\tilde{\theta}; x_i)}{\partial \theta^2} (\theta' - \theta_i) \leq \left\| \theta' - \theta_i \right\|_2 \left\| \frac{\partial^2 f(\tilde{\theta}; x_i)}{\partial \theta^2} \right\|_2 \leq c_H \left\| \theta' - \theta_i \right\|_2^2 \frac{1}{\sqrt{m}}.
\]

Now, note that

\[
I_2 = \frac{1}{n} \sum_{i=1}^{n} \ell'_i Q_{t,i}
\]

\[
\geq - \sum_{i=1}^{n} \left( \frac{1}{\sqrt{n}} \ell'_i \right) \left( \frac{1}{\sqrt{n}} Q_{t,i} \right)
\]

\[
(\text{a}) \quad \geq - \left( \frac{1}{n} \left\| \ell'_i \right\|_2 \right)^{1/2} \left( \frac{1}{n} \sum_{i=1}^{n} Q_{t,i}^2 \right)^{1/2}
\]

\[
\geq - \sqrt{\lambda_t c_H} \left\| \theta' - \theta_i \right\|_2^2 \frac{1}{\sqrt{m}},
\]
where (a) follows by Cauchy-Schwartz inequality. Putting the lower bounds on $I_1$ and $I_2$ back, with $g_i = \frac{1}{n} \sum_{t=1}^n \frac{\partial f(\theta; x_i)}{\partial x_i}$, we have

$$
(\theta' - \theta_1) \mathbf{g} \geq \left( \alpha n^2 \|\mathbf{g}\|^2_2 - \frac{2 \alpha \rho c H \|\theta' - \theta_1\|_2 + c_H \sqrt{\lambda_1}}{\sqrt{m}} \right) \|\theta' - \theta_1\|^2_2.
$$

Now, since $\theta' \in B_{\rho_2}(\theta_1)$, $\|\theta' - \theta_1\|_2 \leq \rho_2$, so we have

$$
(\theta' - \theta_1) \mathbf{g} \geq \left( \alpha n^2 \|\mathbf{g}\|^2_2 - \frac{2 \alpha \rho c H \rho_2 + c_H \sqrt{\lambda_1}}{\sqrt{m}} \right) \|\theta' - \theta_1\|^2_2.
$$

That completes the proof. □

**Theorem 5.2 (Local Smoothness for Square Loss).** For square loss, under Assumptions 1 and 2, with probability at least $1 - 2(L+1)/m$, $\forall \theta, \theta' \in B_{\rho,\rho_1}^\text{Spec}(\theta_0)$,

$$
\mathcal{L}(\theta') \leq \mathcal{L}(\theta) + (\theta' - \theta, \nabla \mathcal{L}(\theta)) + \frac{\beta}{2} \|\theta' - \theta\|^2_2,
$$

with $\beta = 2\rho^2 + \frac{2c_H \sqrt{c_{\rho_1,\gamma}}}{\sqrt{m}}$. Consequently, $\mathcal{L}$ is locally $\beta$-smooth. Moreover, if $\gamma$ (and so $\sigma_1$ and $\rho$) is chosen according to the desirable operating regimes (see Remark 4.7) and $\rho_1 = O(\text{poly}(L))$, then $\beta = O(\text{poly}(L))$.

**Proof.** By the second order Taylor expansion about $\hat{\theta} \in B_{\rho,\rho_1}^\text{Spec}(\theta_0)$, we have $\mathcal{L}(\theta') = \mathcal{L}(\hat{\theta}) + (\theta' - \hat{\theta}, \nabla \mathcal{L}(\hat{\theta})) + \frac{1}{2} (\theta' - \hat{\theta})^T \frac{\partial^2 \mathcal{L}(\hat{\theta})}{\partial \theta^2} (\theta' - \hat{\theta})$, where $\hat{\theta} = \xi \theta' + (1 - \xi) \hat{\theta}$ for some $\xi \in [0, 1]$. Then,

$$
(\theta' - \hat{\theta}) \mathbf{g} = (\theta' - \hat{\theta}) \mathbf{g} = \left( \sum_{i=1}^n \ell_i'' \left( \theta' - \hat{\theta}, \frac{\partial f(\hat{\theta}; x_i)}{\partial \theta} \right) \right) \|\theta' - \hat{\theta}\|^2_2,
$$

where $\ell_i = \ell(y_i, f(\hat{\theta}; x_i))$, $\ell_i' = \frac{\partial \ell(y_i, z)}{\partial z} \big|_{z = f(\hat{\theta}; x_i)}$, and $\ell_i'' = \frac{\partial^2 \ell(y_i, z)}{\partial z^2} \big|_{z = f(\hat{\theta}; x_i)}$. Now, note that

$$
I_1 = \frac{1}{n} \sum_{i=1}^n \ell_i'' \left( \theta' - \hat{\theta}, \frac{\partial f(\hat{\theta}; x_i)}{\partial \theta} \right)^2 \leq \frac{2 n}{n} \sum_{i=1}^n \left( \frac{\partial f(\hat{\theta}; x_i)}{\partial \theta} \right)^2 \|\theta' - \hat{\theta}\|^2_2 \leq 2 \rho^2 \|\theta' - \hat{\theta}\|^2_2.
$$

For $I_2$, first note that for square loss, $\ell_i' = 2(\bar{y}_i - y_i)$ with $\bar{y}_i = f(\hat{\theta}; x_i)$, so that for the vector $[\ell_i']$, we have $\frac{1}{n} \|\ell_i'\|^2_2 = \frac{4}{n} \sum_{i=1}^n (\bar{y}_i - y_i)^2 = 4 \mathcal{L}(\hat{\theta})$. Further, with

$$
Q_1 = (\theta' - \hat{\theta}) \frac{\partial^2 f(\hat{\theta}; x_i)}{\partial \theta^2} (\theta' - \hat{\theta}),
$$

we have

$$
|Q_1| = \left\| (\theta' - \hat{\theta}) \frac{\partial^2 f(\hat{\theta}; x_i)}{\partial \theta^2} (\theta' - \hat{\theta}) \right\| \leq \|\theta' - \hat{\theta}\|^2_2 \left\| \frac{\partial^2 f(\hat{\theta}; x_i)}{\partial \theta^2} \right\|_2 \leq c_H \|\theta' - \hat{\theta}\|^2_2 \sqrt{m}.
$$
Then, we have
\[
I_2 = \frac{1}{n} \sum_{i=1}^{n} \ell'_i (\theta' - \bar{\theta})^{\top} \frac{\partial^2 f (\tilde{\theta}; x_i)}{\partial \theta^2} (\theta' - \bar{\theta})
\]
\[
\leq \left( \frac{1}{n} \mathbb{E} \left[ \ell'_i^2 \right] \right)^{1/2} \left( \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ Q_i^2 \right] \right)^{1/2}
\]
\[
\leq 2 \sqrt{\mathcal{L}(\bar{\theta})} \frac{c_H \| \theta' - \bar{\theta} \|_{2}^{2}}{\sqrt{n}},
\]
where (a) follows by Cauchy-Schwartz. Putting the upper bounds on \( I_1 \) and \( I_2 \) back, we have
\[
(\theta' - \bar{\theta})^{\top} \frac{\partial^2 \mathcal{L}(\bar{\theta})}{\partial \theta^2} (\theta' - \bar{\theta}) \leq \left[ 2 \varrho^2 + \frac{2c_H \sqrt{\mathcal{L}(\bar{\theta})}}{\sqrt{n}} \right] \| \theta' - \bar{\theta} \|_{2}^{2}
\]
\[
\leq \left[ 2 \varrho^2 + \frac{2c_H \sqrt{\mathcal{L}(\bar{\theta})}}{\sqrt{n}} \right] \| \theta' - \bar{\theta} \|_{2}^{2},
\]
where the last inequality follows from Lemma [4.2]. This proves the first statement of the theorem.

Now, the second statement simply follows from the fact that by choosing \( \gamma \) (and so \( \rho \) and \( \sigma_1 \)) according to the desirable operating regimes (see Remark [4.1]) and by choosing \( \rho_1 \) according to Theorem [4.1], we obtain \( c_H = O(\text{poly}(L)) \), \( \rho^2 = O(\text{poly}(L)) \) and \( c_{\rho_1, \gamma} = O(\text{poly}(L)) \). This completes the proof. \( \square \)

Next, we state and prove the smoothness result for general losses.

**Theorem B.2 (Local Smoothness of Loss).** Under Assumptions [3.7, 3.8, 3.9, and 3.10] with probability at least \( \left( 1 - \frac{2(L+1)}{m} \right) \), \( \mathcal{L}(\bar{\theta}), \bar{\theta} \in B_{p, p_1}^{\text{Spec}} (\theta_0) \), is \( \beta \)-smooth with \( \beta = \varrho^2 + \frac{\sqrt{\mathcal{L}(\bar{\theta})}}{\sqrt{n}} \) with \( \varrho \) as in Lemma [4.1] and
\[
\lambda_t = \frac{1}{n} \sum_{i=1}^{n} (\ell'_i)^2 \quad \text{with} \quad \ell'_i = \left. \frac{\partial \ell(y, z)}{\partial z} \right|_{z=f(\hat{\theta}, x_i)} \quad \text{and} \quad \bar{\theta} \in B_{p, p_1}^{\text{Spec}} \text{ being some point in the segment that joins } \theta' \text{ and } \theta \text{ as in (11)}.
\]

**Proof.** By the second order Taylor expansion about \( \bar{\theta} \), we have \( \mathcal{L}(\theta') = \mathcal{L}(\bar{\theta}) + (\theta' - \bar{\theta}, \nabla_{\theta} \mathcal{L}(\bar{\theta})) + \frac{1}{2} (\theta' - \bar{\theta})^{\top} \frac{\partial^2 \mathcal{L}(\bar{\theta})}{\partial \theta^2} (\theta' - \bar{\theta}) \), where \( \hat{\theta} = \xi \theta' + (1 - \xi) \bar{\theta} \) for some \( \xi \in [0, 1] \). Then,
\[
(\theta' - \bar{\theta})^{\top} \frac{\partial^2 \mathcal{L}(\bar{\theta})}{\partial \theta^2} (\theta' - \bar{\theta})
\]
\[
= (\theta' - \bar{\theta})^{\top} \left( \frac{1}{n} \sum_{i=1}^{n} \left[ \ell'_i \frac{\partial f (\tilde{\theta}; x_i)}{\partial \theta} \frac{\partial f (\tilde{\bar{\theta}}; x_i)}{\partial \theta}^{\top} + \ell'_i \frac{\partial^2 f (\tilde{\theta}; x_i)}{\partial \theta^2} \right] \right) (\theta' - \bar{\theta})
\]
\[
= \frac{1}{n} \sum_{i=1}^{n} \ell'_i (\theta' - \bar{\theta})^{\top} \frac{\partial f (\tilde{\theta}; x_i)}{\partial \theta} + \frac{1}{n} \sum_{i=1}^{n} \ell'_i (\theta' - \bar{\theta})^{\top} \frac{\partial^2 f (\tilde{\theta}; x_i)}{\partial \theta^2} (\theta' - \bar{\theta})
\]
where \( \ell_i = \ell(y_i, f (\tilde{\theta}, x_i)) \), \( \ell'_i = \left. \frac{\partial \ell(y, z)}{\partial z} \right|_{z=f(\hat{\theta}, x_i)} \), and \( \ell''_i = \left. \frac{\partial^2 \ell(y, z)}{\partial z^2} \right|_{z=f(\hat{\theta}, x_i)} \). Now, note that
\[
I_1 = \frac{1}{n} \sum_{i=1}^{n} \left( \theta' - \bar{\theta} \right)^{\top} \frac{\partial f (\tilde{\theta}; x_i)}{\partial \theta} \left( \theta' - \bar{\theta} \right)
\]
\[
\leq \frac{b \varrho^2 \| \theta' - \bar{\theta} \|_{2}^{2}}{\sqrt{n}}
\]
and
\[
I_2 = \frac{1}{n} \sum_{i=1}^{n} \left( \theta' - \bar{\theta} \right)^{\top} \frac{\partial^2 f (\tilde{\theta}; x_i)}{\partial \theta^2} (\theta' - \bar{\theta})
\]
\[
\leq \frac{b \varrho^2 \| \theta' - \bar{\theta} \|_{2}^{2}}{\sqrt{n}}.
\]
where (a) follows by the Cauchy-Schwartz inequality and (b) from Lemma 4.1.

For $I_2$, let $\lambda_t = \frac{1}{n} ||\tilde{c}^t_i||_2^2$. Further, with $Q_{t,i} = (\theta' - \bar{\theta})^\top \frac{\partial^2 f(\hat{\theta}; x_i)}{\partial \theta^2}(\theta' - \bar{\theta})$, we have

$$I_2 = \frac{1}{n} \sum_{i=1}^{n} \tilde{c}^t_i (\theta' - \bar{\theta})^\top \frac{\partial^2 f(\hat{\theta}; x_i)}{\partial \theta^2}(\theta' - \bar{\theta})$$

Then, we have

$$I_2 \leq \frac{1}{n} \sum_{i=1}^{n} \left( \frac{1}{\sqrt{n}} \tilde{c}^t_i \right) \left( \frac{1}{\sqrt{n}} Q_i \right) \leq \frac{1}{n} \sum_{i=1}^{n} Q_i^2 \leq \sqrt{n} \frac{c_H \|\theta' - \bar{\theta}\|^2}{\sqrt{m}}$$

where (a) follows by Cauchy-Schwartz. Putting the upper bounds on $I_1$ and $I_2$ back, we have

$$(\theta' - \bar{\theta})^\top \frac{\partial^2 \mathcal{L}(\bar{\theta})}{\partial \theta^2}(\theta' - \bar{\theta}) \leq \frac{b_0^2 + c_H \sqrt{n}}{\sqrt{m}} \|\theta' - \bar{\theta}\|^2.$$ This completes the proof. \qed

Lemma 5.1 (RSC $\Rightarrow$ RPL). Let $B_t := Q_t^1 \cap B_{\rho_1}^{\text{Spec}}(\theta_t) \cap B_{\rho_2}^{\text{Lin}}(\theta_t)$. In the setting of Theorem 5.1, if $\alpha_t > 0$, then the tuple $(B_t, \theta_t)$ satisfies the Restricted Polyak-Łojasiewicz (RPL) condition, i.e.,

$$\mathcal{L}(\theta_t) - \inf_{\theta \in B_t} \mathcal{L}(\theta) \leq \frac{1}{2\alpha_t} \|\nabla \mathcal{L}(\theta_t)\|^2_2,$$ with probability at least $(1 - \frac{2(L+1)}{m})$.

Proof. Define

$$\hat{\mathcal{L}}_{\theta_t}(\theta) := \mathcal{L}(\theta_t) + \langle \theta - \theta_t, \nabla \mathcal{L}(\theta_t) \rangle + \frac{\alpha_t}{2} \|\theta - \theta_t\|^2_2.$$ By Theorem 5.1, for all $\theta' \in B_t$, we have

$$\mathcal{L}(\theta') \geq \hat{\mathcal{L}}_{\theta_t}(\theta').$$ Further, note that $\hat{\mathcal{L}}_{\theta_t}(\theta)$ is minimized at $\hat{\theta}_{t+1} := \theta_t - \nabla \mathcal{L}(\theta_t)/\alpha_t$ and the minimum value is:

$$\inf_{\theta} \hat{\mathcal{L}}_{\theta_t}(\theta) = \hat{\mathcal{L}}_{\theta_t}(\hat{\theta}_{t+1}) = \mathcal{L}(\theta_t) - \frac{1}{2\alpha_t} \|\nabla \mathcal{L}(\theta_t)\|^2_2.$$ Then, we have

$$\inf_{\theta \in B_t} \mathcal{L}(\theta) \overset{(a)}{=} \inf_{\theta \in B_t} \hat{\mathcal{L}}_{\theta_t}(\theta) \geq \inf_{\theta} \hat{\mathcal{L}}_{\theta_t}(\theta) = \mathcal{L}(\theta_t) - \frac{1}{2\alpha_t} \|\nabla \mathcal{L}(\theta_t)\|^2_2,$$ where (a) follows from (39). Rearranging terms completes the proof. \qed

B.2 Convergence with Restricted Strong Convexity

Lemma 5.2 (Local Loss Reduction in $B_t$). Let $\alpha_t, \beta$ be as in Theorems 5.1 and 5.2, respectively, and $B_t := Q_{\rho_1} \cap B_{\rho_1}^{\text{Spec}}(\theta_t) \cap B_{\rho_2}^{\text{Lin}}(\theta_t)$. Consider Assumptions 4 and 4 and gradient descent with step size $\eta_t = \frac{b}{\beta^2}, \omega_t \in (0, 2)$. Then, for any $\tilde{\theta}_{t+1} \in \text{arginf}_{\theta \in B_t} \mathcal{L}(\theta)$, we have with probability at least $(1 - \frac{2(L+1)}{m})$,

$$\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\tilde{\theta}_{t+1}) \leq \left(1 - \frac{\alpha_t \omega_t}{\beta}(2 - \omega_t)\right) (\mathcal{L}(\theta_t) - \mathcal{L}(\tilde{\theta}_{t+1})).$$
Proof. Since $\mathcal{L}$ is $\beta$-smooth by Theorem 5.2, we have

$$\mathcal{L}(\theta_{t+1}) \leq \mathcal{L}(\theta_t) + (\theta_{t+1} - \theta_t, \nabla_{\theta} \mathcal{L}(\theta_t)) + \frac{\beta}{2} \|\theta_{t+1} - \theta_t\|_2^2$$

$$= \mathcal{L}(\theta_t) - \eta_t \|\nabla_{\theta} \mathcal{L}(\theta_t)\|_2^2 + \frac{\beta \eta_t^2}{2} \|\nabla_{\theta} \mathcal{L}(\theta_t)\|_2^2$$

$$= \mathcal{L}(\theta_t) - \eta_t \left(1 - \frac{\beta \eta_t}{2}\right) \|\nabla_{\theta} \mathcal{L}(\theta_t)\|_2^2 \tag{40}$$

Since $\bar{\theta}_{t+1} \in \arg\min_{\theta \in B_t} \mathcal{L}(\theta)$ and $\alpha_t > 0$ by assumption, from Lemma 5.1 we obtain

$$-\|\nabla_{\theta} \mathcal{L}(\theta_t)\|_2^2 \leq -2\alpha_t (\mathcal{L}(\theta_t) - \mathcal{L}(\bar{\theta}_{t+1})).$$

Hence

$$\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\bar{\theta}_{t+1}) \leq \mathcal{L}(\theta_t) - \mathcal{L}(\bar{\theta}_{t+1}) - \eta_t \left(1 - \frac{\beta \eta_t}{2}\right) \|\nabla_{\theta} \mathcal{L}(\theta_t)\|_2^2$$

$$\leq \mathcal{L}(\theta_t) - \mathcal{L}(\bar{\theta}_{t+1}) - \eta_t \left(1 - \frac{\beta \eta_t}{2}\right) 2\alpha_t (\mathcal{L}(\theta_t) - \mathcal{L}(\bar{\theta}_{t+1}))$$

$$= \left(1 - 2\alpha_t \eta_t \left(1 - \frac{\beta \eta_t}{2}\right) (\mathcal{L}(\theta_t) - \mathcal{L}(\bar{\theta}_{t+1})) \right) \mathcal{L}(\theta_t).$$

where (a) follows for any $\eta_t \leq \frac{2}{\beta}$ because this implies $1 - \frac{\beta \eta_t}{2} \geq 0$. Choosing $\eta_t = \frac{\alpha_t}{\beta}, \omega_t \in (0, 2)$,

$$\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\bar{\theta}_{t+1}) \leq \left(1 - \frac{\alpha_t \omega_t}{\beta}(2 - \omega_t)\right) (\mathcal{L}(\theta_t) - \mathcal{L}(\bar{\theta}_{t+1})). \tag{14}$$

This completes the proof. $\square$

Theorem 5.3 (Global Loss Reduction in $B^\text{Spec}(\theta_0)$). Let $\alpha_t, \beta$ be as in Theorems 5.1 and 5.2 respectively, and $B_t := Q_t^0 \cap B^\text{Spec}(\theta_0) \cap B^\text{RSC}(\theta_t)$. Let $\theta^* \in \arg\min_{\theta \in B^\text{Spec}(\theta_0)} \mathcal{L}(\theta)$, $\bar{\theta}_{t+1} \in \arg\min_{\theta \in B_t} \mathcal{L}(\theta)$, and $\gamma_t := \frac{\mathcal{L}(\bar{\theta}_{t+1}) - \mathcal{L}(\theta^*)}{\mathcal{L}(\theta_t) - \mathcal{L}(\theta^*)}$. Let $\alpha_t, \beta$ be as in Theorems 5.1 and 5.2 respectively, and $B_t := Q_t^0 \cap B^\text{Spec}(\theta_0) \cap B^\text{RSC}(\theta_t)$. Consider Assumptions 2 and 4 and gradient descent with step size $\eta_t = \frac{\alpha_t}{\beta}, \omega_t \in (0, 2)$. Then, with probability at least $(1 - \frac{2\alpha_t \omega_t}{m})$, we have we have $\gamma_t \in [0, 1)$ and

$$\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\theta^*) \leq \left(1 - \frac{\alpha_t \omega_t}{\beta}(1 - \gamma_t)(2 - \omega_t)\right) (\mathcal{L}(\theta_t) - \mathcal{L}(\theta^*)). \tag{14}$$

Proof. We start by showing $\gamma_t = \frac{\mathcal{L}(\bar{\theta}_{t+1}) - \mathcal{L}(\theta^*)}{\mathcal{L}(\theta_t) - \mathcal{L}(\theta^*)}$ satisfies $0 \leq \gamma_t < 1$. Since $\theta^* \in \arg\min_{\theta \in B^\text{Spec}(\theta_0)} \mathcal{L}(\theta)$ and $\theta_{t+1} \in B^\text{Spec}(\theta_0)$ by the definition of gradient descent and Assumption 4, we have

$$\mathcal{L}(\theta^*) \leq \mathcal{L}(\theta_{t+1}) \leq \mathcal{L}(\theta_t) - \frac{\omega_t}{\beta} \left(1 - \frac{\omega_t}{2}\right) \|\nabla_{\theta} \mathcal{L}(\theta_t)\|_2^2 < \mathcal{L}(\theta_t),$$

where (a) follows from (40). Since $\mathcal{L}(\bar{\theta}_{t+1}) \geq \mathcal{L}(\theta^*)$ and $\mathcal{L}(\theta_t) > \mathcal{L}(\theta^*)$, we have $\gamma_t \geq 0$. Further, we have $\mathcal{L}(\bar{\theta}_{t+1}) < \mathcal{L}(\theta_t)$, and so we have $\gamma_t < 1$. To see this, consider two cases: (i) $\theta_{t+1} \in B_t$ and (ii) $\theta_{t+1} \notin B_t$. When $\theta_{t+1} \in B_t$, we have $\mathcal{L}(\theta_{t+1}) \leq \mathcal{L}(\theta_t)$. When $\theta_{t+1} \notin B_t$, we only consider the case $\mathcal{L}(\theta_{t+1}) < \mathcal{L}(\theta_t)$. When $\theta_{t+1} \notin B_t$, we have $\mathcal{L}(\theta_{t+1}) \leq \mathcal{L}(\theta_t)$; otherwise, if $\mathcal{L}(\theta_{t+1}) \geq \mathcal{L}(\theta_t)$ then it follows $\mathcal{L}(\theta_{t+1}) < \mathcal{L}(\theta_t)$ by (40). So, let us consider level sets of the loss between $\mathcal{L}(\theta_{t+1})$ and $\mathcal{L}(\theta_t)$. Because of the definition of $Q^0_t$ (which defines a cone), the RSC property due to $\theta' \in B_t$, and the smoothness of the loss, we will have some $\theta \in B_t$ living in one of those level sets such that $\mathcal{L}(\theta_{t+1}) \leq \mathcal{L}(\theta') < \mathcal{L}(\theta_t)$, but then $\mathcal{L}(\theta_{t+1}) \leq \mathcal{L}(\theta')$ by definition, implying $\mathcal{L}(\theta_{t+1}) < \mathcal{L}(\theta_t)$. Hence, $\gamma_t < 1$. 

29
Now, with $\omega_t \in (0, 2)$, we have
\[
\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\theta^*) \\
= \mathcal{L}(\theta_{t+1}) - \mathcal{L}(\hat{\theta}_{t+1}) + \mathcal{L}(\hat{\theta}_{t+1}) - \mathcal{L}(\theta^*) \\
\leq \left( 1 - \frac{\alpha_t \omega_t}{\beta} (2 - \omega_t) \right) (\mathcal{L}(\theta_t) - \mathcal{L}(\hat{\theta}_{t+1})) + \left( 1 - \frac{\alpha_t \omega_t}{\beta} (2 - \omega_t) \right) (\mathcal{L}(\hat{\theta}_{t+1}) - \mathcal{L}(\theta^*)) \\
+ \left( \mathcal{L}(\hat{\theta}_{t+1}) - \left( 1 - \frac{\alpha_t \omega_t}{\beta} (2 - \omega_t) \right) \mathcal{L}(\hat{\theta}_{t+1}) - \mathcal{L}(\theta^*) \right) \\
\left( 1 - \frac{\alpha_t \omega_t}{\beta} (2 - \omega_t) \right) (\mathcal{L}(\theta_t) - \mathcal{L}(\theta^*)) \\
\left( 1 - \frac{\alpha_t \omega_t}{\beta} (2 - \omega_t) \right) (\mathcal{L}(\hat{\theta}_{t+1}) - \mathcal{L}(\theta^*)) \\
\left( 1 - \frac{\alpha_t \omega_t}{\beta} (2 - \omega_t) \right) (\mathcal{L}(\hat{\theta}_{t+1}) - \mathcal{L}(\theta^*)) .
\]
That completes the proof. \(\square\)

C  ANALYSIS FOR MODELS WITH $k$ OUTPUTS

In this section, we illustrate that our results can be extended to neural models with $k$ outputs.

C.1  OPTIMIZATION SETUP WITH $k$ OUTPUTS

Consider a training set $\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n, \mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^d, \mathbf{y}_i \in \mathcal{Y} \subseteq \mathbb{R}^k, k \geq 1$. We will denote by $X \in \mathbb{R}^{n \times d}$ the matrix whose $i$th row is $\mathbf{x}_i^\top$. The goal is to minimize the square loss:
\[
\mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^n \| \mathbf{y}_i - \hat{\mathbf{y}}_i \|^2 = \frac{1}{n} \sum_{i=1}^n \sum_{h=1}^k (\mathbf{y}_{ih} - f_h(\theta; \mathbf{x}_i))^2 ,
\]
where the prediction $\hat{\mathbf{y}}_i := \mathbf{f}(\theta; \mathbf{x}_i) \in \mathbb{R}^k$ is from a deep model, $f_h(\theta; \mathbf{x}_i), h \in [k]$ denotes the $h^{\text{th}}$ output, and $\theta \in \mathbb{R}^p$ denotes the parameter vector. In our setting $\mathbf{f}$ is a feed-forward multi-layer (fully-connected) neural network with depth $L$ and widths $m_l, l \in [L] := \{1, \ldots, L\}$ given by
\[
\alpha^{(0)}(\mathbf{x}) = \mathbf{x} ,
\]
\[
\alpha^{(l)}(\mathbf{x}) = \phi \left( \frac{1}{\sqrt{m_{l-1}}} \mathbf{W}^{(l)} \alpha^{(l-1)}(\mathbf{x}) \right) , \quad l = 1, \ldots, L ,
\]
\[
f(\theta; \mathbf{x}) = \alpha^{(L+1)}(\mathbf{x}) = \frac{1}{\sqrt{m_L}} \mathbf{V}^\top \alpha^{(L)}(\mathbf{x}) ,
\]
where $\mathbf{W}^{(l)} \in \mathbb{R}^{m_1 \times m_{l-1}}, l \in [L]$ are layer-wise weight matrices, $\mathbf{V} \in \mathbb{R}^{m_L \times k}$ is the last layer matrix, $\phi(\cdot)$ is the smooth (pointwise) activation function, and the total set of parameters
\[
\theta := \text{vec}(\mathbf{W}^{(1)})^\top, \ldots, \text{vec}(\mathbf{W}^{(L)})^\top, \mathbf{V}^\top \in \mathbb{R}^{\sum_{l=1}^L m_l m_{l-1} + km_L} ,
\]
with $m_0 = d$. Note that the total number of parameters is $p = \sum_{l=1}^L m_l m_{l-1} + km_L$. For simplicity, we will assume that the width of all the layers is the same, i.e., $m_l = m, l \in [L]$.

Define the pointwise loss $\ell_{ih} := (\mathbf{y}_{ih} - \hat{\mathbf{y}}_{ih})^2$ and denote its first- and second-derivative w.r.t. $\mathbf{y}_{ih}$ as $\ell'_{ih} := -2(\mathbf{y}_{ih} - \hat{\mathbf{y}}_{ih})$ and $\ell''_{ih} := 2$. Let $f_{ih}(\theta; \mathbf{x})$, $h \in [k]$ denote the $h^{\text{th}}$ output, and let the gradient and Hessian of $f_{ih}(\theta; \mathbf{x})$ w.r.t. $\theta$ be denoted as
\[
\nabla f_{ih} := \frac{\partial f_{ih}(\theta; \mathbf{x}_i)}{\partial \theta} , \quad \nabla^2 f_{ih} := \frac{\partial^2 f_{ih}(\theta; \mathbf{x}_i)}{\partial \theta^2} .
\]
By chain rule, the gradient and Hessian of the empirical loss w.r.t. $\theta$ are given by

$$
\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = \frac{1}{n} \sum_{i=1}^{n} \sum_{h=1}^{k} \ell_i' \nabla f_{ih},
$$

$$
\frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta^2} = \frac{1}{n} \sum_{i=1}^{n} \sum_{h=1}^{k} \left[ \ell_i'' \nabla f_{ih} \nabla f_{ih}^T + \ell_i' \nabla^2 f_{ih} \right].
$$

For the last layer, note that

$$
f_h(\theta; x_i) = \frac{1}{\sqrt{m}} v_h^T \alpha^{(L)}(x)
$$

where $v_h \in \mathbb{R}^m$ is the last layer vector corresponding to output $f_h$ and $V = [v_h] \in \mathbb{R}^{m \times k}$ is the last layer matrix. For the analysis, we update the definition of the spectral norm ball to work with each last layer vector $v_h$:

$$
B_{\rho, \rho_1}^{\text{Spec}}(\theta) := \left\{ \theta \in \mathbb{R}^p \text{ as in } (42) \mid \|W^{(\ell)} - \overline{W}^{(\ell)}\|_2 \leq \rho, \ell \in [L], \|v_h - \bar{v}_h\|_2 \leq \rho_1, h \in [k] \right\},
$$

Similarly, we update the initialization so each of the last layer vectors are unit norm.

**Assumption 5 (Initialization).** The initialization weights $w_{0,l,i,j}^{(l)} \sim \mathcal{N}(0, \sigma_n^2)$ for $l \in [L]$ where $\sigma_0 = \frac{\sigma_1}{2(1 + \frac{\sigma_n}{\sqrt{\tau_m}})}$, $\sigma_1 > 0$, and $v_{0,h}, h \in [k]$ are random unit vectors with $\|v_{0,h}\|_2 = 1$. Further, we assume the input data satisfies: $\|x_i\|_2 = \sqrt{d}, i \in [n]$.

Based on the setup, following Theorem 4.1 we get the following result for the Hessian of each $f_h$:

**Theorem C.1 (Hessian Spectral Norm Bound).** Under Assumptions 7 and 5 for $\theta \in B_{\rho, \rho_1}^{\text{Spec}}(\theta_0)$ as in (43), $\rho_1 = O(\text{poly}(L))$, with probability at least $(1 - \frac{2(L+1)}{m})$, for any $x_i, i \in [n]$, we have

$$
\|\nabla^2 f(\theta; x_i)\|_2 \leq \frac{c_H}{\sqrt{m}},
$$

with $c_H = O(\text{poly}(L)(1 + \gamma^2L))$ where $\gamma := \sigma_1 + \frac{\rho_1}{\sqrt{m}}$.

### C.2 Restricted Strong Convexity and Smoothness

Let us assume we have a sequence of iterates $\{\theta_t\}_{t \geq 0}$ from gradient descent. Our RSC analysis will rely on the following $\tilde{Q}_k^t$-sets at step $t$, which avoids directions almost orthogonal to the average gradient of the predictor.

**Definition C.1 ($\tilde{Q}_k^t$ sets).** For iterate $\theta_t \in \mathbb{R}^p$, let $\tilde{g}_t = \frac{1}{n} \sum_{i=1}^{n} \sum_{h=1}^{k} \nabla f_h(\theta_t; x_i)$. For any $\kappa \in (0, 1)$, define $\tilde{Q}_k^t := \{ \theta \mid |\cos(\theta - \theta_t, \tilde{g}_t)| \geq \kappa \}$.

We define the set $B_t := \tilde{Q}_k^t \cap B_{\rho, \rho_1}^{\text{Spec}}(\theta_0) \cap B_{\rho_2}^{\text{Euc}}(\theta_t)$. We focus on establishing RSC w.r.t. the tuple $(B_t, \theta_t)$, where $B_{\rho, \rho_1}^{\text{Spec}}(\theta_0)$ becomes the feasible set for the optimization and $B_{\rho_2}^{\text{Euc}}(\theta_t)$ is an Euclidean ball around the current iterate.

**Theorem C.2 (RSC for $k$-output Square Loss).** For square loss, under Assumptions 7 and 5 with probability at least $(1 - \frac{2(L+1)}{m})$, $\forall \theta' \in \tilde{Q}_k^t \cap B_{\rho, \rho_1}^{\text{Spec}}(\theta_0) \cap B_{\rho_2}^{\text{Euc}}(\theta_t)$,

$$
\mathcal{L}(\theta') \geq \mathcal{L}(\theta_t) + (\theta' - \theta_t, \nabla \mathcal{L}(\theta_t)) + \frac{\alpha_t}{2} \|\theta' - \theta_t\|_2^2,
$$

with $\alpha_t = c_1 k \|\tilde{g}_t\|_2^2 - \frac{kc_2}{\sqrt{m}}$

where $\tilde{g}_t = \frac{1}{n} \sum_{i=1}^{n} \sum_{h=1}^{k} \nabla f_h(\theta_t; x_i)$, $c_1 = 2\kappa^2$ and $c_2 = 2c_H(2\rho_2 + \sqrt{kc_1})$, with $c_H$ is as in Theorem 4.1, $\theta$ as in Lemma 4.1 and $\rho_{1, \gamma}$ as in Lemma 4.2. Consequently, $\mathcal{L}$ satisfies RSC w.r.t. $(\tilde{Q}_k^t \cap B_{\rho, \rho_1}^{\text{Spec}}(\theta_0) \cap B_{\rho_2}^{\text{Euc}}(\theta_t), \theta_t)$ whenever $\alpha_t > 0$. 

31
Proof. For any $\theta' \in \hat{Q}_t^\kappa \cap B_\rho(\theta_0)$, by the second order Taylor expansion around $\theta_t$, we have

$$L(\theta') = L(\theta_t) + \langle \theta' - \theta_t, \nabla_\theta L(\theta_t) \rangle + \frac{1}{2} (\theta' - \theta_t)^T \frac{\partial^2 L(\hat{\theta}_t)}{\partial \theta^2} (\theta' - \theta_t),$$

where $\hat{\theta}_t = \xi \theta' + (1 - \xi) \theta_t$ for some $\xi \in [0, 1]$. We note that $\hat{\theta}_t \in B_{\rho, \rho_1}(\theta_0)$ since,

- $\|\hat{\theta}_t^{(l)} - \theta_0^{(l)}\|_2 \leq \rho$, for any $l \in [L]$, where the last inequality follows from our assumption $\hat{\theta}_t^{(l)} \in B_{\rho, \rho_1}(\theta_0)$; and
- $\|\hat{\theta}_t^{(L+1)} - \theta_0^{(L+1)}\|_2 = \|\hat{V}_0 - V_0\|_F \leq k \rho_1$, by following a similar derivation as in the previous point.

Focusing on the quadratic form in the Taylor expansion and recalling the form of the Hessian, we get

$$\left(\theta' - \theta_t\right)^T \frac{\partial^2 L(\hat{\theta}_t)}{\partial \theta^2} (\theta' - \theta_t)$$

$$= \left(\theta' - \theta_t\right)^T \frac{1}{n} \sum_{i=1}^{n} \sum_{h=1}^{k} \left[ \ell_{ih} \sum_{i=1}^{n} \frac{\partial f_h(\hat{\theta}_t; x_i)}{\partial \hat{\theta}_t} \frac{\partial f_h(\hat{\theta}_t; x_i)}{\partial \theta_t} + \ell_{ih}^2 \frac{\partial^2 f_h(\hat{\theta}_t; x_i)}{\partial \theta^2} \right] \left(\theta' - \theta_t\right)$$

$$= \left\{ \left(\theta' - \theta_t\right)^T \frac{1}{n} \sum_{i=1}^{n} \sum_{h=1}^{k} \ell_{ih}^2 \left(\theta' - \theta_t\right) \right\}_{I_1} + \left\{ \left(\theta' - \theta_t\right)^T \frac{1}{n} \sum_{i=1}^{n} \sum_{h=1}^{k} \ell_{ih} \left(\theta' - \theta_t\right) \right\}_{I_2}$$

where $\ell_{ih} = \ell(y_{ih}, f_h(\hat{\theta}_t; x_i))$, $\ell_{ih}' = \frac{\partial \ell(y_{ih}, z)}{\partial z} \bigg|_{z = f_h(\hat{\theta}_t; x_i)}$, and $\ell_{ih}'' = \frac{\partial^2 \ell(y_{ih}, z)}{\partial z^2} \bigg|_{z = f_h(\hat{\theta}_t; x_i)}$. 

32
Now, note that

\[ I_1 = \frac{1}{n} \sum_{i=1}^{n} \sum_{h=1}^{k} \ell''_{ih} \left( \theta' - \theta_t, \frac{\partial f_h(\tilde{\theta}_t; x_i)}{\partial \theta} \right)^2 \geq 2 \frac{n}{n} \sum_{i=1}^{n} \sum_{h=1}^{k} \left( \theta' - \theta_t, \frac{\partial f_h(\tilde{\theta}_t; x_i)}{\partial \theta} \right)^2 + 2 \frac{n}{n} \sum_{i=1}^{n} \sum_{h=1}^{k} \left( \theta' - \theta_t, - \frac{\partial f_h(\tilde{\theta}_t; x_i)}{\partial \theta} \right)^2 + 4 \frac{n}{n} \sum_{i=1}^{n} \sum_{h=1}^{k} \left( \theta' - \theta_t, \frac{\partial f_h(\tilde{\theta}_t; x_i)}{\partial \theta} \right) \left( \theta' - \theta_t, - \frac{\partial f_h(\tilde{\theta}_t; x_i)}{\partial \theta} \right) \right) \]

\[ \geq 2 \frac{n}{n} \sum_{i=1}^{n} \sum_{h=1}^{k} \left( \theta' - \theta_t, \frac{\partial f_h(\tilde{\theta}_t; x_i)}{\partial \theta} \right)^2 - 4 \frac{n}{n} \sum_{i=1}^{n} \sum_{h=1}^{k} \left( \frac{\partial f_h(\tilde{\theta}_t; x_i)}{\partial \theta} - \frac{\partial f_h(\tilde{\theta}_t; x_i)}{\partial \theta} \right) \left( \frac{\partial f_h(\tilde{\theta}_t; x_i)}{\partial \theta} - \frac{\partial f_h(\tilde{\theta}_t; x_i)}{\partial \theta} \right) \left\| \theta' - \theta_t \right\|^2 \]

where (a) follows by Cauchy-Schwartz inequality; (b) follows by Jensen’s inequality (first term) and the use of Theorem 4.1 and Lemma 4.1 due to \( \tilde{\theta}_t \in B_{\text{spec}}(\theta_0) \); (c) follows from \( \left\| \theta_t - \theta_t \right\|^2 = \left\| \xi \theta' + (1 - \xi) \theta_t - \theta_t \right\|^2 = \xi \left\| \theta' - \theta_t \right\|^2 \leq \left\| \theta' - \theta_t \right\|^2 \); (d) follows since \( \theta' \in Q_k \) and from the fact that \( p \oplus q = \cos(p, q) \|p\| \|q\| \) for any vectors \( p, q \).

For analyzing \( I_2 \), first note that for square loss, \( \ell_{ih,t} = 2(\tilde{y}_{ih,t} - y_{ih}) \) is \( \cdot \) with \( \tilde{y}_{ih,t} = f_h(\theta_t; x_i) \), so that for the vector \( [\ell_{ih,t}] \), we have \( \frac{1}{n} \left\| [\ell_{ih,t}] \right\|^2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{h=1}^{k} (\tilde{y}_{ih,t} - y_{ih})^2 = 4C(\theta_t) \). Further, with \( Q_{ih,t} = (\theta' - \theta_t) \times \frac{\partial^2 f_h(\tilde{\theta}_t; x_i)}{\partial \theta^2}(\theta' - \theta_t) \), we have

\[ |Q_{ih,t}| = \left( \theta' - \theta_t \right) \times \frac{\partial^2 f_h(\tilde{\theta}_t; x_i)}{\partial \theta^2}(\theta' - \theta_t) \leq \left\| \theta' - \theta_t \right\|^2 \left( \frac{\partial^2 f_h(\tilde{\theta}_t; x_i)}{\partial \theta^2} \right) \leq \frac{C_H^2 \left\| \theta' - \theta_t \right\|^2}{\sqrt{m}} \]

Now, note that

\[ I_2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{h=1}^{k} \ell''_{ih} Q_{ih,t} \]

\[ \geq - \sum_{i=1}^{n} \sum_{h=1}^{k} \left( \frac{1}{\sqrt{n}} \ell''_{ih} \right) \left( \frac{1}{\sqrt{n}} Q_{ih,t} \right) \]

\[ \geq \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{h=1}^{k} \ell''_{ih} \right)^{1/2} \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{h=1}^{k} Q_{ih,t}^2 \right)^{1/2} \]

\[ \geq -2 \sqrt{\mathcal{L}(\theta_t)} \frac{C_H^{1/2} \left\| \theta' - \theta_t \right\|^2}{\sqrt{m}} \]
where (a) follows by Cauchy-Schwartz inequality.

Putting the lower bounds on $I_1$ and $I_2$ back, with $\bar{g}_t = \frac{1}{nk} \sum_{i=1}^n \sum_{h=1}^k \frac{\partial f_h(\theta; x_i)}{\partial \theta^2} \cdot \bar{\partial}_t\bar{\ell}_t$, we have

$$(\theta' - \theta_t) \frac{\partial^2 \mathcal{L}(\hat{\theta}_t)}{\partial \theta^2} (\theta' - \theta_t) \geq k \left( 2k^2 \| \bar{g}_t \|^2_2 - \frac{4q_c H \| \theta' - \theta_t \|^2_2 + 2c_H \sqrt{\mathcal{L}(\hat{\theta}_t)}}{\sqrt{m}} \right) \| \theta' - \theta_t \|^2_2.$$ 

Now, since $\theta' \in B_{\rho_2}^{\text{Spec}}(\theta_t)$, $\| \theta' - \theta_t \|^2 \leq \rho_2$, so we have

$$(\theta' - \theta_t)^\top \frac{\partial^2 \mathcal{L}(\hat{\theta}_t)}{\partial \theta^2} (\theta' - \theta_t) \geq k \left( 2k^2 \| \bar{g}_t \|^2_2 - \frac{4q_c H \rho_2 + 2c_H \sqrt{\mathcal{L}(\hat{\theta}_t)}}{\sqrt{m}} \right) \| \theta' - \theta_t \|^2_2 \geq k \left( 2k^2 \| \bar{g}_t \|^2_2 - \frac{4q_c H \rho_2 + 2c_H \sqrt{kc_{\rho_1,\gamma}}}{\sqrt{m}} \right) \| \theta' - \theta_t \|^2_2,$$

where the last inequality follows from Lemma 4.2. That completes the proof. \hfill \Box

**Theorem C.3 (Local Smoothness for $k$-output Square Loss).** Under Assumptions 2, 3 and 4 with probability at least $(1 - \frac{2(L+1)}{m})$, $\forall \theta, \theta' \in B_{\rho_1,\rho_2}^{\text{Spec}}(\theta_0)$,

$$\mathcal{L}(\theta') \leq \mathcal{L}(\theta) + (\theta' - \theta, \nabla_\theta \mathcal{L}(\theta)) + \frac{\beta}{2} \| \theta' - \theta \|^2_2,$$

where $\beta = 2k\rho^2 + \frac{2kc_H \sqrt{kc_{\rho_1,\gamma}}}{\sqrt{m}}$, with $c_H$ as in Theorem 4.2 as in Lemma 4.2 and $c_{\rho_1,\gamma}$ as in Lemma 4.2. Consequently, $\mathcal{L}$ is locally $\beta$-smooth.

**Proof.** By the second order Taylor expansion about $\check{\theta} \in B_{\rho_1,\rho_2}^{\text{Spec}}(\theta_0)$, we have $\mathcal{L}(\theta') = \mathcal{L}(\check{\theta}) + (\theta' - \check{\theta}, \nabla_\theta \mathcal{L}(\check{\theta})) + \frac{1}{2} (\theta' - \check{\theta}) \top \frac{\partial^2 \mathcal{L}(\check{\theta})}{\partial \theta^2} (\theta' - \check{\theta})$, where $\check{\theta} = \xi \theta' + (1 - \xi) \check{\theta}$ for some $\xi \in [0, 1]$. Then,

$$(\theta' - \check{\theta}) \top \frac{\partial^2 \mathcal{L}(\check{\theta})}{\partial \theta^2} (\theta' - \check{\theta}) = (\theta' - \check{\theta}) \top \frac{1}{n} \sum_{i=1}^n \sum_{h=1}^k \left[ \ell''_{ih} \frac{\partial f_h(\check{\theta}; x_i)}{\partial \theta} \frac{\partial f_h(\check{\theta}; x_i)}{\partial \theta} + \ell''_{ih} \frac{\partial^2 f_h(\check{\theta}; x_i)}{\partial \theta^2} \right] (\theta' - \check{\theta})$$

$$= \frac{1}{n} \sum_{i=1}^n \sum_{h=1}^k \ell''_{ih} \left( \theta' - \check{\theta}, \frac{\partial f_h(\check{\theta}; x_i)}{\partial \theta} \right) \left( \theta' - \check{\theta}, \frac{\partial^2 f_h(\check{\theta}; x_i)}{\partial \theta^2} \right),$$

where $\ell_{ih} = \ell(y_i, f_h(\check{\theta}; x_i))$, $\ell''_{ih} = \left. \frac{\partial^2 f_h(\check{\theta}; x_i)}{\partial z^2} \right|_{z=f_h(\check{\theta}; x_i)}$, and $\ell''_{ih} = \left. \frac{\partial^2 f_h(\check{\theta}; x_i)}{\partial z^2} \right|_{z=f_h(\check{\theta}; x_i)}$. Now, note that

$$I_1 = \frac{1}{n} \sum_{i=1}^n \sum_{h=1}^k \ell''_{ih} \left( \theta' - \check{\theta}, \frac{\partial f_h(\check{\theta}; x_i)}{\partial \theta} \right)^2 \leq (a) \frac{2}{n} \sum_{i=1}^n \sum_{h=1}^k \left\| \frac{\partial f_h(\check{\theta}; x_i)}{\partial \theta} \right\|^2_2 \| \theta' - \check{\theta} \|^2_2 \leq (b) \frac{2k\rho^2}{\sqrt{m}} \left\| \theta' - \check{\theta} \right\|^2_2,$$

where (a) follows by the Cauchy-Schwartz inequality and (b) from Lemma 4.1.

For $I_2$, first note that for square loss, $\ell''_{ih} = 2(\hat{y}_{ih} - y_{ih})$ with $\hat{y}_{ih} = f_h(\check{\theta}; x_i)$, so that for the vector $[\ell''_{ih}]$, we have $\frac{1}{n} \| [\ell''_{ih}] \|^2_2 = \frac{4}{n} \sum_{i=1}^n \sum_{h=1}^k (\hat{y}_{ih} - y_{ih})^2 = 4\mathcal{L}(\check{\theta})$. Further, with

$$Q_{ih} = (\theta' - \check{\theta}) \top \frac{\partial^2 f_h(\check{\theta}; x_i)}{\partial \theta^2} (\theta' - \check{\theta}),$$

Published as a conference paper at ICLR 2023
we have
\[ |Q_{ih}| = \left| (\theta' - \bar{\theta})^\top \frac{\partial^2 f_h(\hat{\theta}; \mathbf{x}_i)}{\partial \theta^2} (\theta' - \bar{\theta}) \right| \leq \|\theta' - \bar{\theta}\|_2 \left\| \frac{\partial^2 f_h(\hat{\theta}; \mathbf{x}_i)}{\partial \theta^2} \right\|_2 \leq c_H \|\theta' - \bar{\theta}\|_2^2. \]

Then, we have
\[
I_2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{h=1}^{k} \ell'_{ih} (\theta' - \bar{\theta})^\top \frac{\partial^2 f_h(\hat{\theta}; \mathbf{x}_i)}{\partial \theta^2} (\theta' - \bar{\theta}) \\
\leq \left| \sum_{i=1}^{n} \sum_{h=1}^{k} \frac{1}{\sqrt{n}} \ell'_{ih} \left( \frac{1}{\sqrt{n}} Q_{ih} \right) \right| \\
\leq \left( \frac{1}{n} \sum_{h=1}^{k} \|\ell'_{ih}\|_2^2 \right)^{1/2} \left( \frac{1}{n} \sum_{i=1}^{n} \sum_{h=1}^{k} Q_{ih}^2 \right)^{1/2} \\
\leq 2 \sqrt{\mathcal{L}(\bar{\theta}) k c_H} \|\theta' - \bar{\theta}\|_2^2 \sqrt{m},
\]

where (a) follows by Cauchy-Schwartz. Putting the upper bounds on \( I_1 \) and \( I_2 \) back, we have
\[
(\theta' - \bar{\theta})^\top \frac{\partial^2 \mathcal{L}(\bar{\theta})}{\partial \theta^2} (\theta' - \bar{\theta}) \leq \left[ 2 k o^2 + \frac{2 k c_H \sqrt{\mathcal{L}(\bar{\theta})}}{\sqrt{m}} \right] \|\theta' - \bar{\theta}\|_2^2 \\
\leq \left[ 2 k o^2 + \frac{2 k c_H \sqrt{k c_H \rho_1 \gamma}}{\sqrt{m}} \right] \|\theta' - \bar{\theta}\|_2^2,
\]

where the last inequality follows from Lemma 4.2. That completes the proof.

Note that we now have the RSC and smoothness results for the \( k \)-output case similar to the single output case. With these properties, the rest of the convergence analysis for the \( k \)-output case stays the same as before.