A Generalised Weak Convexity and Two Layer Neural Networks

In this section we generalise the notation of weak convexity considered within the main body of the manuscript. This section proceeds as follows. Section A.1 presents the generalised notation of weak convexity. Section A.2 gives bounds on the **Optimisation & Approximation Error** for two layer neural networks utilising generalised weak convexity.

A.1 Generalised Weak Convexity

As highlighted in Remark 4, a drawback of Theorem 2 when optimising both layers for two layer neural network is the **Approximation Error**'s dependence on the squared Euclidean norm of the population risk minimiser $\|\omega^\star\|_2^2$. This stems from the weak convexity assumption (Assumption 3) penalising all co-ordinates equally i.e. adding $\|\omega\|_2^2$. To improve upon this we introduce the following assumption which aims to better encode the Hessian's structure.

Assumption 5 *For a convex set $\mathcal{X} \subseteq \mathbb{R}^p$ there exists non-negative function $G_{\mathcal{X}} : \mathbb{R}^p \to \mathbb{R}$ such that almost surely*

$$u^\top \nabla^2 R(\omega) u \geq -G_{\mathcal{X}}(u) \qquad \text{for any } u \in \mathbb{R}^p, \omega \in \mathcal{X}$$

Assumption 5 is a weakening of the weak convexity Assumption 3 in two respects. The first is that we have restricted ourselves to a convex set $\mathcal{X}$ while Assumption 3 considers case $\mathcal{X} = \mathbb{R}^p$. The second difference, is that the quadratic form of the Hessian with vectors u is lower bounded by a function $G_{\mathcal{X}}(u)$. This function can then encode additional structure of the Hessian which can be utilised to obtain tighter control on the **Approximation Error**. We note by considering the function $G_{\mathcal{X}}(u) = \frac{\epsilon_{\mathcal{X}}}{2}\|u\|_2^2$ for some $\epsilon_{\mathcal{X}} \geq 0$, we recover Assumption 3 restricted to the convex set $\mathcal{X}$. Meanwhile more generally, $G_{\mathcal{X}}(u)$ may depend upon other norms which penalise subsets of co-ordinates as well as interactions. Given this assumption, we are present the following proposition which bounds the **Optimisation & Approximation Error** that appears within the test error decomposition (2).

Proposition 1 (Opt. & Approx. Error) *Consider Assumption 1 and 5, step size $\eta_s = \eta$ for all $s \geq 1$ and $\widetilde{\omega} \in \mathcal{X}$. Suppose that $\omega_s \in \mathcal{X}$ for all $s \geq 1$. If $\eta\beta \leq 1/2$ then*

$$\underbrace{\mathbf{E}_I[E[R(\widehat{\omega}_I]] - r(\omega^\star)}_{\textit{Opt. \& Approx. Error}} \leq \underbrace{\frac{\mathbf{E}[\|\widehat{\omega}_0 - \widetilde{\omega}\|_2^2]}{2\eta t}}_{\textit{Optimisation Error}} + \underbrace{\frac{1}{2}\frac{1}{t}\sum_{s=1}^{t}\mathbf{E}[G_{\mathcal{X}}(\widetilde{\omega} - \widehat{\omega}_s)] + \mathbf{E}[R(\widetilde{\omega})] - r(\omega^\star)}_{\textit{Approximation Error}}$$

Observe that within Proposition 1 the bound now depends on the Hessian structure through $G_{\mathcal{X}}(\cdot)$ as well as now a free parameter $\widetilde{\omega} \in \mathcal{X}$. The right most term can then be interpreted as an **Approximation Error**, as the series involving $G_{\mathcal{X}}(\widetilde{\omega} - \widehat{\omega}_s)$ can be viewed as a regulariser, with $\widetilde{\omega}$ chosen thereafter to minimise a penalised empirical risk. Although, we note that the regulariser depends upon the iterates of gradient descent i.e. the difference $\widetilde{\omega} - \widehat{\omega}_s$, as such, to proceed we must decouple $\widetilde{\omega}$ and $\widehat{\omega}_s$ by utilising the structure of $G_{\mathcal{X}}(\cdot)$. We therefore now consider the case of a two layer neural network.

A.2 Optimisation and Approximation Error for Two Layer Neural Networks

In this section we demonstrate how the generalised notation of weak convexity just described in Section A.1 can be applied to a two layer neural network when optimising both layers. Specifically, we consider the setting described in Section 4.1, where the loss is a composition of a convex function $g(\cdot, y)$ and a Two Layer Neural network f. Let us begin by introducing some additional notation. For a vector $z \in \mathbb{R}^q$ let $\mathrm{Diag}(z) \in \mathbb{R}^{q\times q}$ be the square diagonal matrix where $\mathrm{Diag}(z)_{ii} = z_i$ for $i = 1, \ldots, q$. Moreover, let us denote the vector $u = (A(u), v(u)) \in \mathbb{R}^{Md+M}$ which is composed in a manner matching $\omega = (A, v)$, so that $A(u) \in \mathbb{R}^{M\times d}$ is a matrix associated to the first layer of weights and $v(u) \in \mathbb{R}^M$ is the vector associated to the second layer of weights. Let us denote the maximum between two real numbers $a, b \in \mathbb{R}$ as $a \vee b = \max\{a, b\}$.

To consider the generalised weak convexity Assumption 5 we first introduce the appropriate set $\mathcal{X}$ and functional $G_{\mathcal{X}}$. For $L_v \geq 0$ consider the set $\mathcal{X}_{L_v} \subseteq \mathbb{R}^{Md+d}$ as well as for $C \geq 0$ the function $G_C : \mathbb{R}^p \to \mathbb{R}^+$. The

set is then defined as $\mathcal{X}_{L_v} := \{\omega \in \mathbb{R}^{Md+d} : \omega = (A, v), \|v\|_\infty \leq L_v\}$, while the functional is defined for $u = (A(u), v(u)) \in \mathbb{R}^{Md+M}$ as

$$G_C(u) := \frac{C}{M^c}\|A(u)\|_F^2 + \frac{C}{M^c}\max_{\|z\|_\infty \leq 1} \big\|A(u)^\top \mathrm{Diag}(z)v(u)\big\|_2.$$

The set $\mathcal{X}_{L_v}$ in our case is technical and is included to, in short, control the second layer weights that arises within second derivative of first layer, see also Theorem 3. Meanwhile, the function G_C now incorporates the structure of the empirical risk Hessian. In particular, there is no term depending on the squared norm $\|v\|_2^2$, since the Hessian restricted to the second layer is zero. The second term in $G_C(\cdot)$ then arises to control the interaction between the first and second layers, with the maximum over vectors $\|z\|_\infty \leq 1$ coming from the activation $\sigma(\cdot)$.

Given the function G_C we must now introduce a related function $H : \mathbb{R}^{Md+M} \to \mathbb{R}$ which encodes the structure of regularisation. It is similarly defined as follows

$$H(u) := \|A(u)\|_F^2 + \sqrt{\eta t}\|v(u)\|_2 + \max_{\|z\|_\infty \leq 1} \|A(u)^\top \mathrm{Diag}(z)v(u)\|_2.$$

Note the structure of $H(\cdot)$ closely aligns with $G_C(\cdot)$ although an extra term $\sqrt{\eta t}\|v(u)\|_2$ is included. This arises due to the coupling between the gradient descent iterates and place holder $\widetilde{\omega}$ i.e. $G_C(\widetilde{\omega} - \widehat{\omega}_s)$. Given this, it will be convenient to denote for $\lambda \geq 0$ a minimiser $\widehat{\omega}_\lambda \in \mathrm{argmin}_{\omega \in \mathcal{X}_{L_v}} \big\{R(\omega) + \lambda H(\omega - \widehat{\omega}_0)\big\}$. The following theorem then utilises Proposition 1 to bound the **Optimisation & Approximation Error** within the test error decomposition (2).

Theorem 4 (Two Layer Neural Network) *Consider loss regularity Assumption 1 alongside the setting of Theorem 3 with bounded activation $|\sigma(\cdot)| \leq L_\sigma$. Initialise gradient descent at $\widehat{\omega}_0 = (A^0, v^0) \in \mathbb{R}^{Md+M}$. If the following holds almost surely*

$$L_v \geq \|v^\star\|_\infty \vee \Big(\|v^0\|_\infty + \frac{\eta t L_{g'} L_\sigma}{M^c}\Big)$$
$$C \geq 2L_{g'}\Big[\Big(L_{\sigma'}\sqrt{\mathrm{Tr}\left(\widehat{\Sigma}\right)}\Big) \vee (L_{\sigma''}L_v\|\widehat{\Sigma}\|_2)\Big],$$

then the optimisation and approximation error is bounded

$$\underbrace{\mathbf{E}_I[\mathbf{E}[R(\widehat{\omega}_I)]] - r(\omega^\star)}_{\textbf{\textit{Opt. \& Approx. Error}}} \leq \underbrace{\frac{\mathbf{E}[\|\widehat{\omega}_\lambda - \widehat{\omega}_0\|_2^2]}{2\eta t}}_{\textbf{\textit{Opt. Error}}} + \underbrace{\frac{3C\mathbf{E}[R(\omega_0) - R(\widehat{\omega}^\star)]\eta t}{M^c}}_{\textbf{\textit{Opt. Component.}}} + \underbrace{\mathbf{E}[\lambda]H(\omega^\star - \widehat{\omega}_0)}_{\textbf{\textit{Stat. Approx.}}}$$

where $\lambda = 3C(\sqrt{R(\widehat{\omega}_0) - R(\widehat{\omega}^\star)} \vee 1)/M^c$.

Observe in Theorem 4 that the parameter L_v, which controls the constraint set $\mathcal{X}_{L_v}$, is required to grow as $O(1 + \eta t/M^c)$, and therefore, is constant for sufficiently wide neural networks (see also discussion in Section 3.2). Meanwhile, the parameter C, which controls the functional, must almost surely upper bound spectral quantities related to the covariates covariance, namely, the trace $\mathrm{Tr}\left(\widehat{\Sigma}\right)$. This can then be more refined than assuming almost surely bounded co-ordinates i.e. $\|x\|_\infty$, as done within the main body of the manuscript. The resulting bound on the **Optimisation and Approximation Error** then consists of two new terms: **Opt. Component** and **Stat. Approx.**. The **Opt. Component.** is order $O(d\eta t/M^c)$ and arises from the dependence on gradient descent iterates $\widehat{\omega}_s$. Meanwhile, the **Stat. Approx.** depends upon the population risk minimiser evaluated at the regulariser i.e. $\mathbf{E}[\lambda]H(\omega^\star - \widehat{\omega}_0)$, and can be interpreted as a statistical bias resulting from the non-convexity. Note it now depends upon $H(\cdot)$ and is scaled by λ which is $O\Big(\sqrt{\mathrm{Tr}(\widehat{\Sigma})}/M^c\Big)$.

It is natural to investigate (picking $\widehat{\omega}_0 = 0$) the size of the **Stat. Approx.** for a particular problem instance. Following the limitations of the weakly convex setting highlighted within remark 4 in the main body of the manuscript, we focus on whether **Stat. Approx.** can be smaller than the Total Weight of the network $\mathrm{TW}(f)$. Denoting $\omega^\star = (A^\star, v^\star)$ a minimiser of the population risk, we focus on upper bounding $\sqrt{\mathrm{Tr}(\widehat{\Sigma})}H(\omega^\star)/M^c$ by the Total Weight of the network, since $\sqrt{\mathrm{Tr}(\widehat{\Sigma})}H(\omega^\star)/M^c$ equals the **Stat. Approx.** up to constants

Begin by noting that $H(\cdot)$ depends upon the interaction between the first and second layer, therefore, introduce the single valued decomposition $A^\star = \Gamma\Lambda\Xi^\top = \sum_{\ell=1}^d \lambda_\ell \gamma_\ell \xi_\ell^\top$ where left-singular vectors $\{\gamma_\ell\}_{\ell=1}^d$ are the columns of $\Gamma \in \mathbb{R}^{M\times d}$, the right-singular vectors $\{\xi\}_{\ell=1}^d$ are the columns of $\Xi \in \mathbb{R}^{d\times d}$ and $\{\lambda_i\}_{i=1}^d$ are the singular values. Let us also denote the element wise multiplication of two vectors $u, v \in \mathbb{R}^p$ as $u \odot v = (u_1 v_1, \dots, u_p v_p) \in \mathbb{R}^p$. With the single valued decomposition the regulariser can then be written as follows

$$\frac{1}{M^c} H(\omega^\star) = \underbrace{\frac{1}{M^c}\big(\|A^\star\|_F^2 + \sqrt{\eta t}\|v^\star\|_2\big)}_{\textbf{Norm Condition}} + \underbrace{\max_{\|z\|_\infty \le 1} \frac{1}{M^c}\sqrt{\sum_{j=1}^d |\lambda_j|^2 |\langle z, v^\star \odot \gamma_j\rangle|^2}}_{\textbf{Interaction}}.$$

The first two terms above depend upon the norm of the population risk minimiser $\omega^\star$, while the third term now encodes the interaction between the first and second layer. For clarity, let $a_1, a_2 \in \mathbb{R}$ and assume each of the second layers weights are the same magnitude so $|v_j^\star| = a_1$ for $j = 1, \dots, M$ as well as the singular values so $|\lambda_j| = a_2$ for $j = 1, \dots, d$. The **Interaction** term can then be upper bounded by taking the maximum $\max_{\|z\|_\infty \le 1}$ inside the series

$$\textbf{Interaction} = \frac{1}{M^c}\max_{\|z\|_\infty \le 1}\sqrt{\sum_{j=1}^d |\lambda_j|^2 |\langle z, v^\star \odot \gamma_j\rangle|^2} \le \frac{1}{M^c}\sqrt{\sum_{j=1}^d |\lambda_j|^2 \|v^\star \odot \gamma_j\|_1^2} = a_1 a_2 \frac{1}{M^c}\sqrt{\sum_{j=1}^d \|\gamma_j\|_1^2}.$$

Note that this quantity now aligns with the $\ell_{1,2}$ element-wise matrix norm on the left-singular vectors of $A^\star$ i.e. Γ. Therefore, let us assume that the left-singular vectors $\{\gamma_i\}_{i=1}^d$ are supported on disjoint sets of size $M/d \ge s \ge 1$ and are such that $|(\gamma_i)_j| = \frac{1}{\sqrt{s}}$ for $i = 1, \dots, d$, $j \in \mathrm{Supp}(\gamma_i)$ (where $\mathrm{Supp}(\cdot)$ denotes the support of a vector) and zero otherwise. Noting that $\|v^\star\|_2 = a_1\sqrt{ds}$ since the second layer of weights will be supported on the non-zero rows of $A^\star$ as well as that $\|\gamma_i\|_1 = \sqrt{s}$, yields the upper bound on the **Stat. Approx.** term

$$\frac{\sqrt{\mathrm{Tr}\left(\widehat{\Sigma}\right)}H(\omega^\star)}{M^c} \le \sqrt{\mathrm{Tr}\left(\widehat{\Sigma}\right)}\Big(\underbrace{\frac{da_2}{M^c} + \frac{a_1\sqrt{\eta t}\sqrt{ds}}{M^c}}_{\textbf{Norm Condition}} + \underbrace{\frac{a_1 a_2\sqrt{ds}}{M^c}}_{\textbf{Interaction}}\Big).$$

Let us now consider the Total Weight with this particular choice of weights $\omega^\star = (v^\star, A^\star)$. Precisely, for this particular choice of second layer weights $v^\star$, singular values $\{\lambda_j\}_{j=1}^d$ and singular vectors $\{\gamma_i\}_{i=1}^d$, the Total Weight aligns with the $\ell_{2,1}$ element-wise matrix norm of $\Gamma^\top$

$$TW(f) = \frac{1}{M^c}\sum_{j=1}^M |v_j^\star| \|A_j^\star\|_2 = \frac{a_1 a_2}{M^c}\sum_{j=1}^M \sqrt{\sum_{i=1}^d (\gamma_i)_j^2} = \frac{a_1 a_2}{M^c}\sum_{j=1}^{d\times s}\sqrt{\frac{1}{s}} = a_1 a_2 \frac{d\sqrt{s}}{M^c}.$$

Where we note that the first equality arises from the single valued decomposition $A_j^\star = \sum_{i=1}^d \lambda_i (\gamma_i)_j \xi_i$ and thus $\|A_j^\star\|_2 = \sqrt{\sum_{i=1}^d \lambda_i^2 (\gamma_i)_j^2}$. Meanwhile for the second equality, for each $j = 1, \dots, d\times s$ we have $j \in \mathrm{Supp}(\gamma_k)$ for at most one $k \in \{1, \dots, d\}$ since $\{\gamma_i\}_{i=1}^d$ are supported on disjoint sets. Meanwhile for $j = d\times s + 1, \dots, M$ we have $j \notin \mathrm{Supp}(\gamma_k)$ since the supports are of size at most s. This means for $j = 1, \dots, d\times s$ we get $\sum_{i=1}^d (\gamma_i)_j^2 = \sum_{i: j\in \mathrm{Supp}(\gamma_i)} (\gamma_i)_j^2 = \frac{1}{s}$, and then zero otherwise. Dividing the **Stat. Approx.** by the Total Weight we have

$$\frac{1}{TW(f)}\frac{\sqrt{\mathrm{Tr}\left(\widehat{\Sigma}\right)}H(\omega^\star)}{M^c} = \sqrt{\mathrm{Tr}\left(\widehat{\Sigma}\right)}\Big(\frac{1}{a_1\sqrt{s}} + \frac{\sqrt{\eta t}}{a_2\sqrt{d}} + \frac{1}{\sqrt{d}}\Big).$$

Now, if $s > d \ge 9$, $a_1, a_2 \ge 1$ and $\eta t\,\mathrm{Tr}(\widehat{\Sigma}) \le d/9$ then the **Stat. Approx.** error is upper bounded by the Total Weight $\sqrt{\mathrm{Tr}(\widehat{\Sigma})}H(\omega^\star)/M^c \le TW(f)$, as required.

B Proof of Generalisation Error Bounds under Weak Convexity

In this section we present the proofs related to the first half of the manuscript which gives generalisation error bounds for gradient descent under pointwise weak convexity Assumption 2 and standard weak convexity

Assumption 3. We begin by presenting the proof of Theorem 1 in Section B.1. Section B.2 present the proof of Lemma 1 which is technical result used within the proof of Theorem B.1. Section B.3 then presents and proves a generalisation error bound for gradient descent under standard weak convexity Assumption 3.

B.1 Proof of Theorem 1

In this section we give the proof of Theorem 1. Using equation (1) we can then bound the generalisation error for gradient descent in terms of the difference between gradient descent with and without a resampled datapoint. Specifically, using that the loss is L-Lipschitz as well as Jensen's Inequality to take the absolute value inside the expectation we get

$$|\mathbf{E}[R(\widehat{w}_t) - r(\widehat{w}_t)]| \leq \frac{1}{N}\sum_{i=1}^{N} |\mathbf{E}[\ell(\widehat{w}_t^{(i)}, Z_i') - \ell(\widehat{w}_t, Z_i')]| \leq \frac{L}{N}\sum_{i=1}^{N} \mathbf{E}[\|\widehat{w}_t^{(i)} - \widehat{w}_t\|_2].$$

For $i = 1, \ldots, N$ it then suffices to bound the deviation $\|\widehat{w}_t - \widehat{w}_t^{(i)}\|_2$. Using that the gradient of the empirical risk can be denoted $\nabla R^{(i)}(w) = \nabla R(w) + \frac{1}{N}(\nabla \ell(w, Z_i') - \nabla \ell(w, Z_i))$ alongside the Lipschitz assumption we get for any $k \geq 1$,

$$\|\widehat{w}_k - \widehat{w}_k^{(i)}\|_2 \leq \|\widehat{w}_{k-1} - \widehat{w}_{k-1}^{(i)} - \eta_{k-1}\big(\nabla R(\widehat{w}_{k-1}) - \nabla R(\widehat{w}_{k-1}^{(i)})\big)\|_2 + \frac{2\eta_{k-1}L}{N}. \tag{4}$$

The first term on the right hand side is then referred to as the expansiveness of the gradient update. Note, for $k = 1$ it is zero since the iterates with and without the resampled data point are initialised at the same location $\widehat{\omega}_0 = \widehat{\omega}_0^{(i)}$, and thus,

$$\|\widehat{w}_1 - \widehat{w}_1^{(i)}\|_2 \leq \frac{2\eta_0 L}{N}. \tag{5}$$

Therefore, let us consider the difference $\|\widehat{w}_k - \widehat{w}_k^{(i)}\|_2$ for $k \geq 2$. Expanding the expansiveness of the gradient update term we get

$$\begin{aligned}
&\|\widehat{w}_{k-1} - \widehat{w}_{k-1}^{(i)} - \eta_{k-1}\big(\nabla R(\widehat{w}_{k-1}) - \nabla R(\widehat{w}_{k-1}^{(i)})\big)\|_2^2 \\
&= \|\widehat{w}_{k-1} - \widehat{w}_{k-1}^{(i)}\|_2^2 + \eta_{k-1}^2\|\nabla R(\widehat{w}_{k-1}) - \nabla R(\widehat{w}_{k-1}^{(i)}\|_2^2 - 2\eta_{k-1}\langle \nabla R(\widehat{w}_{k-1}) - \nabla R(\widehat{w}_{k-1}^{(i)}), \widehat{w}_{k-1} - \widehat{w}_{k-1}^{(i)}\rangle.
\end{aligned}$$

Now we must lower bound $\langle \nabla R(\widehat{w}_{k-1}) - \nabla R(\widehat{w}_{k-1}^{(i)}), \widehat{w}_{k-1} - \widehat{w}_{k-1}^{(i)}\rangle$ utilising both the loss regularity (Assumption 1) and the pointwise weak convexity (Assumption 2). These steps are summarised within the following lemma.

Lemma 1 *Consider assumptions 1 and 2. Then for $s \geq 1$ and $\eta \geq 0$*

$$\begin{aligned}
\langle \nabla R(\widehat{w}_s) - \nabla R(\widehat{w}_s^{(i)}), \widehat{w}_s - \widehat{w}_s^{(i)}\rangle \geq\ & 2\eta\Big(1 - \frac{\beta\eta}{2}\Big)\|\nabla R(\widehat{w}_s) - \nabla R(\widehat{w}_s^{(i)})\|_2^2 \\
& - \big(\epsilon_s + \frac{2\beta}{N}\big)\|\widehat{w}_s - \widehat{w}_s^{(i)} - \eta\big(\nabla R(\widehat{w}_s) - \nabla R(\widehat{w}_s^{(i)})\big)\|_2^2 \\
& - \frac{\rho}{3}\|\widehat{w}_s - \widehat{w}_s^{(i)} - \eta\big(\nabla R(\widehat{w}_s) - \nabla R(\widehat{w}_s^{(i)})\big)\|_2^3
\end{aligned}$$

Utilising Lemma 1 with $s = k-1$ and $\eta = \eta_{k-1}$ the expansiveness of the gradient update term can then be upper bounded

$$\begin{aligned}
\|\widehat{w}_{k-1} - \widehat{w}_{k-1}^{(i)} - \eta_{k-1}\big(\nabla R(\widehat{w}_{k-1}) - \nabla R(\widehat{w}_{k-1}^{(i)})\big)\|_2^2 \leq\ & \|\widehat{w}_{k-1} - \widehat{w}_{k-1}^{(i)}\|_2^2 \\
& + \eta_{k-1}^2\big(1 - 4\big(1 - \frac{\beta\eta_{k-1}}{2}\big)\big)\|\nabla R(\widehat{w}_{k-1}) - \nabla R(\widehat{w}_{k-1}^{(i)})\|_2^2 \\
& + 2\eta_{k-1}\big(\epsilon_{k-1} + \frac{2\beta}{N}\big)\|\widehat{w}_{k-1} - \widehat{w}_{k-1}^{(i)} - \eta_{k-1}\big(\nabla R(\widehat{w}_{k-1}) - \nabla R(\widehat{w}_{k-1}^{(i)})\big)\|_2^2 \\
& + 2\eta_{k-1}\frac{\rho}{3}\|\widehat{w}_{k-1} - \widehat{w}_{k-1}^{(i)} - \eta_{k-1}\big(\nabla R(\widehat{w}_{k-1}) - \nabla R(\widehat{w}_{k-1}^{(i)})\big)\|_2^3.
\end{aligned}$$

Note from assumptions within the theorem that $\eta_{k-1} \leq \frac{3}{2\beta}$ so the second term on the right hand side is negative. Meanwhile, if we denote $\Delta(k) = \widehat{w}_{k-1} - \widehat{w}_{k-1}^{(i)} - \eta_{k-1}\big(\nabla R(\widehat{w}_{k-1}) - \nabla R(\widehat{w}_{k-1}^{(i)})\big)$, the third term can be bounded using Young's inequality $ab \leq \frac{1}{2}a^2 + \frac{1}{2}b^2$ as

$$\begin{aligned}2\eta_{k-1}\frac{\rho}{3}\|\Delta(k)\|_2^3 &= (\sqrt{2}\eta_{k-1}^{\frac{1}{2\alpha}}\|\Delta(k)\|_2)(\eta_{k-1}^{1-\frac{1}{2\alpha}}\frac{\sqrt{2}\rho}{3}\|\Delta(k)\|_2^2)\\ &\leq \eta_{k-1}^{\frac{1}{\alpha}}\|\Delta(k)\|_2^2 + \eta_{k-1}^{2(1-\frac{1}{2\alpha})}\frac{\rho^2}{9}\|\Delta(k)\|_2^4.\end{aligned}$$

Collecting the squared terms and taking on the left hand side, the expansiveness of the gradient update term can then be upper bounded

$$\begin{aligned}(1 - 2\eta_{k-1}(\epsilon_{k-1} + \frac{2\beta}{N}) - \eta_{k-1}^{\frac{1}{\alpha}})\|\Delta(k)\|_2^2 &\leq \|\widehat{w}_{k-1} - \widehat{w}_{k-1}^{(i)}\|_2^2 + \eta^{2(1-\frac{1}{2\alpha})}\frac{\rho^2}{9}\|\Delta(k)\|_2^4\\ &\leq \|\widehat{w}_{k-1} - \widehat{w}_{k-1}^{(i)}\|_2^2 + 9\eta_{k-1}^{2(1-\frac{1}{2\alpha})}\rho^2\|\widehat{w}_{k-1} - \widehat{w}_{k-1}^{(i)}\|_2^4. \qquad (6)\end{aligned}$$

Note that the second inequality above uses the Lipschitz property of the loss's gradient and that $\eta_{k-1}\beta \leq 3/2$ to say

$$\begin{aligned}\|\widehat{w}_{k-1} - \widehat{w}_{k-1}^{(i)} - \eta_{k-1}\big(\nabla R(\widehat{w}_{k-1}) - \nabla R(\widehat{w}_{k-1}^{(i)})\big)\|_2 &\leq (1 + \eta_{k-1}\beta)\|\widehat{w}_{k-1} - \widehat{w}_{k-1}^{(i)}\|_2\\ &\leq 3\|\widehat{w}_{k-1} - \widehat{w}_{k-1}^{(i)}\|_2.\end{aligned}$$

Dividing both sides of (6) by $1 - 2\eta_{k-1}(\epsilon_{k-1} + \frac{2\beta}{N}) - \eta_{k-1}^{\frac{1}{\alpha}}$, taking care this quantity is non-negative from an assumption within the theorem, applying the square root and plugging into (4) then yields the recursion

$$\begin{aligned}\|\widehat{w}_k - \widehat{w}_k^{(i)}\|_2 \leq \Big(\frac{1}{1 - 2\eta_{k-1}(\epsilon_{k-1} + \frac{2\beta}{N}) - \eta_{k-1}^{\frac{1}{\alpha}}}\Big)^{1/2}\|\widehat{w}_{k-1} - \widehat{w}_{k-1}^{(i)}\|_2 + \frac{2\eta_{k-1}L}{N}\\ + 3\eta_{k-1}^{1-\frac{1}{2\alpha}}\rho\Big(\frac{1}{1 - 2\eta_{k-1}(\epsilon_{k-1} + \frac{2\beta}{N}) - \eta_{k-1}^{\frac{1}{\alpha}}}\Big)^{1/2}\|\widehat{w}_{k-1} - \widehat{w}_{k-1}^{(i)}\|_2^2.\end{aligned}$$

Unravelling the iterates with the convention $\prod_{s=k}^{k-1}\big(\frac{1}{1-2\eta_s(\epsilon_s+\frac{2\beta}{N})-\eta_s^{\frac{1}{\alpha}}}\big)^{1/2} = 1$ gives

$$\begin{aligned}\|\widehat{w}_k - \widehat{w}_k^{(i)}\|_2 \leq \frac{2L}{N}\sum_{j=0}^{k-1}\prod_{s=j+1}^{k-1}\big(\frac{1}{1 - 2\eta_s(\epsilon_s + \frac{2\beta}{N}) - \eta_s^{\frac{1}{\alpha}}}\big)^{1/2}\eta_j\\ + 3\rho\sum_{j=1}^{k-1}\prod_{s=j}^{k-1}\big(\frac{1}{1 - 2\eta_s(\epsilon_s + \frac{2\beta}{N}) - \eta_s^{\frac{1}{\alpha}}}\big)^{1/2}\eta_j^{1-\frac{1}{2\alpha}}\|\widehat{w}_j - w_j^{(i)}\|_2^2\end{aligned}$$

We must now bound the product of terms. Using the assumption within the theorem that $2\eta_s(\epsilon_s + \frac{2\beta}{N}) + \eta_s^{\frac{1}{\alpha}} < 1/2$, as well as th inequality $1 + x \leq e^x$, we get the upper bound

$$\begin{aligned}\sum_{j=0}^{k-1}\prod_{s=j+1}^{k-1}\big(\frac{1}{1 - 2\eta_s(\epsilon_s + \frac{2\beta}{N}) - \eta_s^{\frac{1}{\alpha}}}\big)^{1/2}\eta_j &= \sum_{j=0}^{k-1}\prod_{s=j+1}^{k-1}\big(1 + \frac{2\eta_s(\epsilon_s + \frac{2\beta}{N}) + \eta_s^{\frac{1}{\alpha}}}{1 - 2\eta_s(\epsilon_s + \frac{2\beta}{N}) - \eta_s^{\frac{1}{\alpha}}}\big)^{1/2}\eta_j\\ &\leq \sum_{j=0}^{k-1}\exp\big(2\sum_{s=j+1}^{k-1}\eta_s\epsilon_s + \frac{4\sum_{s=j+1}^{k-1}\eta_s\beta}{N} + \sum_{s=j+1}^{k-1}\eta_s^{\frac{1}{\alpha}}\big)\eta_j\end{aligned}$$

where we have adopted the convention $\sum_{s=k}^{k-1}\eta_s = 0$. This then leads to following upper bound for $k \geq 2$

$$\begin{aligned}\|\widehat{w}_k - \widehat{w}_k^{(i)}\|_2 \leq \frac{2L}{N}\sum_{j=0}^{k-1}\exp\big(2\sum_{s=j+1}^{k-1}\eta_s\epsilon_s + \frac{4\sum_{s=j+1}^{k-1}\eta_s\beta}{N} + \sum_{s=j+1}^{k-1}\eta_s^{\frac{1}{\alpha}}\big)\eta_j \qquad (7)\\ + 3\rho\sum_{j=1}^{k-1}\prod_{s=j}^{k-1}\big(\frac{1}{1 - 2\eta_s(\epsilon_s + \frac{2\beta}{N}) - \eta_s^{\frac{1}{\alpha}}}\big)^{1/2}\eta_j^{1-\frac{1}{2\alpha}}\|\widehat{w}_j - w_j^{(i)}\|_2^2.\end{aligned}$$

Observe that the above bound depends on higher order terms from previous time steps $\|\widehat{w}_j - w_j^{(i)}\|_2^2$. To control $\|\widehat{w}_k - \widehat{w}_k^{(i)}\|_2$ for some $k \geq 2$, we now utilise the fact that the difference from earlier iterations $\|\widehat{w}_j - \widehat{w}_j^{(i)}\|_2$ for $j = 1, \ldots, k-1$ can also be small. To this end, we use the upper bounds (5) and (7) to show inductively, under the assumptions of the theorem, that the following holds for $t \geq k \geq 2$

$$\|\widehat{w}_k - \widehat{w}_k^{(i)}\|_2 \leq \frac{4L}{N} \sum_{j=0}^{k-1} \exp\big(2 \sum_{s=j+1}^{k-1} \eta_s \epsilon_s + \frac{4\sum_{s=j+1}^{k-1} \eta_s \beta}{N} + \sum_{s=j+1}^{k-1} \eta_s^{\frac{1}{\alpha}}\big)\eta_j.$$

Showing the above would then imply the bound presented within the theorem. Let us begin by proving the base case $k = 2$. Looking to (7) and plugging in the upper bound on $\|\widehat{w}_1 - \widehat{w}_1^{(i)}\|_2 \leq 2\eta_0 L/N$ from (5) yields

$$\begin{aligned}
\|\widehat{w}_2 - \widehat{w}_2^{(i)}\|_2 &\leq \frac{2L}{N} \sum_{j=0}^{1} \exp\big(2 \sum_{s=j+1}^{1} \eta_s \epsilon_s + \frac{4\sum_{s=j+1}^{1} \eta_s \beta}{N} + \sum_{s=j+1}^{1} \eta_s^{\frac{1}{\alpha}}\big)\eta_j \\
&\qquad + 3\rho\big(\frac{1}{1 - 2\eta_1(\epsilon_1 + \frac{2\beta}{N}) - \eta_1^{\frac{1}{\alpha}}}\big)^{1/2} \eta_1^{1-\frac{1}{2\alpha}} \frac{4\eta_0^2 L^2}{N^2} \\
&= \frac{2L}{N} \exp\big(2\eta_1\epsilon_1 + \frac{4\eta_1\beta}{N} + \eta_1^{\frac{1}{\alpha}}\big)\eta_0 + \frac{2L\eta_1}{N} + 3\rho\big(\frac{1}{1 - 2\eta_1(\epsilon_1 + \frac{2\beta}{N}) - \eta_1^{\frac{1}{\alpha}}}\big)^{1/2} \eta_1^{1-\frac{1}{2\alpha}} \frac{4\eta_0^2 L^2}{N^2}.
\end{aligned}$$

Note from the assumption within the theorem $2\eta_1(\epsilon_1 + \frac{2\beta}{N}) - \eta_1^{\frac{1}{\alpha}} \leq 1/2$ the third term can be upper bounded in a similar manner to previously

$$\big(\frac{1}{1 - 2\eta_1(\epsilon_1 + \frac{2\beta}{N}) - \eta_1^{\frac{1}{\alpha}}}\big)^{1/2} \eta_1^{1-\frac{1}{2\alpha}} \frac{4\eta_0^2 L^2}{N^2} \leq \exp(2\eta_1(\epsilon_1 + \frac{2\beta}{N}) + \eta_1^{\frac{1}{\alpha}}) \eta_1^{1-\frac{1}{2\alpha}} \frac{4\eta_0^2 L^2}{N^2},$$

and thus

$$\|\widehat{w}_2 - \widehat{w}_2^{(i)}\|_2 \leq \frac{2L}{N} \exp\big(2\eta_1\epsilon_1 + \frac{4\eta_1\beta}{N} + \eta_1^{\frac{1}{\alpha}}\big)\eta_0 \Big(1 + \underbrace{6\rho\eta_1^{1-\frac{1}{2\alpha}} \frac{\eta_0 L}{N}}_{\textbf{Remainder Term}}\Big) + \frac{2L\eta_1}{N}.$$

It is then sufficient to show **Remainder Term** ≤ 1 for the base case to hold. Note that this is then implied by the condition on the sample size within the theorem, namely that

$$\begin{aligned}
N &\geq 24\rho L \exp\Big(2\sum_{s=1}^{t} \eta_s\big(\epsilon_s + \frac{4\beta}{N}\big) + \eta_s^{\frac{1}{\alpha}}\Big) \sum_{j=1}^{t} \eta_j^{1-\frac{1}{2\alpha}} \sum_{\ell=0}^{j-1} \eta_\ell \\
&\geq 6\rho L \eta_1^{1-\frac{1}{2\alpha}} \eta_0.
\end{aligned}$$

Let us now assume the inductive hypothesis holds up-to k and consider the case $k+1$. Utilising the inductive hypothesis for $u = 1, \ldots, k$ as well as multiplying and dividing by $\big(\sum_{j=0}^{u-1} \eta_j\big)^2$ allows the squared deviation to be bounded

$$\begin{aligned}
\|\widehat{w}_u - \widehat{w}_u^{(i)}\|_2^2 &\leq \big(\sum_{j=0}^{u-1} \eta_j\big)^2 \Big(\frac{4L}{N} \sum_{j=0}^{u-1} \exp\big(2 \sum_{s=j+1}^{u-1} \eta_s \epsilon_s + \frac{4\sum_{s=j+1}^{u-1} \eta_s \beta}{N} + \sum_{s=j+1}^{u-1} \eta_s^{\frac{1}{\alpha}}\big) \frac{\eta_j}{\sum_{j=0}^{u-1} \eta_j}\Big)^2 \\
&\leq \big(\sum_{j=0}^{u-1} \eta_j\big) \frac{16L^2}{N^2} \sum_{j=0}^{u-1} \eta_j \exp\big(4 \sum_{s=j+1}^{u-1} \eta_s \epsilon_s + \frac{8\sum_{s=j+1}^{u-1} \eta_s \beta}{N} + 2 \sum_{s=j+1}^{u-1} \eta_s^{\frac{1}{\alpha}}\big) \\
&\leq \big(\sum_{j=0}^{u-1} \eta_j\big) \frac{16L^2}{N^2} \exp\big(2 \sum_{s=1}^{u-1} \eta_s \epsilon_s + \frac{4\sum_{s=1}^{u-1} \eta_s \beta}{N} + \sum_{s=1}^{u-1} \eta_s^{\frac{1}{\alpha}}\big) \\
&\qquad \times \Big(\sum_{j=0}^{k} \eta_j \exp\big(2 \sum_{s=j+1}^{k} \eta_s \epsilon_s + \frac{4\sum_{s=j+1}^{k} \eta_s \beta}{N} + \sum_{s=j+1}^{k} \eta_s^{\frac{1}{\alpha}}\big)\Big),
\end{aligned}$$

where we note the second inequality arises from convexity of the squared function, and the third inequality from adding positive terms within the exponentials. Plugging the above into the squared terms of (7) for $u = 1, \dots, k$, as well as factoring out $\frac{2L}{N}\sum_{j=0}^{k} \exp\big(2\sum_{s=j+1}^{k}\eta_s\epsilon_s + \frac{4\sum_{s=j+1}^{k}\eta_s\beta}{N} + \sum_{s=j+1}^{k}\eta_s^{\frac{1}{\alpha}}\big)\eta_j$ allows the deviation at time $k+1$ to be bounded

$$\|\widehat{w}_{k+1} - \widehat{w}_{k+1}^{(i)}\|_2 \leq \Big(\frac{2L}{N}\sum_{j=0}^{k}\exp\big(2\sum_{s=j+1}^{k}\eta_s\epsilon_s + \frac{4\sum_{s=j+1}^{k}\eta_s\beta}{N} + \sum_{s=j+1}^{k}\eta_s^{\frac{1}{\alpha}}\big)\eta_j\Big)$$
$$\times \Big(\underbrace{1+\frac{24\rho L}{N}\sum_{j=1}^{k}\prod_{s=j}^{k}\Big(\frac{1}{1-2\eta_s(\epsilon_s+\frac{2\beta}{N})-\eta_s^{\frac{1}{\alpha}}}\Big)^{1/2}\eta_j^{1-\frac{1}{2\alpha}}\Big(\sum_{\ell=0}^{j-1}\eta_\ell\Big)\exp\big(2\sum_{s=1}^{j-1}\eta_s\epsilon_s+\frac{4\sum_{s=1}^{j-1}\eta_s\beta}{N}+\sum_{s=1}^{j-1}\eta_s^{\frac{1}{\alpha}}\big)}_{\textbf{Remainder Term}}\Big).$$

To prove the inductive hypothesis holds for $k+1$ we must show that **Remainder Term** ≤ 1. To this end, follow the previous steps to bound the product of terms for $j = 1, \dots, k$

$$\prod_{s=j}^{k}\Big(\frac{1}{1-2\eta_s(\epsilon_s+\frac{2\beta}{N})-\eta_s^{\frac{1}{\alpha}}}\Big)^{1/2} \leq \exp\big(2\sum_{s=j}^{k}\eta_s\epsilon_s + \frac{4\sum_{s=j}^{k}\eta_s\beta}{N} + \sum_{s=j}^{k}\eta_s^{\frac{1}{\alpha}}\big)$$

to upper bound the **Remainder Term**

$$\textbf{Remainder Term} \leq \frac{24\rho L}{N}\sum_{j=1}^{k}\eta_j^{1-\frac{1}{2\alpha}}\Big(\sum_{\ell=0}^{j-1}\eta_\ell\Big)\exp\big(2\sum_{s=1}^{k}\eta_s\epsilon_s + \frac{4\sum_{s=1}^{k}\eta_s\beta}{N} + \sum_{s=1}^{k}\eta_s^{\frac{1}{\alpha}}.\big)$$

Following the assumption with the theorem, we then have when $k \leq t$

$$N \geq 24\rho L\sum_{j=1}^{t}\eta_j^{1-\frac{1}{2\alpha}}\Big(\sum_{\ell=0}^{j-1}\eta_\ell\Big)\exp\big(2\sum_{s=1}^{t}\eta_s\epsilon_s + \frac{4\sum_{s=1}^{t}\eta_s\beta}{N} + \sum_{s=1}^{t}\eta_s^{\frac{1}{\alpha}}\big)$$
$$\geq 24\rho L\sum_{j=1}^{k}\eta_j^{1-\frac{1}{2\alpha}}\Big(\sum_{\ell=0}^{j-1}\eta_\ell\Big)\exp\big(2\sum_{s=1}^{k}\eta_s\epsilon_s + \frac{4\sum_{s=1}^{k}\eta_s\beta}{N} + \sum_{s=1}^{k}\eta_s^{\frac{1}{\alpha}}\big)$$

and therefore **Remainder Term** ≤ 1 as required and the inductive hypothesis holds. This completes the proof.

B.2 Proof of Lemma 1

In this section we give the proof of Lemma 1. We begin recalling a function f is β-smooth if for any x, y we have

$$|f(y) - f(x) - \langle\nabla f(x), y - x\rangle| \leq \frac{\beta}{2}\|x - y\|_2^2$$

Moreover, to save on notational burden let us denote $x = \widehat{w}_s$ and $y = \widehat{w}_s^{(i)}$.

We begin by following the standard proof for showing co-coercivity in the smooth convex setting, see for instance (Nesterov, 2013). Let us define the functions $\phi_x(\omega) = R(\omega) - \langle R(x), \omega\rangle$ and $\phi_y(\omega) = R(\omega) - \langle R(y), \omega\rangle$. It is then clear that ϕ_x, ϕ_y are both β-smooth. As such using smoothness we get

$$\phi_x(y - \eta\nabla\phi_x(y)) \leq \phi_x(y) + \langle\nabla\phi_x(y), y - \eta\nabla\phi(y) - y\rangle + \frac{\beta\eta^2}{2}\|\nabla\phi_x(y)\|_2^2$$
$$= \phi_x(y) + \eta\big(\frac{\eta\beta}{2} - 1\big)\|\nabla\phi_x(y)\|_2^2$$

Plugging in the definition of ϕ_x, and repeating the steps for with x, y swapped we get the two inequalities

$$R(y - \eta\big(\nabla R(y) - \nabla R(x)\big)) - \langle\nabla R(x), y - \eta\big(\nabla R(y) - \nabla R(x)\big)\rangle \leq R(y) - \langle R(x), y\rangle$$
$$+ \eta\big(\frac{\eta\beta}{2} - 1\big)\|\nabla\phi_x(y)\|_2^2$$
$$R(x - \eta\big(\nabla R(x) - \nabla R(y)\big)) - \langle\nabla R(y), x - \eta\big(\nabla R(x) - \nabla R(y)\big)\rangle \leq R(x) - \langle R(y), x\rangle$$
$$+ \eta\big(\frac{\eta\beta}{2} - 1\big)\|\nabla\phi_y(x)\|_2^2$$

We would now like to lower bound the left side of each of these inequalities. In the convex setting this is immediate from the definition of convexity. In our case, we wish to use the local convexity of the iterates $\widehat{w}_s$ on the objective $\nabla R(\cdot)$. We begin with lower bounding the left side of the first inequality.

Lower Bounding First Inequality Let us denote $\widetilde{y} = y - \eta\big(\nabla R(y) - \nabla R(x)\big)$. We then must lower bound

$$R(\widetilde{y}) - \langle \nabla R(x), \widetilde{y} \rangle.$$

Recall from Assumption 2 that the Hessian $\nabla^2 R(\cdot)$ has minimum Eigenvalue lower bounded by $-\epsilon_s$ at the point $x = \widehat{\omega}_s$. To utilise this let us define the function for $\alpha \in [0,1]$

$$g(\alpha) = R(x + \alpha(\widetilde{y} - x)) + \frac{\epsilon_s}{2}\|x + \alpha(\widetilde{y} - x)\|_2^2 + \frac{\alpha^3}{6}\rho\|x - \widetilde{y}\|_2^3.$$

Taking derivatives of g with respect to α we observe that

$$\begin{aligned}
g''(\alpha) &= (\widetilde{y} - x)^\top \nabla^2 R(x + \alpha(\widetilde{y} - x))(\widetilde{y} - x) + \epsilon_s\|\widetilde{y} - x\|_2^2 + \alpha\rho\|x - \widetilde{y}\|_2^3 \\
&= (\widetilde{y} - x)^\top (\nabla^2 R(x) + \epsilon_s I)(\widetilde{y} - x) \\
&\qquad + (\widetilde{y} - x)^\top (\nabla^2 R(x + \alpha(\widetilde{y} - x)) - \nabla^2 R(x))(\widetilde{y} - x) + \alpha\rho\|x - \widetilde{y}\|_2^3 \\
&\geq 0 - \|\widetilde{y} - x\|_2^2 \|\nabla^2 R(x + \alpha(\widetilde{y} - x)) - \nabla^2 R(x)\|_2 + \alpha\rho\|x - \widetilde{y}\|_2^3 \\
&\geq 0
\end{aligned}$$

where we have added and subtracted $(\widetilde{y} - x)^\top \nabla^2 R(x)(\widetilde{y} - x)$ on the second equality, and note that the first term is lower bounded from the pointwise weak convexity $\nabla^2 R(x) \succeq -\epsilon_s I$.

Therefore g is convex in $\alpha \in [0,1]$, and thus for $\alpha' \in [0,1]$ we get $0 \leq g(\alpha) - g(\alpha') - g'(\alpha')(\alpha - \alpha')$. Picking $\alpha = 1$ and $\alpha' = 0$ and plugging in the definition of g we have

$$\begin{aligned}
0 &\leq R(\widetilde{y}) + \frac{\epsilon_s}{2}\|\widetilde{y}\|_2^2 + \frac{\rho}{6}\|x - \widetilde{y}\|_2^3 - R(x) - \frac{\epsilon_s}{2}\|x\|_2^2 - (\widetilde{y} - x)^\top(\nabla R(x) + \epsilon_s x) \\
&= R(\widetilde{y}) - R(x) - \langle \nabla R(x), \widetilde{y} - x \rangle + \frac{\epsilon_s}{2}\|x - \widetilde{y}\|_2^2 + \frac{\rho}{6}\|x - \widetilde{y}\|_2^3.
\end{aligned}$$

Rearranging the above then results in the lower bound

$$R(\widetilde{y}) - \langle \nabla R(x), \widetilde{y} \rangle \geq R(x) - \langle \nabla R(x), x \rangle - \frac{\epsilon_s}{2}\|x - \widetilde{y}\|_2^2 - \frac{\rho}{6}\|x - \widetilde{y}\|_2^3,$$

which will be the lower bound that we will use.

Lower Bounding Second Inequality With $\widetilde{x} = x - \eta(\nabla R(x) - \nabla R(y))$ we now lower bound

$$R(\widetilde{x}) - \langle \nabla R(y), \widetilde{x} \rangle$$

This lower bound will be slightly more technical as the minimum Eigenvalue of $\nabla^2 R(y) = \nabla^2 R(\widehat{w}_s^{(i)})$ is not immediately lower bounded from our assumptions. Although, note that we have for any vector $v \in \mathbb{R}^p$ that

$$\begin{aligned}
v^\top \nabla^2 R(y) v &= v^\top \nabla^2 R^{(i)}(y) v + v^\top \nabla^2 R^{(i)}(y) - \nabla^2 R(y) v \\
&= v^\top \nabla^2 R^{(i)}(y) v + \frac{1}{N} v^\top \big(\nabla^2 \ell(y, Z_i') - \nabla^2 \ell(y, Z_i)\big) v \\
&\geq -\Big(\epsilon_s + \frac{2\beta}{N}\Big)\|v\|_2^2
\end{aligned}$$

Therefore $\nabla^2 R(y)$ has minimum Eigenvalue lower bounded by $-\big(\epsilon_s + \frac{2\beta}{N}\big)$. Following an identical set of arguments to the lower bound for the first inequality with ϵ_s swapped with $\epsilon_s + \frac{2\beta}{N}$ we then get

$$R(\widetilde{x}) - \langle \nabla R(y), \widetilde{x} \rangle \geq R(y) - \langle \nabla R(y), y \rangle - \frac{\epsilon_s + 2\beta/N}{2}\|y - \widetilde{x}\|_2^2 - \frac{\rho}{6}\|\widetilde{x} - y\|_2^3$$

Using Lower Bounds Given the two lower bounds we arrive at after rearranging

$$\langle \nabla R(x), x-y\rangle \geq R(x)-R(y)+\eta\big(1-\frac{\eta\beta}{2}\big)\|\nabla\phi_x(y)\|_2^2-\frac{\epsilon_s}{2}\|x-\widetilde{y}\|_2^2-\frac{\rho}{6}\|x-\widetilde{y}\|_2^3$$
$$\langle \nabla R(y), y-x\rangle \geq R(y)-R(x)+\eta\big(1-\frac{\eta\beta}{2}\big)\|\nabla\phi_y(x)\|_2^2-\frac{\epsilon_s+2\beta/N}{2}\|y-\widetilde{x}\|_2^2-\frac{\rho}{6}\|\widetilde{x}-y\|_2^3$$

The result is then arrived at by adding together the two above bounds and substituting in the definitions of $\phi_x, \phi_y, \widetilde{x}, \widetilde{y}, x, y$.

B.3 Generalisation Error Bound for Gradient Descent under Standard Weak Convexity

In this section we present and prove a generalisation error bound of gradient descent under standard weak convexity Assumption 3. The following theorem presents the generalisation error bound.

Theorem 5 (Generalisation Error Bound Standard Weak Convexity) *Consider Assumptions 1 and 3. If $\eta\beta \leq 3/2$ and $2\eta\epsilon < 1$, then the generalisation error of gradient descent satisfies*

$$\mathbf{E}[R(\widehat{w}_t)-r(\widehat{w}_t)] \leq \frac{2\eta L^2}{N}\sum_{k=0}^{t-1}\exp\big(\frac{\eta\epsilon k}{1-2\eta\epsilon}\big)$$

The proof of Theorem 5 closely follows the steps in proving Theorem 1. Therefore in the proof we focus on the key differences. We begin with the following lemma which lower bounds the co-coercivity of weakly convex losses.

Lemma 2 *Consider Assumptions 1 and 3. Then for $\eta \geq 0$ and $x, y \in \mathbb{R}^d$*

$$\langle \nabla R(x)-\nabla R(y), x-y\rangle \geq 2\eta\Big(1-\frac{\beta\eta}{2}\Big)\|\nabla R(x)-\nabla R(y)\|_2^2-\epsilon\|x-y-\eta\big(\nabla R(x)-\nabla R(y)\big)\|_2^2$$

We now provide the proof of this Lemma.

Proof 1 (Lemma 2) *Follow the proof of Lemma 1 to the point of lower bounding $R(\widetilde{y})-\langle \nabla R(x), \widetilde{y}\rangle$ where $\widetilde{y}=y-\eta\big(\nabla R(y)-\nabla R(x)\big)$. Now, let us alternatively choose*

$$g(\alpha)=R(x+\alpha(\widetilde{y}-x))+\frac{\epsilon}{2}\|x+\alpha(\widetilde{y}-x)\|_2^2$$

We then immediate see that $g''(\alpha)=(\widetilde{y}-x)^\top\nabla^2 R(x+\alpha(\widetilde{y}-x))(\widetilde{y}-x)+\epsilon\|\widetilde{y}-x\|_2^2 \geq 0$ since Assumption 3 states that $\nabla^2 R(\omega) \succeq -\epsilon I$ for every $\omega \in \mathbb{R}^d$. Therefore $g(\alpha)$ is convex on $\alpha \in [0,1]$. Using that $0 \leq g(1)-g(0)-g'(0)$ and rearranging we get the lower bound

$$R(\widetilde{y})-\langle R(x), \widetilde{y}\rangle \geq R(x)-\langle \nabla R(x), x\rangle-\frac{\epsilon}{2}\|x-\widetilde{y}\|_2^2.$$

Performing an identical set of steps for $R(\widetilde{x})-\langle R(y), \widetilde{x}\rangle$ where $\widetilde{x}=x-\eta\big(\nabla R(x)-\nabla R(y)\big)$ yields the lower bound

$$R(\widetilde{x})-\langle R(y), \widetilde{x}\rangle \geq R(y)-\langle \nabla R(y), y\rangle-\frac{\epsilon}{2}\|\widetilde{x}-y\|_2^2.$$

Following the steps in Lemma 1 then yields the result.

Given this proof, we now provide the proof of Theorem 5.

Proof 2 (Theorem 5) *Let us begin by bounding the expansiveness of the gradient update. Note for $3/(2\beta) \geq \eta \geq 0$ we have when using Lemma 2*

$$\begin{aligned}\|x-y-\eta\big(\nabla R(x)-\nabla R(y)\big)\|_2^2 &= \|x-y\|_2^2+\eta^2\|\nabla R(x)-\nabla R(y)\|_2^2-2\eta\langle x-y, \nabla R(x)-\nabla R(y)\rangle\\ &\leq \|x-y\|_2^2+\eta^2\big(1-4\big(1-\frac{\eta\beta}{2}\big)\big)\|\nabla R(x)-\nabla R(y)\|_2^2\\ &\quad+2\eta\epsilon\|x-y-\eta\big(\nabla R(x)-\nabla R(y)\big)\|_2^2\\ &\leq \|x-y\|_2^2+2\eta\epsilon\|x-y-\eta\big(\nabla R(x)-\nabla R(y)\big)\|_2^2\end{aligned}$$

It is then clear that the expansiveness of the gradient update can be upper bounded

$$\|x - y - \eta\big(\nabla R(x) - \nabla R(y)\big)\|_2 \leq \frac{1}{\sqrt{1-2\eta\epsilon}}\|x-y\|_2$$

Following the steps in the proof of Theorem 1 (similarly (Hardt et al., 2016)) we immediately get

$$\begin{aligned}\|\widehat{\omega}_t - \widehat{\omega}_t^{(i)}\|_2 &\leq \frac{1}{\sqrt{1-2\eta\epsilon}}\|\widehat{\omega}_{t-1} - \widehat{\omega}_{t-1}^{(i)}\|_2 + \frac{2\eta L}{N} \\ &\leq \frac{2\eta L}{N}\sum_{k=0}^{t-1}\Big(\frac{1}{\sqrt{1-2\eta\epsilon}}\Big)^k \\ &\leq \frac{2\eta L}{N}\sum_{k=0}^{t-1}\exp\big(\frac{\eta\epsilon k}{1-2\eta\epsilon}\big)\end{aligned}$$

where we have used that $1/(1-u) = 1 + \frac{u}{1-u} \leq e^{\frac{u}{1-u}}$. *This yields the result.*

C Proofs of Test Error Bounds for General Loss Functions

In this section we present proofs related to test error bounds for general loss functions. This section proceeds as follows. Section C.1 presents the proof of Theorem 2, which gives a test error bound for weakly convex losses. Section C.2 presents the proof of Proposition 1 which is presented within Appendix A. Section C.3 gives the proofs of additional lemmas used within this section.

C.1 Proof of Test Error Bounds for Weakly Convex Losses (Theorem 2)

In this section we present the proof of Theorem 2. We begin by introducing the penalised population objective $r_\epsilon(\omega) := r(\omega) + \frac{\epsilon}{2}\|\widehat{\omega}_0 - \omega\|_2^2$ as well as one of its minimiser $\omega_\epsilon^\star \in \operatorname{argmin}_\omega r_\epsilon(\omega)$. The test error is then decomposed as follows

$$\begin{aligned}&\mathbf{E}_I[\mathbf{E}[r(\widehat{\omega}_I)]] - \min_\omega r(\omega) \\ &= \mathbf{E}_I[\mathbf{E}[r(\widehat{\omega}_I) - R(\widehat{\omega}_I)]] + \mathbf{E}[\mathbf{E}_I[R(\widehat{\omega}_I)] - R(\widehat{\omega}_\epsilon^\star)] + \mathbf{E}[R(\widehat{\omega}_\epsilon^\star)] - r_\epsilon(\omega_\epsilon^\star) + r_\epsilon(\omega_\epsilon^\star) - \min_\omega r(\omega) \\ &\leq \underbrace{\max_{k=1,\ldots,t}\big\{\mathbf{E}[r(\widehat{\omega}_k) - R(\widehat{\omega}_k)]\big\}}_{\textbf{Generalisation Error}} + \underbrace{\mathbf{E}[\mathbf{E}_I[R(\widehat{\omega}_I)] - R(\widehat{\omega}_\epsilon^\star)]}_{\textbf{Term 1}} + \underbrace{\mathbf{E}[R(\widehat{\omega}_\epsilon^\star)] - r_\epsilon(\omega_\epsilon^\star) + r_\epsilon(\omega_\epsilon^\star) - \min_\omega r(\omega)}_{\textbf{Term 2}}\end{aligned}$$

which has three terms. The **Generalisation Error** is bounded by Theorem 5 within Appendix B.3, since we assume standard weak convexity. Note the **Generalisation Error** term above depends upon the maximum from $k = 1, \ldots, t$ therefore, apply the bound within Theorem 1 for each instance $k = 1, \ldots, t$ and take the largest at $k = t$. We now set out to bound **Term 1** and **Term 2**, beginning with **Term 1**. To do so, we introduce the following lemma which is proved in Section C.3.

Lemma 3 *Suppose assumption 1 and 3 hold,* $\eta_s = \eta$ *for all* $s \geq 0$ *and* $\eta\beta \leq 1/2$. *Then*

$$\mathbf{E}_I[R(\widehat{\omega}_I)] = \frac{1}{t}\sum_{s=0}^{t-1} R(\widehat{\omega}_{s+1}) \leq R(\widehat{\omega}_\epsilon^\star) + \frac{\|\widehat{\omega}_0 - \widehat{\omega}_\epsilon^\star\|_2^2}{2\eta t} + \frac{\epsilon}{2}\frac{1}{t}\sum_{s=0}^{t-1}\|\widehat{\omega}_{s+1} - \widehat{\omega}_\epsilon^\star\|_2^2$$

and $\|\widehat{\omega}_{s+1} - \widehat{\omega}_0\|_2 \leq \sqrt{2\eta s(R(\widehat{\omega}_0) - R(\widehat{\omega}^\star))}$ *for* $s \geq 0$.

Note for $t-1 \geq s \geq 0$ that the deviation $\|\widehat{\omega}_{s+1} - \widehat{\omega}_\epsilon^\star\|_2^2$ can be bounded by adding and subtracting the initialisation $\widehat{\omega}_0$ and applying the upper bound on $\|\widehat{\omega}_{s+1} - \widehat{\omega}_0\|_2$ from Lemma 3. Specifically, using triangle inequality alongside that $(a+b)^2 \leq 2a^2 + 2b^2$ for $a, b \in \mathbb{R}$ gives

$$\begin{aligned}\|\widehat{\omega}_{s+1} - \omega_\epsilon^\star\|_2^2 &\leq 2\|\widehat{\omega}_{s+1} - \widehat{\omega}_0\|_2^2 + 2\|\widehat{\omega}_0 - \widehat{\omega}_\epsilon^\star\|_2^2 \\ &\leq 2\big(\eta t(R(\widehat{\omega}_0) - R(\widehat{\omega}^\star)) + 2\|\widehat{\omega}_0 - \widehat{\omega}_\epsilon^\star\|_2^2.\end{aligned}$$

Combining this with the first part of Lemma 3 and taking expectation then gives the following upper bound for **Term 1**

$$\textbf{Term 1} \leq \frac{\mathbf{E}[\|\omega_0 - \widehat{\omega}_\epsilon^\star\|_2^2]}{2\eta t} + \epsilon\Big(\eta t \mathbf{E}[R(\widehat{\omega}_0) - R(\widehat{\omega}^\star)] + \mathbf{E}[\|\widehat{\omega}_0 - \widehat{\omega}_\epsilon^\star\|_2^2]\Big).$$

To bound **Term 2** begin by noting that we have the lower bound

$$\begin{aligned} r_\epsilon(\omega_\epsilon^\star) &= \mathbf{E}[\ell(\omega_\epsilon^\star, Z) + \epsilon\|\widehat{\omega}_0 - \omega_\epsilon^\star\|_2^2] \\ &= \mathbf{E}\Big[\frac{1}{N}\sum_{i=1}^N \ell(\omega_\epsilon^\star, Z_i) + \epsilon\|\widehat{\omega}_0 - \omega_\epsilon^\star\|_2^2\Big] \\ &= \mathbf{E}[R(\omega_\epsilon^\star) + \epsilon\|\widehat{\omega}_0 - \omega_\epsilon^\star\|_2^2] \\ &\geq \mathbf{E}[R(\widehat{\omega}_\epsilon^\star) + \epsilon\|\widehat{\omega}_0 - \widehat{\omega}_\epsilon^\star\|_2^2], \end{aligned}$$

where we recall that $\widehat{\omega}_\epsilon^\star = \text{argmin}_\omega \left(R(\omega) + \epsilon\|\widehat{\omega}_0 - \omega\|_2^2\right)$. Moreover, note that we also have $r_\epsilon(\omega_\epsilon^\star) \leq r_\epsilon(\omega^\star)$ where $\omega^\star \in \text{argmin}_\omega r(\omega)$. By adding and subtracting $\epsilon\mathbf{E}[\|\widehat{\omega}_0 - \widehat{\omega}_\epsilon^\star\|_2^2]$ and using these two facts then yields the following upper bound on **Term 2**

$$\begin{aligned} \textbf{Term 2} &= \underbrace{\mathbf{E}[R(\widehat{\omega}_\epsilon^\star) + \epsilon\|\widehat{\omega}_0 - \widehat{\omega}_\epsilon^\star\|_2^2] - r_\epsilon(\omega_\epsilon^\star)}_{\leq 0} + r_\epsilon(\omega_\epsilon^\star) - r(\omega^\star) - \epsilon\mathbf{E}[\|\widehat{\omega}_0 - \widehat{\omega}_\epsilon^\star\|_2^2] \\ &\leq r_\epsilon(\omega_\epsilon^\star) - r(\omega^\star) - \epsilon\mathbf{E}[\|\widehat{\omega}_0 - \widehat{\omega}_\epsilon^\star\|_2^2] \\ &\leq \epsilon\|\widehat{\omega}_0 - \omega^\star\|_2^2 - \epsilon\mathbf{E}[\|\widehat{\omega}_0 - \widehat{\omega}_\epsilon^\star\|_2^2]. \end{aligned}$$

Bringing together the bounds for the **Generalisation Error**, **Term 1** and **Term 2** and noting that $\epsilon\mathbf{E}[\|\widehat{\omega}_0 - \widehat{\omega}_\epsilon^\star\|_2^2]$ in **Term 1** and **Term 2** cancel, yields the result.

C.2 Proof of Optimisation and Approximation Error Bounds under Generalised Weak Convexity (Proposition 1)

In this section we present the proof of Proposition 1. Following the proof of Theorem 2 in the previous section, we begin with the following decomposition of the **Optimisation & Approximation Error**. Specifically, for $\widetilde{\omega} \in \mathcal{X}$

$$\underbrace{\mathbf{E}_I[E[R(\widehat{\omega}_I]] - r(\omega^\star)}_{\textbf{Opt. \& Approx. Error}} = \underbrace{\mathbf{E}[\mathbf{E}_I[R(\widehat{\omega}_I)] - R(\widetilde{\omega})]}_{\textbf{Term 1}} + \underbrace{\mathbf{E}[R(\widetilde{\omega})] - r(\omega^\star)}_{\textbf{Term 2}}.$$

Where we have labelled the **Term 1** and **Term 2**. We now proceed to bound **Term 1**. To do so we will use the following lemma, which is a generalisation of Lemma 3 given previously.

Lemma 4 *Suppose assumption 1 and 5 hold, $\eta_s = \eta$ for $s \geq 0$ and $\eta\beta \leq 1/2$. Furthermore suppose that $\widehat{\omega}_s \in \mathcal{X}$ for $s \geq 0$. Then for $\widetilde{\omega} \in \mathcal{X}$ we have*

$$\mathbf{E}_I[R(\widehat{\omega}_I)] = \frac{1}{t}\sum_{s=0}^{t-1} R(\widehat{\omega}_{s+1}) \leq R(\widetilde{\omega}) + \frac{\|\widehat{\omega}_0 - \widetilde{\omega}\|_2^2}{2\eta t} + \frac{1}{2}\frac{1}{t}\sum_{s=0}^{t-1} G_{\mathcal{X}}(\widetilde{\omega} - \widehat{\omega}_{s+1})$$

and $\|\widehat{\omega}_{s+1} - \widehat{\omega}_0\|_2 \leq \sqrt{2\eta s(R(\widehat{\omega}_0) - R(\widehat{\omega}^\star))}$ for $s \geq 0$.

Plugging the bound from the first part of Lemma 4 and taking expectation gives the result.

C.3 Proof of Lemma 3 and 4

In this section we provide the proof of both Lemma 3 and 4. We note that Lemma 4 holds under a weaker assumption than Lemma 3. We therefore begin with the proof of Lemma 4, with the proof of Lemma 3 given after.

Proof 3 (Lemma 4) *Recall that $R(\cdot)$ is β-smooth, and therefore, using smoothness and the first iteration of gradient descent for $s \geq 0$ i.e. $\widehat{\omega}_{s+1} = \widehat{\omega}_s - \eta \nabla R(\widehat{\omega}_s)$ we get*

$$\begin{aligned} R(\widehat{\omega}_{s+1}) &\leq R(\widehat{\omega}_s) + \langle \nabla R(\widehat{\omega}_s), \widehat{\omega}_{s+1} - \widehat{\omega}_s \rangle + \frac{\beta}{2} \|\widehat{\omega}_{s+1} - \widehat{\omega}_s\|_2^2 \\ &= R(\widehat{\omega}_s) - \eta \|\nabla R(\widehat{\omega}_s)\|_2^2 + \frac{\eta^2 \beta}{2} \|\nabla R(\widehat{\omega}_s)\|_2^2 \\ &\leq R(\widehat{\omega}_s) - \frac{\eta}{2} \|\nabla R(\widehat{\omega}_s)\|_2^2 \end{aligned} \tag{8}$$

where at the end we used that $\eta\beta \leq 1$. We now wish to upper bound $R(\widehat{\omega}_s)$, for which we will use our Assumption 5. In particular, let us define the function for $\alpha \in [0,1]$ as

$$g(\alpha) = R(\widehat{\omega}_s + \alpha(\widetilde{\omega} - \widehat{\omega}_s)) + \frac{\alpha^2}{2} G_{\mathcal{X}}(\widetilde{\omega} - \widehat{\omega}_s).$$

Differentiating with respect to α twice we get

$$\begin{aligned} g'(\alpha) &= (\widetilde{\omega} - \widehat{\omega}_s)^\top \nabla R(\widehat{\omega}_s + \alpha(\widetilde{\omega} - \widehat{\omega}_s)) + \alpha G_{\mathcal{X}}(\widetilde{\omega} - \widehat{\omega}_s) \\ g''(\alpha) &= (\widetilde{\omega} - \widehat{\omega}_t)^\top \nabla^2 R(\widehat{\omega}_s + \alpha(\widetilde{\omega} - \widehat{\omega}_s))(\widetilde{\omega} - \widehat{\omega}_s) + G_{\mathcal{X}}(\widetilde{\omega} - \widehat{\omega}_s). \end{aligned}$$

Observe that $g''(\alpha) \geq 0$ for $\alpha \in [0,1]$ from Assumption 5. That is $\widehat{\omega}_s + \alpha(\widetilde{\omega} - \widehat{\omega}_s) \in \mathcal{X}$, since both $\widehat{\omega}_s, \widetilde{\omega} \in \mathcal{X}$ and therefore so is the linear combination $(1-\alpha)\widehat{\omega}_s + \alpha\widetilde{\omega}$ provided $\alpha \in [0,1]$ by the convexity of the set $\mathcal{X}$. Since $g(\alpha)$ is convex on $\alpha \in [0,1]$ we then have the inequality $g(0) \leq g(1) - g'(0)$. Plugging in the definition of $g(\cdot)$ into this then gives

$$R(\widehat{\omega}_s) \leq R(\widetilde{\omega}) + \frac{1}{2} G_{\mathcal{X}}(\widetilde{\omega} - \widehat{\omega}_s) - (\widetilde{\omega} - \widehat{\omega}_s)^\top \nabla R(\widehat{\omega}_s).$$

Using this upper bound with (8) and following the standard steps for proving the convergence of gradient descent yields

$$\begin{aligned} R(\widehat{\omega}_{s+1}) &\leq R(\widetilde{\omega}) + \nabla R(\widehat{\omega}_s)^\top (\widehat{\omega}_s - \widetilde{\omega}) - \frac{\eta}{2} \|\nabla R(\widehat{\omega}_s)\|_2^2 + \frac{1}{2} G_{\mathcal{X}}(\widetilde{\omega} - \widehat{\omega}_s) \\ &= R(\widetilde{\omega}) + \frac{1}{\eta} \big(\widehat{\omega}_s - \widehat{\omega}_{s+1}\big)^\top (\widehat{\omega}_s - \widetilde{\omega}) - \frac{1}{2\eta} \|\widehat{\omega}_{s+1} - \widehat{\omega}_s\|_2^2 + \frac{1}{2} G_{\mathcal{X}}(\widetilde{\omega} - \widehat{\omega}_s) \\ &= R(\widetilde{\omega}) + \frac{1}{2\eta} \big(\|\widehat{\omega}_s - \widetilde{\omega}\|_2^2 - \|\widehat{\omega}_{s+1} - \widetilde{\omega}\|_2^2\big) + \frac{1}{2} G_{\mathcal{X}}(\widetilde{\omega} - \widehat{\omega}_s). \end{aligned}$$

Summing up this bound for $s = 0, \ldots, t-1$ and diving by t gives

$$\mathbf{E}_I[R(\omega_I)] = \frac{1}{t} \sum_{s=0}^{t-1} R(\widehat{\omega}_{s+1}) \leq R(\widetilde{\omega}) + \frac{\|\widehat{\omega}_0 - \widetilde{\omega}\|_2^2}{2\eta t} + \frac{1}{2} \frac{1}{t} \sum_{s=0}^{t-1} G_{\mathcal{X}}(\widetilde{\omega} - \widehat{\omega}_{s+1}),$$

which proves the first part of the lemma. To show the second part of the lemma, upper bounding $\|\widehat{\omega}_s - \widehat{\omega}_0\|_2$, we look back to (8). Recalling that $\eta \nabla R(\widehat{\omega}_s) = \widehat{\omega}_s - \widehat{\omega}_{s+1}$, and rearranging we get the upper bound

$$\|\widehat{\omega}_{s+1} - \widehat{\omega}_s\|_2^2 \leq 2\eta \big(R(\widehat{\omega}_s) - R(\widehat{\omega}_{s+1})\big).$$

Summing up both sides yields

$$\sum_{s=0}^{t-1} \|\widehat{\omega}_{s+1} - \widehat{\omega}_s\|_2^2 \leq \eta \big(R(\widehat{\omega}_0) - R(\widehat{\omega}_t)\big) \leq \eta \big(R(\widehat{\omega}_0) - R(\widehat{\omega}^\star)\big),$$

where we have plugged in a minimiser $\widehat{\omega}^\star \in \operatorname{argmin}_\omega R(\omega)$. Note that convexity of the squared norm $\|\cdot\|_2^2$ gives us the lower bound $\frac{1}{t}\sum_{s=0}^t \|\widehat{\omega}_{s+1} - \widehat{\omega}_s\|_2^2 \geq \|\frac{1}{t}\sum_{s=0}^t \widehat{\omega}_{s+1} - \widehat{\omega}_s\|_2^2 = \frac{1}{t^2}\|\widehat{\omega}_{t+1} - \widehat{\omega}_0\|_2^2$. Using this to lower bound the right hand side of the above when multiplying then proves the second part of the lemma.

We now proceed to the present the proof of Lemma 3, which come as an application of Lemma 4.

Proof 4 (Lemma 3) *Note Assumption 3 implies Assumption 5 with $\mathcal{X} = \mathbb{R}^p$ and $G_{\mathcal{X}}(u) = \epsilon\|u\|_2^2$. Plugging in the bound from Lemma 4 immediately yields the result.*

D Proof of Results for Two and Three Layer Neural Networks

In this section we present proofs of the result within Section 4 as well as Theorem 4 from Appendix A. Recall that we consider the supervised learning setting described within Section 4.1 where the loss function is a composition of a function $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, which is convex and smooth in its first argument $g(\cdot, y) : \mathbb{R} \to \mathbb{R}$ for any $y \in \mathbb{R}$, as well as a prediction function parameterised by $\omega \in \mathbb{R}^p$ such that $f(\cdot, \omega) : \mathbb{R}^d \to \mathbb{R}$. For an observation $Z = (x, y)$ the loss function is then $\ell(\omega, Z) = g(f(x, \omega), y)$.

Recall the minimum Eigenvalue of a matrix can be defined as the minimum quadratic form with vectors on a unit ball i.e. for $A \in \mathbb{R}^{p\times p}$ the quantity $\min_{u\in\mathbb{R}^p \|u\|_2=1} u^\top Au$ (see for instance Section 6.1.1 in (Wainwright, 2019)). Therefore, we are focused on lower bounding quadratic forms of the empirical risk Hessian i.e. $u^\top \nabla^2 R(\omega)u$ with $u \in \mathbb{R}^p$. It will therefore be convenient to write out the gradient and Hessian of the empirical risk in this case

$$\nabla R(\omega) = \frac{1}{N}\sum_{i=1}^N g'(f(x_i, \omega), y_i)\nabla f(x_i, \omega)$$
$$\nabla^2 R(\omega) = \frac{1}{N}\sum_{i=1}^N g''(f(x_i, \omega), y_i)\nabla f(x_i, \omega) f(x_i, \omega)^\top + \frac{1}{N}\sum_{i=1}^N g'(f(x_i, \omega), y_i)\nabla^2 f(x_i, \omega).$$

Moreover, note for $u \in \mathbb{R}^p$ the quadratic form of the Hessian then decomposes as

$$u^\top \nabla^2 R(\omega)u = \underbrace{\frac{1}{N}\sum_{i=1}^N g''(f(x_i, \omega), y_i)\langle u, \nabla f(x_i, \omega)\rangle^2}_{\geq 0} + \frac{1}{N}\sum_{i=1}^N g'(f(x_i, \omega), y_i)u^\top \nabla^2 f(x_i, \omega)u. \tag{9}$$

Since the first quantity above is non-negative, to lower bound the quadratic form $u^\top \nabla^2 R(\omega)u$ it is sufficient to lower bound the second term only.

The remainder of this section is then as follows. Section D.1 presents the proof of Theorem 3 which considers the minimum Eigenvalue of the empirical risk in the case of a two layer neural network. Section D.2 presents the proof of Theorem 6 which considers a three layer neural network when both the first and third layers are optimised.

D.1 Bounding Empirical Risk Hessian Minimum Eigenvalue for a Two Layer Neural Network (Theorem 3)

Recall Section 4, where the prediction function is a two layer neural network. This is defined for width $M \geq 1$ and scaling $1/2 \leq c \leq 1$ as $f(x, \omega) := \frac{1}{M^c}\sum_{j=1}^M v_j \sigma(\langle A_j, x\rangle)$. The parameter is then the concatenation $\omega = (A, v) \in \mathbb{R}^{dM+M}$ where the first layer $A \in \mathbb{R}^{M\times d}$ has been vectorised in a row-major manner. For a parameter ω and input $x \in \mathbb{R}^d$ the gradient and Hessian of the prediction function are, respectively, a vector $\nabla f(x, \omega) \in \mathbb{R}^{Md+M}$ and matrix $\nabla^2 f(x, \omega) \in \mathbb{R}^{(Md+M)\times(Md+M)}$. We now go on to calculate these quantities, noting for $k = 1, \ldots, d$ that the kth co-ordinate of the input x will be denoted $(x)_k$.

Let us decompose co-ordinates of the gradient $(\nabla f(x))_i$ for $i = 1, \ldots, Md + M$ into two parts associated to the first and second layer. Specifically, for the first layer consider $i = (j-1)d + k$ with $j = 1, \ldots, M$ and $k = 1, \ldots, d$. With this indexing, the ith co-ordinate aligns with the gradient of the jth neuron and kth input, which is

$$(\nabla f(x, \omega))_i = \frac{\partial f(x, \omega)}{\partial A_{jk}} = \frac{1}{M^c} v_j \sigma'(\langle A_j, x\rangle)(x)_k.$$

Meanwhile, for the second layer consider $i = Md + k$ with $k = 1, \ldots, M$. The ith co-ordinate then aligns with the gradient of the kth weight in the second layer, which is

$$(\nabla f(x, \omega))_i = \frac{\partial f(x, \omega)}{\partial v_k} = \frac{1}{M^c}\sigma(\langle A_k, x\rangle).$$

Let us now decompose the entries of the Hessian $(\nabla^2 f(x, \omega))_{i\ell}$ for $i, \ell = 1, \ldots, Md + M$ into three parts associated to the second derivative of the first layer only, second derivative of the second layer only, and the derivative with

respect to both the first and second layers. Specifically, for the second derivative with respect to the first layer consider $\ell = (\widetilde{j}-1)d + \widetilde{k}$ and $i = (j-1)d + k$ with $j, \widetilde{j} = 1, \ldots, M$ and $k, \widetilde{k} = 1, \ldots, d$. For these indices the Hessian is

$$(\nabla^2 f(x,\omega))_{i\ell} = \frac{\partial^2 f(x,\omega)}{\partial A_{jk} \partial A_{\widetilde{j}\widetilde{k}}} = \begin{cases} \frac{1}{M^c} v_j \sigma''(\langle A_j, x\rangle)(x)_k(x)_{\widetilde{k}} & \text{if } j = \widetilde{j} \\ 0 & \text{otherwise} \end{cases}.$$

For the second derivative with respect to the second layer, consider the indices $i = Md + j$ and $\ell = Md + \widetilde{j}$ with $j, \widetilde{j} = 1, \ldots, M$. For these indices Hessian is

$$(\nabla^2 f(x,\omega))_{i\ell} = \frac{f(x,\omega)}{\partial v_j \partial v_{\widetilde{j}}} = 0.$$

Finally, for the derivative with respect to the first and second layers consider the indicies $i = (j-1)d + k$ and $\ell = Md + \widetilde{j}$ with $j, \widetilde{j} = 1, \ldots, M$ and $k = 1, \ldots, d$. For these indices the Hessian is

$$(\nabla^2 f(x,\omega))_{i\ell} = \frac{\partial^2 f(x,\omega)}{\partial A_{jk} \partial v_{\widetilde{j}}} = \begin{cases} \frac{1}{M^c} \sigma'(\langle A_j, x\rangle)(x)_k & \text{if } j = \widetilde{j} \\ 0 & \text{Otherwise} \end{cases}.$$

Now let the vector $u \in \mathbb{R}^{Md+M}$ have unit norm $\|u\|_2 = 1$ and be composed in a manner matching the parameter $\omega = (A, v)$ so that $u = (A(u), v(u))$ where $A(u) \in \mathbb{R}^{M \times d}$ has been vectorised in a row-major manner and $u(v) \in \mathbb{R}^M$. Following the previous definitions, the quadratic form of the prediction function Hessian is

$$\begin{aligned} u^\top \nabla^2 f(x,\omega) u &= \sum_{\widetilde{j},j=1}^{M} \sum_{k,\widetilde{k}=1}^{d} A(u)_{jk} A(u)_{\widetilde{j}\widetilde{k}} \frac{\partial^2 f(x,\omega)}{\partial A_{jk} \partial A_{\widetilde{j}\widetilde{k}}} + 2 \sum_{j,\widetilde{j}=1}^{M} \sum_{k=1}^{d} A(u)_{jk} v(u)_{\widetilde{j}} \frac{\partial^2 f(x,\omega)}{\partial A_{jk} \partial v_{\widetilde{j}}} \\ &\quad + \sum_{j,\widetilde{j}=1}^{M} v(u)_j v(u)_{\widetilde{j}} \frac{\partial^2 f(x,\omega)}{\partial v_j \partial v_{\widetilde{j}}} \\ &= \frac{1}{M^c} \sum_{j=1}^{M} v_j \sigma''(\langle A_j, x\rangle) \langle x, A(u)_j \rangle^2 + 2 \frac{1}{M^c} \sum_{j=1}^{M} v(u)_j \sigma'(\langle A_j, x\rangle) \langle x, A(u)_j \rangle. \end{aligned}$$

Let us now lower bound the quadratic form of the empirical risk's Hessian $u^\top \nabla^2 R(\omega) u$, and thus, the minimum Eigenvalue. Recall the discussion around equation (9), in that the first term in the decomposition of the quadratic form $u^\top \nabla^2 R(\omega) u$ is non-negative. Utilising this and plugging in the above gives the following lower bound

$$\begin{aligned} u^\top \nabla^2 R(\omega) u &\geq \frac{1}{N} \sum_{i=1}^{N} g'(f(x_i,\omega), y_i) u^\top \nabla^2 f(x_i,\omega) u \\ &= \frac{1}{M^c} \frac{1}{N} \sum_{i=1}^{N} g'(f(x_i,\omega), y_i) \sum_{j=1}^{M} v_j \sigma''(\langle A_j, x_i\rangle) \langle x_i, A(u)_j \rangle^2 \\ &\quad + 2 \frac{1}{M^c} \frac{1}{N} \sum_{i=1}^{N} g'(f(x_i,\omega), y_i) \sum_{j=1}^{M} v(u)_j \sigma'(\langle A_j, x_i\rangle) \langle x_i, A(u)_j \rangle. \end{aligned} \tag{10}$$

We now set to lower bound each of the terms in (10). For the first term note we have, using upper bounds on the

derivative of g as well as the activation σ, the lower bound

$$\begin{aligned}
&\frac{1}{M^c}\frac{1}{N}\sum_{i=1}^{N} g'(f(x_i,\omega),y_i)\sum_{j=1}^{M} v_j\sigma''(\langle A_j,x_i\rangle)\langle x_i,A(u)_j\rangle^2\\
&\geq -\frac{1}{M^c}\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}|g'(f(x_i,A),y_i)v_j\sigma''(\langle A_j,x_i\rangle)|\langle x_i,A(u)_j\rangle^2\\
&\geq -\frac{L_{g'}L_{\sigma''}\|v\|_\infty}{M^c}\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}\langle x_i,A(u)_j\rangle^2\\
&= -\frac{L_{g'}L_{\sigma''}\|v\|_\infty}{M^c}\sum_{j=1}^{M}A(u)_j^\top\Big(\frac{1}{N}\sum_{i=1}^{N}x_ix_i^\top\Big)A(u)_j\\
&\geq -\frac{L_{g'}L_{\sigma''}\|v\|_\infty}{M^c}\|\widehat{\Sigma}\|_2\sum_{j=1}^{M}\|A(u)_j\|_2^2
\end{aligned}$$

Let us now consider the second term in (10). Bringing out the summation over $j = 1\ldots,M$ and using Cauchy-Schwarz we get

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N} g'(f(x_i,\omega),y_i)\sum_{j=1}^{M} v(u)_j\sigma'(\langle A_j,x_i\rangle)\langle x_i,A(u)_j\rangle\\
&=\sum_{j=1}^{M} v(u)_j\frac{1}{N}\sum_{i=1}^{N} g'(f(x_i,\omega),y_i)\sigma'(\langle A_j,x_i\rangle)\langle x_i,A(u)_j\rangle\\
&\geq -\Big|\sum_{j=1}^{M} v(u)_j\frac{1}{N}\sum_{i=1}^{N} g'(f(x_i,\omega),y_i)\sigma'(\langle A_j,x_i\rangle)\langle x_i,A(u)_j\rangle\Big|\\
&\geq -\sqrt{\sum_{j=1}^{M} v(u)_j^2}\sqrt{\sum_{j=1}^{M}\Big(\frac{1}{N}\sum_{i=1}^{N} g'(f(x_i,\omega),y_i)\sigma'(\langle A_j,x_i\rangle)\langle x_i,A(u)_j\rangle\Big)^2}\\
&\geq -\sqrt{\sum_{j=1}^{M} v(u)_j^2}\sqrt{\sum_{j=1}^{M}\frac{1}{N}\sum_{i=1}^{N}\big(g'(f(x_i,\omega),y_i)\sigma'(\langle A_j,x_i\rangle)\big)^2\langle x_i,A(u)_j\rangle^2}\\
&\geq -L_{\sigma'}L_{g'}\sqrt{\sum_{j=1}^{M} v(u)_j^2}\sqrt{\sum_{j=1}^{M}\frac{1}{N}\sum_{i=1}^{N}\langle x_i,A(u)_j\rangle^2}\\
&\geq -L_{\sigma'}L_{g'}\sqrt{\sum_{j=1}^{M} v(u)_j^2}\sqrt{\|\widehat{\Sigma}\|_2}\sqrt{\sum_{j=1}^{M}\|A(u)_j\|_2^2}
\end{aligned}$$

where we note some key steps within the above calculation. The third inequality arises from convexity of the squared function. Meanwhile the fourth inequality follows from the steps to bound the first term. Plugging each of these bounds in (10) then immediately yields the lower bound

$$\begin{aligned}
u^\top R(\omega)u &\geq -\frac{L_{g'}L_{\sigma''}\|\widehat{\Sigma}\|_2\|v\|_\infty}{M^c}\sum_{j=1}^{M}\|A(u)_j\|_2^2 - 2\frac{L_{\sigma'}L_{g'}\sqrt{\|\widehat{\Sigma}\|_2}}{M^c}\|v(u)\|_2\sqrt{\sum_{j=1}^{M}\|A(u)_j\|_2^2}\\
&\geq -\frac{L_{g'}L_{\sigma''}\|\widehat{\Sigma}\|_2\|v\|_\infty}{M^c} - 2\frac{L_{\sigma'}L_{g'}\sqrt{\|\widehat{\Sigma}\|_2}}{M^c},
\end{aligned}$$

where at the end we used that $\|u\|_2^2 = \|v(u)\|_2^2 + \sum_{j=1}^{M}\|A(u)_j\|_2^2 = 1$. This is then a lower bound on the minimum Eigenvalue, as required.

D.2 Bounding Empirical Risk Hessian Minimum Eigenvalue for a Three Layer Neural Network

In this section we provide a lower bound on the minimum Eigenvalue of the empirical risk Hessian when the prediction function is a three layer neural network. Specifically, for $M_1, M_2 \geq 1$ and $1/2 \leq c \leq 1$ the prediction function takes the form

$$f(x,\omega) = \frac{1}{M_2^c}\sum_{i=1}^{M_2} v_i \sigma\Big(\frac{1}{M_1^c}\sum_{s=1}^{M_1} A_{is}^{(2)}\langle A_s^{(1)}, x\rangle\Big), \tag{11}$$

where we have denoted the first layer of weights $A^{(1)} \in \mathbb{R}^{M_1 \times d}$, the second layer of weights $A^{(2)} \in \mathbb{R}^{M_2 \times M_1}$ and third layer of weights $v \in \mathbb{R}^{M_2}$. The first activation is linear, while the second activation is more generally $\sigma : \mathbb{R} \to \mathbb{R}$. We consider the setting in which we only optimise the first and third set of weights, and thus, the parameter will be the concatenation $\omega = (A^{(1)}, v) \in \mathbb{R}^{M_1 d + M_2}$ where the first layer of weights $A^{(1)}$ have been vectorised in a row-major manner. It will also be convenient to denote the empirical covariance of the second layer $\widehat{\Sigma}_{A^{(2)}} := \frac{1}{M_2}(A^{(2)})^\top A^{(2)}$. The following theorem then presents the lower bound on the minimum Eigenvalue in this case.

Theorem 6 *Consider the loss function as in Section 4.1 with three layer neural network prediction function* (11)*. Suppose the activation σ has first and second derivative bounded by $L_{\sigma'}$ and $L_{\sigma''}$ respectively. Moreover, suppose the function g has bounded derivative $|g'(\widetilde{y}, y)| < L_{g'}$. Then with $\omega = (A^{(1)}, v)$*

$$\nabla^2 R(\omega) \succeq -\Big(L_{\sigma''}L_{g'}\frac{\|\widehat{\Sigma}\|_2\|\widehat{\Sigma}_{A^{(2)}}\|_2}{M_1^{2c}M_2^{c-1}}\|v\|_\infty + 2L_{g'}L_{\sigma'}\frac{\sqrt{\|\widehat{\Sigma}\|_2\|\widehat{\Sigma}_{A^{(2)}}\|_2}}{M_2^{c-1/2}M_1^c}\Big)I.$$

We now briefly discuss the above theorem, with the proof presented thereafter. We note that the lower bound in Theorem 6 now scales with the product of layer widths i.e. $M_1^{2c}M_2^{c-1}$ and $M_2^{c-1/2}M_1^c$. Let us then suppose that each layer is the same width so $M_1 = M_2 = M$, the spectral norm of the second layer's empirical covariance $\|\widehat{\Sigma}_{A^{(2)}}\|_2$ is constant and the third layer remains bounded $\|v\|_\infty$. In this case the bound is then $O(1/(M^{2c-1/2}))$. In contrast, the bound for a two layer neural network presented in Theorem 3 is $O(1/M^c)$. The lower bounds are then the same order for $c = 1/2$, while for $c > 1/2$, the three layer neural network case is smaller. We now give the proof of Theorem 6.

Proof 5 (Theorem 6) *In a similar manner to the proof of Theorem 3, let us begin by defining the gradient and Hessian of the prediction function with respect to the parameter ω. In this case they are, respectively, a vector $\nabla f(x,\omega) \in \mathbb{R}^{M_1 d + M_2}$ and matrix $\nabla^2 f(x,\omega) \in \mathbb{R}^{(M_1 d + M_2)\times(M_1 d + M_2)}$.*

Once again split the co-ordinates of the gradient $(\nabla f(x,\omega))_i$ for $i = 1, \ldots, M_1 d + M_2$ into two parts associated to the first and third layers. For the first layer, consider $i = (j-1)d + k$ with $j = 1, \ldots, M_1$ and $k = 1, \ldots, d$. This aligns with the gradient of the jth neuron in the first layer and kth input, which is

$$(\nabla f(x,\omega))_i = \frac{\partial f(x,\omega)}{\partial A_{jk}^{(1)}} = \frac{1}{M_2^c}\sum_{i=1}^{M_2} v_i \sigma'\Big(\frac{1}{M_1^c}\sum_{s=1}^{M_1} A_{is}^{(2)}\langle A_s^{(1)}, x\rangle\Big)\frac{A_{ij}^{(2)}}{M_1^c}(x)_k.$$

For the third layer consider $i = M_1 d + k$ with $k = 1, \ldots, M_2$. This aligns with the gradient of the kth weight in the third layer, which is

$$(\nabla f(x,\omega))_i = \frac{\partial f(x,\omega)}{\partial v_k} = \frac{1}{M_2^c}\sigma\Big(\frac{1}{M_1^c}\sum_{s=1}^{M_1} A_{ks}^{(2)}\langle A_s^{(1)}, x\rangle\Big).$$

Let us now decompose the entries of the Hessian $(\nabla^2 f(x,\omega))_{i\ell}$ for $i, \ell = 1, \ldots, M_1 d + M_2$ into three parts associated to the second derivative of the first layer only, second derivative of the third layer only, and the derivative with respect to the first and third layer. For the second deriative of the first layer consider the indices $i = (j-1)d + k$ and $\ell = (\widetilde{j}-1)d + \widetilde{k}$ with $j, \widetilde{j} = 1, \ldots, M_1$ and $k, \widetilde{k} = 1, \ldots, d$. For these indices the Hessian is

$$(\nabla^2 f(x,\omega))_{i\ell} := \frac{\partial^2 f(x,\omega)}{\partial A_{jk}^{(1)}\partial A_{\widetilde{j}\widetilde{k}}^{(1)}} = \frac{1}{M_2^c}\sum_{i=1}^{M_2} v_i \sigma''\Big(\frac{1}{M_1^c}\sum_{s=1}^{M_1} A_{is}^{(2)}\langle A_s, x\rangle\Big)\frac{A_{ij}^{(2)}}{M_1^c}\frac{A_{i\widetilde{j}}^{(2)}}{M_1^c}(x)_k(x)_{\widetilde{k}}.$$

Meanwhile for the Hessian with respect to the third layer consider the indices $i = M_1 d + k$ and $\ell = M_1 d + \widetilde{k}$ with $k, \widetilde{k} = 1, \dots, M_2$. The Hessian in this case is then

$$(\nabla^2 f(x,\omega))_{i\ell} := \frac{\partial^2 f(x,\omega)}{\partial v_k \partial v_{\widetilde{k}}} = 0.$$

Finally, for the derivative with respect to the first and third layers consider the indices $i = (j-1)d + k$ and $\ell = M_1 d + \widetilde{k}$ with $j = 1, \dots, M_1$ and $k = 1, \dots, d$ and $\widetilde{k} = 1, \dots, M_2$. The Hessian in this case is then

$$(\nabla^2 f(x,\omega))_{i\ell} := \frac{\partial^2 f(x,\omega)}{\partial A^{(1)}_{jk} \partial v_{\widetilde{k}}} = \frac{1}{M_2^c} \sigma'\Big(\frac{1}{M_1^c} \sum_{s=1}^{M_1} A^{(2)}_{\widetilde{k}s} \langle A^{(1)}_s, x\rangle\Big) \frac{A^{(2)}_{\widetilde{k}j}}{M_1^c}(x)_k.$$

Let the vector $u \in \mathbb{R}^{M_1 d + M_2}$ have unit norm $\|u\|_2 = 1$ and be composed in a manner matching the parameter $\omega = (A^{(1)}, v)$, so that $u = (A^{(1)}(u), v(u))$ where $A^{(1)}(u) \in \mathbb{R}^{M_1 \times d}$ has been vectorised in a row-major manner and $v(u) \in \mathbb{R}^{M_2}$. Following the previous definitions, the quadratic form of the prediction function Hessian then takes the form

$$\begin{aligned}
u^\top \nabla^2 f(x,\omega) u &= \sum_{j,\widetilde{j}=1}^{M_1} \sum_{k,\widetilde{k}=1}^{d} A^{(1)}(u)_{jk} A^{(1)}(u)_{\widetilde{j}\widetilde{k}} \frac{\partial^2 f(x,\omega)}{\partial A^{(1)}_{jk} \partial A^{(1)}_{\widetilde{j}\widetilde{k}}} + 2 \sum_{j=1}^{M_1} \sum_{k=1}^{d} \sum_{\widetilde{k}=1}^{M_2} A^{(1)}(u)_{jk} v(u)_{\widetilde{k}} \frac{\partial^2 f(x,\omega)}{\partial A^{(1)}_{jk} \partial v_{\widetilde{k}}} \\
&\quad + \sum_{k,\widetilde{k}=1}^{M_2} v(u)_k v(u)_{\widetilde{k}} \frac{\partial^2 f(x,\omega)}{\partial v_k \partial v_{\widetilde{k}}} \\
&= \sum_{j,\widetilde{j}=1}^{M_1} \frac{1}{M_2^c} \sum_{i=1}^{M_2} v_i \sigma''\Big(\frac{1}{M_1^c} \sum_{s=1}^{M_1} A^{(2)}_{is} \langle A^{(1)}_s, x\rangle\Big) \frac{A^{(2)}_{ij}}{M_1^c} \frac{A^{(2)}_{\widetilde{i}\widetilde{j}}}{M_1^c} \langle A^{(1)}(u)_j, x\rangle \langle A^{(1)}(u)_{\widetilde{j}}, x\rangle \\
&\quad + 2 \sum_{j=1}^{M_1} \sum_{\widetilde{k}=1}^{M_2} v(u)_{\widetilde{k}} \frac{1}{M_2^c} \sigma'\Big(\frac{1}{M_1^c} \sum_{s=1}^{M_1} A^{(2)}_{\widetilde{k}s} \langle A^{(1)}_s, x\rangle\Big) \frac{A^{(2)}_{\widetilde{k}j}}{M_1^c} \langle x, A^{(1)}(u)_j\rangle \\
&= \frac{1}{M_2^c} \sum_{i=1}^{M_2} v_i \sigma''\Big(\frac{1}{M_1^c} \sum_{s=1}^{M_1} A^{(2)}_{is} \langle A^{(1)}_s, x\rangle\Big) \frac{1}{M_1^{2c}} \Big\langle \sum_{j=1}^{M_1} A^{(2)}_{ij} A^{(1)}(u)_j, x\Big\rangle^2 \\
&\quad + 2 \sum_{\widetilde{k}=1}^{M_2} v(u)_{\widetilde{k}} \frac{1}{M_2^c} \sigma'\Big(\frac{1}{M_1^c} \sum_{s=1}^{M_1} A^{(2)}_{\widetilde{k}s} \langle A^{(1)}_s, x\rangle\Big) \frac{1}{M_1^c} \Big\langle x, \sum_{j=1}^{M} A^{(2)}_{\widetilde{k}j} A^{(1)}(u)_j\Big\rangle
\end{aligned}$$

where on the second equality we have taken the summation over $j, \widetilde{j} = 1, \dots, M$ inside the inner products. We now set to lower bound the quadratic form involving the empirical risk Hessian $u^\top \nabla^2 R(\omega) u$, and thus, the minimum Eigenvalue. Recalling the discussion around equation (9) *and plugging in the above we get*

$$\begin{aligned}
&u^\top \nabla^2 R(\omega) u \\
&\geq \frac{1}{N} \sum_{\ell=1}^{N} g'(f(x_\ell,\omega), y_\ell) u^\top \nabla^2 f(x_\ell,\omega) u \\
&= \underbrace{\frac{1}{N} \sum_{\ell=1}^{N} g'(f(x_\ell,\omega), y_\ell) \frac{1}{M_2^c} \sum_{i=1}^{M_2} v_i \sigma''\Big(\frac{1}{M_1^c} \sum_{s=1}^{M_1} A^{(2)}_{is} \langle A^{(1)}_s, x_\ell\rangle\Big) \frac{1}{M_1^{2c}} \Big\langle \sum_{j=1}^{M_1} A^{(2)}_{ij} A^{(1)}(u)_j, x_\ell\Big\rangle^2}_{=: \textbf{\textit{Term 1}}} \\
&\quad + \underbrace{2 \frac{1}{N} \sum_{\ell=1}^{N} g'(f(x_\ell,\omega), y_\ell) \sum_{\widetilde{k}=1}^{M_2} v(u)_{\widetilde{k}} \frac{1}{M_2^c} \sigma'\Big(\frac{1}{M_1^c} \sum_{s=1}^{M_1} A^{(2)}_{\widetilde{k}s} \langle A^{(1)}_s, x_\ell\rangle\Big) \frac{1}{M_1^c} \Big\langle x_\ell, \sum_{j=1}^{M} A^{(2)}_{\widetilde{k}j} A^{(1)}(u)_j\Big\rangle}_{=: \textbf{\textit{Term 2}}}.
\end{aligned}$$

Which has then been decomposed into two components labelled ***Term 1*** *and* ***Term 2****, each of which will now be lower bounded separately. For* ***Term 1****, aligning with the Hessian of the first layer, we get using the upper bound*

on the gradient of g i.e $|g'(\cdot, y)| \leq L_{g'}$ and the activation $|\sigma''(\cdot)| \leq L_{\sigma''}$

$$\textbf{\textit{Term 1}} \geq -\frac{L_{g'} L_{\sigma''} \|v\|_\infty}{M_2^c} \frac{1}{N} \frac{1}{M_1^{2c}} \sum_{\ell=1}^{N} \sum_{i=1}^{M_2} \Big\langle \sum_{j=1}^{M_1} A_{ij}^{(2)} A^{(1)}(u)_j, x_\ell \Big\rangle^2$$

Meanwhile, for ***Term 2*** *bring out the sum over $\widetilde{k} = 1, \ldots, M_2$ and apply Cauchy-Schwarz to get*

$$\begin{aligned}
&\textbf{\textit{Term 2}} \\
&= \sum_{\widetilde{k}=1}^{M_2} v(u)_{\widetilde{k}} \frac{1}{N} \sum_{\ell=1}^{N} g'(f(x_\ell, \omega), y_\ell) \frac{1}{M_2^c} \sigma'\Big(\frac{1}{M_1^c} \sum_{s=1}^{M_1} A_{\widetilde{k}s}^{(2)} \langle A_s^{(1)}, x_\ell \rangle\Big) \frac{1}{M_1^c} \Big\langle x_\ell, \sum_{j=1}^{M_1} A_{\widetilde{k}j}^{(2)} A^{(1)}(u)_j \Big\rangle \\
&\geq -\sqrt{\sum_{\widetilde{k}=1}^{M_2} v(u)_{\widetilde{k}}^2} \sqrt{\sum_{\widetilde{k}=1}^{M_2} \Big(\frac{1}{N} \sum_{\ell=1}^{N} g'(f(x_\ell, \omega), y_\ell) \frac{1}{M_2^c} \sigma'\Big(\frac{1}{M_1^c} \sum_{s=1}^{M_1} A_{\widetilde{k}s}^{(2)} \langle A_s^{(1)}, x_\ell \rangle\Big) \frac{1}{M_1^c} \Big\langle x_\ell, \sum_{j=1}^{M} A_{\widetilde{k}j}^{(2)} A^{(1)}(u)_j \Big\rangle \Big)^2} \\
&\geq -\frac{L_{g'} L_{\sigma'}}{M_2^c M_1^c} \sqrt{\sum_{\widetilde{k}=1}^{M_2} v(u)_{\widetilde{k}}^2} \sqrt{\sum_{\widetilde{k}=1}^{M_2} \frac{1}{N} \sum_{\ell=1}^{N} \Big\langle x_\ell, \sum_{j=1}^{M_1} A_{\widetilde{k}j}^{(2)} A^{(1)}(u)_j \Big\rangle^2}.
\end{aligned}$$

Where for the second inequality we have followed a similar set of steps to those within the proof of Theorem 3. In particular, we applied convexity of $x \to x^2$ as well as the upper bounds on the gradient of g and the activation σ. We are now left to upper bound the quantity $\frac{1}{N} \sum_{\ell=1}^{N} \sum_{i=1}^{M_2} \Big\langle \sum_{j=1}^{M_1} A_{ij}^{(2)} A^{(1)}(u)_j, x_\ell \Big\rangle^2$ which appears within the bounds for both ***Term 1*** *and* ***Term 2****. To bound this quantity, let us rewrite the summation as a vector matrix multiplication so $\sum_{j=1}^{M_1} A_{ij}^{(2)} A^{(1)}(u)_j = A_i^{(2)} A^{(1)}(u)$ where $A_i^{(2)}$ is the ith row of $A^{(2)}$. Normalising by M_2, we note this quantity can be bounded in terms of the spectral norm of the empirical covariance $\|\widehat{\Sigma}\|_2$ and covariance of the second layer $\|\widehat{\Sigma}_{A^{(2)}}\|_2$. Specifically,*

$$\begin{aligned}
\frac{1}{M_2} \frac{1}{N} \sum_{\ell=1}^{N} \sum_{i=1}^{M_2} \Big\langle \sum_{j=1}^{M_1} A_{ij}^{(2)} A^{(1)}(u)_j, x_\ell \Big\rangle^2 &= \frac{1}{M_2} \sum_{i=1}^{M_2} \Big(A_i^{(2)} A^{(1)}(u)\Big)^\top \underbrace{\Big(\frac{1}{N} \sum_{\ell=1}^{N} x_\ell x_\ell^\top\Big)}_{\widehat{\Sigma}} \Big(A_i^{(2)} A^{(1)}(u)\Big) \\
&\leq \|\widehat{\Sigma}\|_2 \frac{1}{M_2} \sum_{i=1}^{M_2} \|A_i^{(2)} A^{(1)}(u)\|_2^2 \\
&= \|\widehat{\Sigma}\|_2 \frac{1}{M_2} \sum_{i=1}^{M_2} \mathrm{Tr}\Big(A_i^{(2)} A^{(1)}(u) \big(A^{(1)}(u)\big)^\top \big(A_i^{(2)}\big)^\top\Big) \\
&= \|\widehat{\Sigma}\|_2 \,\mathrm{Tr}\Big(A^{(1)}(u) \big(A^{(1)}(u)\big)^\top \underbrace{\Big(\frac{1}{M_2} \sum_{i=1}^{M_2} \big(A_i^{(2)}\big)^\top A_i^{(2)}\Big)}_{\widehat{\Sigma}_{A^{(2)}}}\Big) \\
&\leq \|\widehat{\Sigma}\|_2 \|\widehat{\Sigma}_{A^{(2)}}\|_2 \|A^{(1)}(u)\|_F^2
\end{aligned}$$

The final lower bound is then arrived at by using the above to lower bound ***Term 1*** *and* ***Term 2*** *and recalling that $\|u\|_2^2 = \|v(u)\|_2^2 + \|A^{(1)}(u)\|_F^2 = 1$.*

D.3 Bounding Optimisation and Generalisation Error for a Two Layer Neural Network (Theorem 4)

In this section we present the proof of Theorem 4 from Appendix A.2. The proof proceeds in two steps. The first step is to apply Proposition 1 from Appendix A.1 to bound the **Opt. & Approx. Error**. The second step involves utilising the structure of $G_C(\cdot)$ to decouple the place holder $\widetilde{\omega}$ and the iterates of gradient descent $\widehat{\omega}_s$. Each of these steps will, respectively, be included in the following paragraphs **Applying Proposition 1** and **Decoupling Gradient Descent Iterates**.

Applying Proposition 1 To applying Proposition 1 we require showing two conditions: Assumption 5 holds for the set $\mathcal{X}_{L_v}$ and function $G_C(\cdot)$ described in Section A.2; and the iterates of gradient descent to remain within the set $\widehat{\omega}_s \in \mathcal{X}_{L_v}$ for $s \geq 0$. Let us begin with the first of these conditions.

To ensure Assumption 5 holds we must lower bound the quadratic form $u^\top \nabla^2 R(\omega) u \geq -G_C(u)$ for $\omega \in \mathcal{X}_{L_v}$ and $u \in \mathbb{R}^{Md+M}$. To this end, let us recall the steps in the proof of Theorem 3 in Appendix D.1. Specifically, looking to equation (10) where we have for $u = (A(u), v(u)) \in \mathbb{R}^{Md+M}$ a lower bound on the quadratic form $u^\top \nabla^2 R(\omega) u$ involving the sum of two terms. For the first term simply consider the lower bound presented within Theorem 3 which is $-L_{g'} L_{\sigma''} \|\widehat{\Sigma}\|_2 \|v\|_\infty \|A(u)\|_F^2 / M^c$. For the second term note it can be lower bounded

$$
\begin{aligned}
&2\frac{1}{M^c}\frac{1}{N}\sum_{i=1}^N g'(f(x_i,\omega),y_i)\sum_{j=1}^M v(u)_j \sigma'(\langle A_j, x_i\rangle)\langle x_i, A(u)_j\rangle \\
&\geq -2\frac{L_{g'}}{M^c}\frac{1}{N}\sum_{i=1}^N \Big|\Big\langle \sum_{j=1}^M v(u)_j \sigma'(\langle A_j, x_i\rangle) A(u)_j, x_i\Big\rangle\Big| \\
&= -2\frac{L_{g'}}{M^c}\frac{1}{N}\sum_{i=1}^N \Big|\Big\langle \sum_{j=1}^M v(u)_j \sigma'(\langle A_j, x_i\rangle) A(u)_j, x_i\Big\rangle\Big| \\
&\geq -2\frac{L_{g'}}{M^c}\frac{1}{N}\sum_{i=1}^N \Big\| \sum_{j=1}^M v(u)_j \sigma'(\langle A_j, x_i\rangle) A(u)_j\Big\|_2 \|x_i\|_2
\end{aligned}
$$

where at the end we applied Hölders inequality. Furthermore, if we denote the vector $\alpha \in \mathbb{R}^M$ such that $\alpha_\ell = \sigma'(\langle A_l, x_i\rangle)$ for $\ell = 1, \ldots, M$, note we can rewrite $\sum_{j=1}^M v(u)_j \sigma'(\langle A_j, x_i\rangle) A(u)_j = A(u)^\top \mathrm{Diag}(\alpha) v(u)$. Since the activation σ has derivative bounded by $L_{\sigma'}$ we then have $\alpha_i / L_{\sigma'} \in [-1, 1]$ and the upper bound $\big\|A(u)^\top \mathrm{Diag}(\alpha) v(u)\big\|_2 \leq L_{\sigma'} \max_{\|z\|_\infty \leq 1} \big\|A(u)^\top \mathrm{Diag}(z) v(u)\big\|_2$. Combined with the upper bound on the covariates $\frac{1}{N}\sum_{i=1}^N \|x_i\|_2 \leq \frac{1}{N}\sqrt{N}\sqrt{\sum_{i=1}^N \|x_i\|_2^2} = \sqrt{\mathrm{Tr}\left(\widehat{\Sigma}\right)}$, we get the lower bound

$$
\begin{aligned}
&2\frac{1}{M^c}\frac{1}{N}\sum_{i=1}^N g'(f(x_i,\omega),y_i)\sum_{j=1}^M v(u)_j \sigma'(\langle A_j, x_i\rangle)\langle x_i, A(u)_j\rangle \\
&\geq -2\frac{L_{g'} L_{\sigma'}\sqrt{\mathrm{Tr}\left(\widehat{\Sigma}\right)}}{M^c} \max_{\|z\|_\infty \leq 1} \big\|A(u)^\top \mathrm{Diag}(z) v(u)\big\|_2.
\end{aligned}
$$

Bringing together the bounds on the two terms yields

$$
u^\top \nabla^2 R(\omega) u \geq -\frac{L_{g'} L_{\sigma''} \|\widehat{\Sigma}\|_2}{M^c}\|v\|_\infty \|A(u)\|_F^2 - 2\frac{L_{g'} L_{\sigma'}\sqrt{\mathrm{Tr}\left(\widehat{\Sigma}\right)}}{M^c} \max_{\|z\|_\infty \leq 1} \Big\|A(u)^\top \mathrm{Diag}(z) v(u)\Big\|_2.
$$

We then see that Assumption 5 is satisfied with $\mathcal{X}_{L_v}$ and $G_C(\cdot)$ provided $C \geq 2L_{g'}\Big[\Big(L_{\sigma'}\sqrt{\mathrm{Tr}\left(\widehat{\Sigma}\right)}\Big) \vee (L_{\sigma''} L_v \|\widehat{\Sigma}\|_2)\Big]$.

Let us now ensure the iterates of gradient descent remain within the set $\mathcal{X}_{L_v}$. To do so we must consider the infinity norm of the second layer throughout training. Let the gradient descent iterates be denoted as $\widehat{\omega}_s = (A^s, v^s)$ for $s \geq 0$ where $A^s \in \mathbb{R}^{M \times d}$ has been vectorised in a row-major manner. Plugging in the definition of the empirical risk gradient, the second layer is updated for $s \geq 0$ as

$$
v_k^{s+1} = v_k^s - \eta \frac{1}{N}\sum_{i=1}^N g'(f(x_i,\omega),y_i)\frac{1}{M^c}\sigma(\langle A_k, x_i\rangle).
$$

Using the upper bound on the derivative $|g'(\cdot)| \leq L_{g'}$ as well as the activation $|\sigma(\cdot)| \leq L_\sigma$ we get $\|v^{s+1}\|_\infty \leq \|v^s\|_\infty + \frac{\eta L_{g'} L_\sigma}{M^c} \leq \|v^0\|_\infty + \frac{\eta t L_{g'} L_\sigma}{M^c}$. Therefore, we can ensure $\widehat{\omega}_s \in \mathcal{X}_{L_v}$ for $t \geq s \geq 0$ if $L_v \geq \|v^0\|_\infty + \frac{\eta t L_{g'} L_\sigma}{M^c}$. We can now apply Proposition 1, to upper bound the **Opt & Approx. Error**.

Decoupling Gradient Descent Iterates: Given the upper bound presented within Proposition 1, we must now decoupling the iterates of gradient descent $\widehat{\omega}_s$ and the placeholder $\widetilde{\omega}$ within the **Approximation Error**,

specifically $G_C(\widetilde{\omega} - \widehat{\omega}_s)$. We will, in short, add and subtract the initialisation of gradient descent $\widehat{\omega}_0$ and apply triangle inequality a number of times. Let us denote the placeholder $\widetilde{\omega} = (\widetilde{A}, \widetilde{v})$ where $\widetilde{A} \in \mathbb{R}^{M \times d}$ has been vectorised in a row-major manner and the second layer of weights is $\widetilde{v} \in \mathbb{R}^M$. For a general $\omega = (A, v)$ and $\widehat{\omega}_0 = (A^0, v^0)$, we then get when adding and subtracting A^0 inside the Frobeinus norm to get

$$\begin{aligned} G_C(\widetilde{\omega} - \omega) \leq \frac{2C}{M^c}\|\widetilde{A} - A^0\|_F^2 + \frac{2C}{M^c}\|A^0 - A\|_F^2 \\ + \frac{C}{M^c}\max_{\|z\|_\infty \leq 1} \left\|(\widetilde{A} - A)^\top \mathrm{Diag}(z)(\widetilde{v} - v)\right\|_2. \end{aligned}$$

We now consider the second term. Adding and subtracting A^0 and v^0 we get the following four terms

$$\begin{aligned} &\left\|(\widetilde{A} - A)^\top \mathrm{Diag}(z)(\widetilde{v} - v)\right\|_2 \\ &\leq \left\|(\widetilde{A} - A^0)^\top \mathrm{Diag}(z)(\widetilde{v} - v)\right\|_2 + \left\|(A^0 - A)^\top \mathrm{Diag}(z)(\widetilde{v} - v)\right\|_2 \\ &\leq \left\|(\widetilde{A} - A^0)^\top \mathrm{Diag}(z)(\widetilde{v} - v^0)\right\|_2 + \left\|(\widetilde{A} - A^0)^\top \mathrm{Diag}(z)(v^0 - v)\right\|_2 \\ &\quad + \left\|(A^0 - A)^\top \mathrm{Diag}(z)(\widetilde{v} - v^0)\right\|_2 + \left\|(A^0 - A)^\top \mathrm{Diag}(z)(v^0 - v)\right\|_2. \end{aligned}$$

The second, third and fourth terms can then be bounded as follows. The second term can be bounded

$$\begin{aligned} \left\|(\widetilde{A} - A^0)^\top \mathrm{Diag}(z)(v^0 - v)\right\|_2 &\leq \left\|(\widetilde{A} - A^0)^\top\right\|_2 \|\mathrm{Diag}(z)\|_2 \|v^0 - v\|_2 \\ &\leq \frac{1}{2}\|\widetilde{A} - A^0\|_F^2 + \frac{1}{2}\|v^0 - v\|_2^2 \\ &\leq \frac{1}{2}\|\widetilde{A} - A^0\|_F^2 + \frac{1}{2}\|\widehat{\omega}_0 - \omega\|_2^2 \end{aligned}$$

where we have used the following set of steps. The first inequality arises from using the consistency of matrix norms to be break the product of matrices into operator norms. The second inequality follows from: noting that $\|\mathrm{Diag}(z)\|_2 \leq 1$ since $\|z\|_\infty \leq 1$; upper bounding the matrix ℓ_2 operator norm by the Frobenius norm; and using young's inequality to break apart the product of norms. The final inequality comes from simply adding the co-ordinates associated to the first layer to the second term. Following a similar set of steps the fourth term can be bounded

$$\begin{aligned} \left\|(A^0 - A)^\top \mathrm{Diag}(z)(v^0 - v)\right\|_2 &\leq \frac{1}{2}\|A^0 - A\|_F^2 + \frac{1}{2}\|v^0 - v\|_2^2 \\ &= \frac{1}{2}\|\widehat{\omega}_0 - \omega\|_2^2. \end{aligned}$$

Meanwhile, the third term is bounded without applying Young's inequality so

$$\begin{aligned} \left\|(A^0 - A)^\top \mathrm{Diag}(z)(\widetilde{v} - v^0)\right\|_2 &\leq \|A^0 - A\|_F \|\widetilde{v} - v^0\|_2 \\ &\leq \|\widehat{\omega}_0 - \omega\|_2 \|\widetilde{v} - v^0\|_2. \end{aligned}$$

Bringing everything together we get the following upper bound on the function $G_C(\widetilde{\omega} - u)$

$$\begin{aligned} G_C(\widetilde{\omega} - \omega) \leq \frac{3C}{M^c}\|\widehat{\omega}_0 - \omega\|_2^2 \\ + \frac{3C}{M^c}\|\widetilde{A} - A^0\|_F^2 + \frac{C}{M^c}\|\widehat{\omega}_0 - \omega\|_2 \|\widetilde{v} - v^0\|_2 \\ + \frac{C}{M^c}\max_{\|z\|_\infty \leq 1} \left\|(\widetilde{A} - A^0)^\top \mathrm{Diag}(z)(\widetilde{v} - v^0)\right\|_2. \end{aligned}$$

We now utilise the above within the **Approximation Error** in Proposition 1. In particular, for $s = 1, \ldots, t$ pick $\omega = \widehat{\omega}_s$ and sum up the above for $s = 1, \ldots, t$. Utilising the bound $\|\widehat{\omega}_s - \widehat{\omega}_0\|_2 \leq \sqrt{2\eta t(R(\widehat{\omega}_0) - R(\widehat{\omega}^\star))}$ (See

Lemma 4) then yields

$$\begin{aligned}\frac{1}{t}\sum_{s=1}^{t} G_C(\widetilde{\omega}-\widehat{\omega}_s) &\leq \frac{6C\eta t(R(\widehat{\omega}_0)-R(\widehat{\omega}^\star))}{M^c}+\frac{3C}{M^c}\|\widetilde{A}-A^0\|_F^2+\frac{C\sqrt{2\eta t((R(\widehat{\omega}_0)-R(\widehat{\omega}^\star))}}{M^c}\|\widetilde{v}-v^0\|_2\\ &\quad+\frac{C}{M^c}\max_{\|z\|_\infty\leq 1}\left\|(\widetilde{A}-A^0)^\top \mathrm{Diag}(z)(\widetilde{v}-v^0)\right\|_2\\ &\leq \frac{6C\eta t(R(\widehat{\omega}_0)-R(\widehat{\omega}^\star))}{M^c}+\frac{3C(\sqrt{(R(\widehat{\omega}_0)-R(\widehat{\omega}^\star))}\vee 1)}{M^c}\\ &\quad\times\underbrace{\left(\|\widetilde{A}-A^0\|_F^2+\sqrt{\eta t}\|\widetilde{v}-v^0\|_2+\max_{\|z\|_\infty\leq 1}\left\|(\widetilde{A}-A^0)^\top \mathrm{Diag}(z)(\widetilde{v}-v^0)\right\|_2\right)}_{H(\widetilde{\omega}-\widehat{\omega}_0)}\end{aligned}$$

where on the second inequality we simply pulled out the constant on the second term. Plugging into the **Approximation Error** within Proposition 1, then yields

$$\begin{aligned}\frac{1}{2}\frac{1}{t}\sum_{s=1}^{t} G_C(\widetilde{\omega}-\widehat{\omega}_s)+R(\widetilde{\omega})-r(\omega^\star) &\leq \frac{3C\eta t(R(\widehat{\omega}_0)-R(\widehat{\omega}^\star))}{M^c}\\ &\quad+\frac{3C(\sqrt{(R(\widehat{\omega}_0)-R(\widehat{\omega}^\star))}\vee 1)}{M^c}H(\widetilde{\omega}-\widehat{\omega}_0)+R(\widetilde{\omega})-r(\omega^\star)\\ &= \frac{3C\eta t(R(\widehat{\omega}_0)-R(\widehat{\omega}^\star))}{M^c}+\lambda H(\widetilde{\omega}-\widehat{\omega}_0)+R(\widetilde{\omega})-r(\omega^\star)\end{aligned}$$

where we have defined $\lambda=\frac{3C(\sqrt{(R(\widehat{\omega}_0)-R(\widehat{\omega}^\star))}\vee 1)}{M^c}$. With a population risk minimiser $\omega^\star=(A^\star,v^\star)$, we recall that $L_v\geq\|v^\star\|_\infty$ and thus, $\omega^\star\in\mathcal{X}_{L_v}$. Therefore, if we pick $\widetilde{\omega}=\widehat{\omega}_\lambda\in\mathrm{argmin}_{\omega\in\mathcal{X}_{L_v}}\left\{R(\omega)+\lambda H(\omega-\widehat{\omega}_0)\right\}$, the right most quantity above can be upper bounded

$$\lambda H(\widetilde{\omega}-\widehat{\omega}_0)+R(\widetilde{\omega})-r(\omega^\star)\leq\lambda H(\omega^\star-\widehat{\omega}_0)+R(\omega^\star)-r(\omega^\star).$$

Taking expectation and noting that $\mathbf{E}[R(\omega^\star)]=r(\omega^\star)$ yields the following bound on the **Approximation Error**

$$\frac{1}{2}\frac{1}{t}\sum_{s=1}^{t}\mathbf{E}[G_{\mathcal{C}}(\widetilde{\omega}-\widehat{\omega}_s)]+\mathbf{E}[R(\widetilde{\omega})]-r(\omega^\star)\leq\frac{3C\eta t\mathbf{E}[R(\widehat{\omega}_0)-R(\widehat{\omega}^\star)]}{M^c}+\mathbf{E}[\lambda]H(\omega^\star-\widehat{\omega}_0).$$

The result is then arrived at by plugging the above into the bound presented within Proposition 1, and recalling that $\widetilde{\omega}=\widehat{\omega}_\lambda$.