Foundations of Deep Learning HW3

Yonatan Ariel Slutzky Eyal Grinberg

1 June 2023

1 Part 1

1.1 Question 1

We will use the under line notation to denote tensors.
We will first prove the following lemma:

1.1.1 Lemma

Let $A \in \mathbb{R}^{n \times m}, x \in \mathbb{R}^m$. Then $\frac{\partial Ax}{\partial A}$ is a $n \times m \times n$ tensor of the following form:

$$\forall i \in [n], j \in [m].(\frac{\partial Ax}{\partial A})_{ij.} = \frac{\partial Ax}{\partial A_{ij}} = \begin{pmatrix} 0 \\ : \\ x_j \\ : \\ 0 \end{pmatrix}$$

The proof follows simple calculus and matrix multiplication rules:

$$\forall i \in [n], j \in [m].(\frac{\partial Ax}{\partial A})_{ij.} = \frac{\partial Ax}{\partial A_{ij}} = \frac{\partial}{\partial A_{ij}} \sum_{j'=1}^{m} x_{j'} A_{.j'} = \frac{\partial}{\partial A_{ij}} x_j A_{.j} = \begin{pmatrix} 0 \\ : \\ x_j \\ : \\ 0 \end{pmatrix} .\Box$$

We'll continue to our main proof. According to the assumption, there exists $\underline{W}' := (W^{(1)'}, .., W^{(N)'})$ such that

$$L(\underline{W}') < L(\underline{0}) = L(0, .., 0)$$

Assume on the contrary that L is convex. We will address $\langle ., . \rangle$ as a generalized dot product which can also be applied to tensors.
By a property of convex functions, we have that

$$L(\underline{W}') \geq L(\underline{0}) + \langle \nabla L(\underline{0}), \underline{W}' - \underline{0} \rangle = L(\underline{0}) + \langle \nabla L(\underline{0}), \underline{W}' \rangle =$$

$$= L(\underline{0}) + \sum_{n=1}^{N} \langle (\frac{\partial}{\partial W^{(n)}} L(\underline{W})|_{\underline{W}=\underline{0}})^\top, W^{(n)'} \rangle = (*)$$

Since L is solely dependant on the input-output mappings of the network, it can be re-written as a function of the result hypothesis. In other words, we can view it as

$$L(\underline{W}) = L(h_{\underline{W}}(x)) = L(W^{(N)}\sigma(..W^{(2)}\sigma(W^{(1)}x)..))$$

Thus, we have that

$$(*) = L(\underline{0}) + \sum_{n=1}^{N} \langle (\frac{\partial}{\partial W^{(n)}} L(W^{(N)}\sigma(..W^{(2)}\sigma(W^{(1)}x)..))|_{\underline{W}=\underline{0}})^{\top}, W^{(n)'} \rangle = (**)$$

Consider $n \in [N]$. Since L and σ are continuously differentiable, we can employ the chain rule and receive that

$$\frac{\partial}{\partial W^{(n)}} L(W^{(N)}\sigma(..W^{(2)}\sigma(W^{(1)}x)..))|_{\underline{W}=\underline{0}} =$$

$$(\frac{\partial}{\partial z} L(z)|_{z=h_{\underline{0}}(x)})^{\top} \cdot \frac{\partial}{\partial W^{(n)}} W^{(N)}\sigma(..W^{(2)}\sigma(W^{(1)}x)..)|_{\underline{W}=\underline{0}} = (***)$$

If $1 < n$, we can employ our lemma and plug in the fact that $\sigma(0) = 0$ and receive that

$$\frac{\partial}{\partial W^{(n)}} W^{(n)}\sigma(..W^{(2)}\sigma(W^{(1)}x)..)|_{\underline{W}=\underline{0}} = 0_{d_n \times d_{n-1} \times d_n}$$

Hence, employing the chain rule, we receive that

$$(***) = (\frac{\partial}{\partial z} L(z)|_{z=h_{\underline{0}}(x)})^{\top} \prod_{k=N}^{n+1} \frac{\partial W^{(k)}\sigma(..W^{(2)}\sigma(W^{(1)}x)..)}{\partial W^{(k-1)}\sigma(..W^{(2)}\sigma(W^{(1)}x)..)}|_{\underline{W}=\underline{0}} \frac{\partial W^{(n)}\sigma(..W^{(2)}\sigma(W^{(1)}x)..)}{\partial W^{(n)}}|_{\underline{W}=\underline{0}} =$$

$$= 0_{d_{n-1} \times d_n}$$

Otherwise if $n = 1$, we can note that by using the linearity of the derivative we receive

$$\frac{\partial W^{(2)}\sigma(W^{(1)}x)}{\partial W^{(1)}} = W^{(2)} \frac{\partial \sigma(W^{(1)}x)}{\partial W^{(1)}}$$

Thus, employing the chain rule we get

$$(***) = (\frac{\partial}{\partial z} L(z)|_{z=h_{\underline{0}}(x)})^{\top} \prod_{k=N}^{3} \frac{\partial W^{(k)}\sigma(..W^{(2)}\sigma(W^{(1)}x)..)}{\partial W^{(k-1)}\sigma(..W^{(2)}\sigma(W^{(1)}x)..)}|_{\underline{W}=\underline{0}} \frac{\partial W^{(2)}\sigma(W^{(1)}x)}{\partial W^{(1)}}|_{\underline{W}=\underline{0}} =$$

$$= (\frac{\partial}{\partial z} L(z)|_{z=h_{\underline{0}}(x)})^{\top} \prod_{k=N}^{3} \frac{\partial W^{(k)}\sigma(..W^{(2)}\sigma(W^{(1)}x)..)}{\partial W^{(k-1)}\sigma(..W^{(2)}\sigma(W^{(1)}x)..)}|_{\underline{W}=\underline{0}} \cdot 0_{d_2 \times d_1} \cdot \frac{\partial \sigma(W^{(1)}x)}{\partial W^{(1)}}|_{\underline{W}=\underline{0}} =$$

$$= 0_{d_0 \times d_1}$$

Finally, we have that

$$(**) = L(\underline{0}) + \sum_{n=1}^{N} \langle (0_{d_n \times d_{n-1}}, W^{(n)'} \rangle = L(0)$$

Contradicting our assumption. □

2 Part 2

2.1 Question 1

Let $t \in \mathbb{N} \cup \{0\}$. Using the lemma we saw in class, we have by definition of SGD that

$$f(w_{t+1}) \leq f(w_t) + \langle \nabla f(w_t), w_{t+1} - w_t \rangle + \frac{\beta}{2}||w_{t+1} - w_t||_2^2 =$$

$$= f(w_t) + \langle \nabla f(w_t), w_t - \eta_t(\nabla f(w_t) + \xi_t) - w_t \rangle + \frac{\beta}{2}||w_t - \eta_t(\nabla f(w_t) + \xi_t) - w_t||_2^2 =$$

$$= f(w_t) - \eta_t \langle \nabla f(w_t), \nabla f(w_t) + \xi_t \rangle + \frac{\eta_t^2 \beta}{2}||\nabla f(w_t) + \xi_t||_2^2 =$$

$$= f(w_t) - \eta_t ||\nabla f(w_t)||_2^2 - \eta_t \langle \nabla f(w_t), \xi_t \rangle + \frac{\eta_t^2 \beta}{2}||\nabla f(w_t)||_2^2 + \eta_t^2 \beta \langle \nabla f(w_t), \xi_t \rangle + \frac{\eta_t^2 \beta}{2}||\xi_t||_2^2 =$$

$$= f(w_t) + \frac{\eta_t^2 \beta - 2\eta_t}{2}||\nabla f(w_t)||_2^2 + (\eta_t^2 \beta - \eta_t)\langle \nabla f(w_t), \xi_t \rangle + \frac{\eta_t^2 \beta}{2}||\xi_t||_2^2 = (*)$$

Plugging the fact that $\beta = \frac{1}{\eta_t}$ results in

$$(*) = f(w_t) - \frac{1}{2\beta}||\nabla f(w_t)||_2^2 + \frac{1}{2\beta}||\xi_t||_2^2$$

Taking expected values on both sides we receive using linearity that

$$E(f(w_{t+1})) \leq E(f(w_t)) - \frac{1}{2\beta}E(||\nabla f(w_t)||_2^2) + \frac{1}{2\beta}E(||\xi_t||_2^2) =$$

$$= E(f(w_t)) - \frac{1}{2\beta}E(||\nabla f(w_t)||_2^2) + \frac{1}{2\beta}\sigma^2$$

Let $\epsilon > 0$. Assume that for steps $t \in [T-1] \cup \{0\}$ we have that

$$E(||\nabla f(w_t)||_2) > \epsilon + \sigma$$

Thus, we receive from Jensen's inequality that

$$E(||\nabla f(w_t)||_2^2) \geq E(||\nabla f(w_t)||_2)^2 > (\epsilon + \sigma)^2$$

Therefore, after T steps the following holds

$$f^* = E(f^*) \leq E(f(w_T)) \leq E(f(w_0)) - \frac{T}{2\beta}(\epsilon + \sigma)^2 + \frac{T}{2\beta}\sigma^2 = E(f(w_0)) - \frac{\epsilon^2 + 2\epsilon\sigma}{2\beta}T =$$

$$= f(w_0) - \frac{\epsilon^2 + 2\epsilon\sigma}{2\beta}T$$

Where the last equality stems from the fact that the init. of w_0 isn't related to the stochastic nature of SGD.
The above implies that

$$T \leq \frac{2\beta}{\epsilon^2 + 2\epsilon\sigma}(f(w_0) - f^*)$$

which is an upper bound on T as desired. □

2.2 Question 2

We established in the previous section that given $t \in \mathbb{N} \cup \{0\}$, the following holds:

$$f(w_{t+1}) \leq f(w_t) + \frac{\eta_t^2\beta - 2\eta_t}{2}||\nabla f(w_t)||_2^2 + (\eta_t^2\beta - \eta_t)\langle \nabla f(w_t), \xi_t \rangle + \frac{\eta_t^2\beta}{2}||\xi_t||_2^2$$

Taking expected values on both sides we receive using linearity that

$$E(f(w_{t+1})) \leq E(f(w_t)) + \frac{\eta_t^2\beta - 2\eta_t}{2}E(||\nabla f(w_t)||_2^2) + (\eta_t^2\beta - \eta_t)E(\langle \nabla f(w_t), \xi_t \rangle) + \frac{\eta_t^2\beta}{2}\sigma^2 =$$

$$= E(f(w_t)) + \frac{\eta_t^2\beta - 2\eta_t}{2}E(||\nabla f(w_t)||_2^2) + (\eta_t^2\beta - \eta_t)\sum_{i=1}^{d} E(\nabla f(w_t)_i, \xi_{t_i}) + \frac{\eta_t^2\beta}{2}\sigma^2 = (*)$$

Using the fact that $\nabla f(w_t)$ and ξ_t are independent and that $E(\xi_t) = 0$, we get that

$$(*) = E(f(w_t)) + \frac{\eta_t^2\beta - 2\eta_t}{2}E(||\nabla f(w_t)||_2^2) + (\eta_t^2\beta - \eta_t)\sum_{i=1}^{d} E(\nabla f(w_t)_i)E(\xi_{t_i}) + \frac{\eta_t^2\beta}{2}\sigma^2 =$$

$$= E(f(w_t)) + \frac{\eta_t^2\beta - 2\eta_t}{2}E(||\nabla f(w_t)||_2^2) + \frac{\eta_t^2\beta}{2}\sigma^2$$

We'll consider the learning rate $\eta_t = min\{\frac{1}{t+1}, \frac{1}{\beta}\}$. This learning rate promises that

$$\frac{\eta_t^2\beta - 2\eta_t}{2} < 0$$

We also receive that for any time T

$$\sum_{t=0}^{T-1} \frac{\eta_t^2\beta}{2} \leq \sum_{t=0}^{T-1} \frac{\beta}{2(t+1)^2} \leq \sum_{t=0}^{\infty} \frac{\beta}{2(t+1)^2} \leq \frac{\beta}{2}\frac{\pi^2}{6}$$

Lastly, using the harmonic sum bounds we get that

$$\sum_{t=0}^{T-1} \eta_t \geq \sum_{t=\lceil \beta - 1 \rceil}^{T-1} \frac{1}{t+1} \geq ln(T) - ln(\beta)$$

Let $\epsilon > 0$. Assume that for steps $t \in [T-1] \cup \{0\}$ we have that

$$E(||\nabla f(w_t)||_2) > \epsilon$$

Thus, we receive from Jensen's inequality that

$$E(||\nabla f(w_t)||_2^2) \geq E(||\nabla f(w_t)||_2)^2 > \epsilon^2$$

Therefore, after T steps the following holds

$$f^* = E(f^*) \leq E(f(w_T)) \leq E(f(w_0)) + \sum_{t=0}^{T-1} \frac{\eta_t^2\beta - 2\eta_t}{2}E(||\nabla f(w_t)||_2^2) + \sum_{t=0}^{T-1} \frac{\eta_t^2\beta}{2}\sigma^2 \leq$$

$$\leq E(f(w_0)) + \sum_{t=0}^{T-1} \frac{\eta_t^2\beta - 2\eta_t}{2}\epsilon^2 + \sum_{t=0}^{T-1} \frac{\eta_t^2\beta}{2}\sigma^2 \leq$$

$$\leq E(f(w_0)) + (\frac{\beta}{2}\frac{\pi^2}{6} - ln(T) + ln(\beta))\epsilon^2 + \frac{\beta}{2}\frac{\pi^2}{6}\sigma^2 \rightarrow$$

$$f^* - E(f(w_0)) - \frac{\beta}{2}\frac{\pi^2}{6}(\epsilon^2 + \sigma^2) - ln(\beta)\epsilon^2 \leq -ln(T)\epsilon^2 \rightarrow$$

$$ln(T) \leq \frac{E(f(w_0)) + \frac{\beta}{2}\frac{\pi^2}{6}(\epsilon^2 + \sigma^2) + ln(\beta)\epsilon^2 - f^*}{\epsilon^2} \rightarrow$$

$$T \leq exp(\frac{f(w_0) + \frac{\beta}{2}\frac{\pi^2}{6}(\epsilon^2 + \sigma^2) + ln(\beta)\epsilon^2 - f^*}{\epsilon^2})$$

In other words, after performing the amount of steps above, it is guaranteed that we have reached an ϵ-stationary point. □

2.3 Question 3

We'll first note that in this question, the gradient of ϕ w.r.t to $\underline{W}$ is a $N \times d \times d$ tensor.

2.3.1 Claim 1

We saw in class that the gradient of ϕ w.r.t the input tensor $\underline{W} \in \mathbb{R}^{N \times d \times d}$ takes the following form for any entry index $j \in [N]$ (i.e., when deriving ϕ w.r.t to W_j):

$$\nabla\phi(W_1, .., W_N)_j = W_{j+1:N}^\top \nabla l(W_{1:N}) W_{1:j-1}^\top$$

By definition, for any index $i \in [N]$, when deriving $\nabla\phi(W_1, .., W_N)_j$ w.r.t to W_i, the following holds:

$$\nabla^2\phi(W_1, .., W_N)_{ji} = \frac{\partial}{\partial W_i}\nabla\phi(W_1, .., W_N)_j = \frac{\partial}{\partial W_i} W_{j+1:N}^\top \nabla l(W_{1:N}) W_{1:j-1}^\top = (*)$$

We assume that $\nabla l(A)$ is a continuously differentiable function w.r.t to its input $A \in \mathbb{R}^{d \times d}$ (each of its elements acts as a continuously differentiable function).
The function $g(W_1, .., W_N) := W_{1:N}$ is a continuously differentiable function w.r.t the input tensor $\underline{W} \in \mathbb{R}^{N \times d \times d}$ (as a polynomial expression of the weight matrices).
Therefore, the function $\nabla l(g(\underline{W})) = \nabla l(W_{1:N})$ is also a continuously differentiable function w.r.t the input tensor $\underline{W} \in \mathbb{R}^{N \times d \times d}$ as a composite function.
Thus, the differentiated expression in $(*)$ is also a continuously differentiable function w.r.t the input tensor $\underline{W} \in \mathbb{R}^{N \times d \times d}$ as a composite function.
Hence, $\nabla^2\phi(W_1, .., W_N)_{ji}$ is continuous w.r.t to the input tensor $\underline{W} \in \mathbb{R}^{N \times d \times d}$.
This implies that when restricting the domain to the hyper-sphere $B \times B \times ... \times B$, we receive that the hessian $\nabla^2\phi(W_1, .., W_N)$ is bounded as it is continuous.
Applying a theorem regarding the relation between a Lipschitz continuous function and its bounded derivative, we have that $\nabla\phi(W_1, .., W_N)$ is a Lipschitz continuous function as desired. □

2.3.2 Claim 2

Suppose l is constant. Then we have that for any $j \in [N]$

$$\nabla\phi(W_1, .., W_N)_j = W_{j+1:N}^{\top} \nabla l(W_{1:N}) W_{1:j-1}^{\top} = W_{j+1:N}^{\top} 0_{d\times d} W_{1:j-1}^{\top} = 0_{d\times d}$$

Hence, for instance, ϕ's gradient is globally 1-Lipschitz and thus ϕ is 1-smooth. Suppose l is affine and $N = 2$. Thus, the gradient of l is a constant matrix which we denote $C \in \mathbb{R}^{d\times d}$. Then we have that

$$\nabla\phi(W_1, .., W_N)_1 = W_{2:2}^{\top} \nabla l(W_{1:2}) W_{1:0}^{\top} = W_2^{\top} \nabla l(W_{1:2}) I_d = W_2^{\top} C$$

$$\nabla\phi(W_1, .., W_N)_2 = W_{3:2}^{\top} \nabla l(W_{1:2}) W_{1:1}^{\top} = I_d \nabla l(W_{1:2}) W_1^{\top} = C W_1^{\top}$$

Thus, the following holds for any $W_1, W_2, \tilde{W}_1, \tilde{W}_2 \in \mathbb{R}^{d\times d}$ by the Cauchy-Schwarz inequality

$$||\nabla\phi(W_1, W_2) - \nabla\phi(\tilde{W}_1, \tilde{W}_2)|| =$$

$$= || \begin{pmatrix} W_2^{\top} C - \tilde{W}_2^{\top} C \\ C W_1^{\top} - C \tilde{W}_1^{\top} \end{pmatrix} || = || \begin{pmatrix} (W_2^{\top} - \tilde{W}_2^{\top}) C \\ C (W_1^{\top} - \tilde{W}_1^{\top}) \end{pmatrix} || =$$

$$|| \begin{pmatrix} C^{\top} (W_2 - \tilde{W}_2) \\ C (W_1^{\top} - \tilde{W}_1^{\top}) \end{pmatrix} || \leq ||C|| \cdot ||\underline{W} - \underline{\tilde{W}}||$$

Therefore by definition, the gradient is $||C||$-Lipschitz and thus ϕ is $||C||$-smooth. Suppose ϕ is β-smooth. Therefore, for any two sets of weight matrices $\underline{W_1}, \underline{W_2}$ the following holds:

$$\beta||\underline{W_1} - \underline{W_2}|| \geq ||\nabla\phi(\underline{W_1}) - \nabla\phi(\underline{W_2})|| \ (*)$$

Assume l isn't constant.
Assume on the contrary that either l isn't affine or $N > 2$.
Suppose l isn't affine. Thus, there exist $W_1, .., W_N, \tilde{W}_1, .., \tilde{W}_N \in \mathbb{R}^{d\times d}$ such that

$$D := \nabla l(W_{1:N}) - \nabla l(\tilde{W}_{1:N}) \neq 0$$

Due to the overparameterized nature of ϕ, the above weight matrices can be chosen WLOG in manner such that the first weight matrices are different - $W_1 \neq \tilde{W}_1$, and for the rest of the indices $j \in [N] \setminus \{1\}$, the weights align - $W_j = \tilde{W}_j$.
Therefore, the following holds from ϕ's β-smoothness and from the development of ϕ's gradient we saw in class:

$$\beta||W_1 - \tilde{W}_1|| = \beta||(W_1, .., W_N) - (\tilde{W}_1, .., \tilde{W}_N)|| \geq$$

$$\geq ||\nabla\phi(W_1, .., W_N) - \nabla\phi(\tilde{W}_1, .., \tilde{W}_N)|| \geq$$

$$\geq ||(\nabla\phi(W_1, .., W_N) - \nabla\phi(\tilde{W}_1, .., \tilde{W}_N))_1|| =$$

$$= ||W_{2:N}^{\top} \nabla l(W_{1:N}) W_{1:0}^{\top} - \tilde{W}_{2:N}^{\top} \nabla l(\tilde{W}_{1:N}) \tilde{W}_{1:0}^{\top}|| =$$

$$= ||W_{2:N}^{\top}(\nabla l(W_{1:N}) - \nabla l(\tilde{W}_{1:N}))|| = ||W_{2:N}^{\top} D||$$

Now, we can set the following for some $\alpha \in \mathbb{R}_{>1}$:

$$\hat{W}_2 := \alpha W_2 =: \bar{W}_2, \hat{W}_1 := \frac{1}{\alpha} W_1, \bar{W}_1 := \frac{1}{\alpha} \tilde{W}_1$$

$$\forall j \in [N] \setminus [2]. \hat{W}_j := W_j =: \bar{W}_j$$

For these, the following equalities will hold:

$$\beta ||\hat{W}_1 - \bar{W}_1|| = \beta \frac{1}{\alpha} ||W_1 - \tilde{W}_1||$$

$$\nabla l(\hat{W}_{1:N}) - \nabla l(\bar{W}_{1:N}) = \nabla l(W_{1:N}) - \nabla l(\tilde{W}_{1:N}) = D \rightarrow$$
$$||(\nabla \phi(\hat{W}_1, .., \hat{W}_N) - \nabla \phi(\bar{W}_1, .., \bar{W}_N))_1|| =$$
$$= ||\hat{W}_{2:N}^{\top} \nabla l(\hat{W}_{1:N}) \hat{W}_{1:0}^{\top} - \bar{W}_{2:N}^{\top} \nabla l(\bar{W}_{1:N}) \bar{W}_{1:0}^{\top}|| =$$
$$= ||\hat{W}_{2:N}^{\top}(\nabla l(\hat{W}_{1:N}) - \nabla l(\bar{W}_{1:N}))|| =$$
$$= ||\hat{W}_{2:N}^{\top} D|| = \alpha ||W_{2:N}^{\top} D||$$

Therefore, we have shown we can shrink the upper bounder by a factor of α, and increase the lower bounder by a factor of α. This implies that we can take $\alpha \rightarrow \infty$ and receive a contradiction to the argument in $(*)$.

Suppose $N > 2$. Since l is affine and isn't constant we have that its gradient w.r.t to any set of weights B is constant and different from 0.

Consider two sets of weights $W_1, .., W_N, \tilde{W}_1, .., \tilde{W}_N \in \mathbb{R}^{d \times d}$ such that the first weight matrices are different - $W_1 \neq \tilde{W}_1$, and for the rest of the indices $j \in [N] \setminus \{1\}$, the weights align - $W_j = \tilde{W}_j$.

Therefore, the following holds from ϕ's β-smoothness and from the development of ϕ's gradient we saw in class:

$$\beta ||W_1 - \tilde{W}_1|| = \beta ||(W_1, .., W_N) - (\tilde{W}_1, .., \tilde{W}_N)|| \geq$$

$$\geq ||\nabla \phi(W_1, .., W_N) - \nabla \phi(\tilde{W}_1, .., \tilde{W}_N)||$$
$$\geq ||(\nabla \phi(W_1, .., W_N) - \nabla \phi(\tilde{W}_1, .., \tilde{W}_N))_2|| =$$
$$= ||W_{3:N}^{\top} \nabla l(W_{1:N}) W_{1:1}^{\top} - \tilde{W}_{3:N}^{\top} \nabla l(\tilde{W}_{1:N}) \tilde{W}_{1:1}^{\top}|| =$$
$$= ||W_{3:N}^{\top} B (W_1 - \tilde{W}_1)^{\top}||$$

Now, we can set the following for some $\alpha \in \mathbb{R}_{>1}$:

$$\hat{W}_1 := W_1, \bar{W}_1 := \tilde{W}_1, \hat{W}_2 = \frac{1}{\alpha} W = \bar{W}_2, \hat{W}_3 := \alpha W_3 =: \bar{W}_3$$

$$\forall j \in [N] \setminus [3]. \hat{W}_j := W_j =: \bar{W}_j$$

For these, the following equalities will hold:

$$\beta ||\hat{W}_1 - \bar{W}_1|| = \beta ||W_1 - \tilde{W}_1||$$

$$||(\nabla\phi(\hat{W}_1,..,\hat{W}_N) - \nabla\phi(\bar{W}_1,..,\bar{W}_N))_2|| =$$
$$= ||\hat{W}_{3:N}^\top B\hat{W}_1^\top - \bar{W}_{3:N}^\top B\bar{W}_1^\top|| =$$
$$= \alpha||W_{3:N}^\top B(W_1 - \tilde{W}_1)^\top||$$

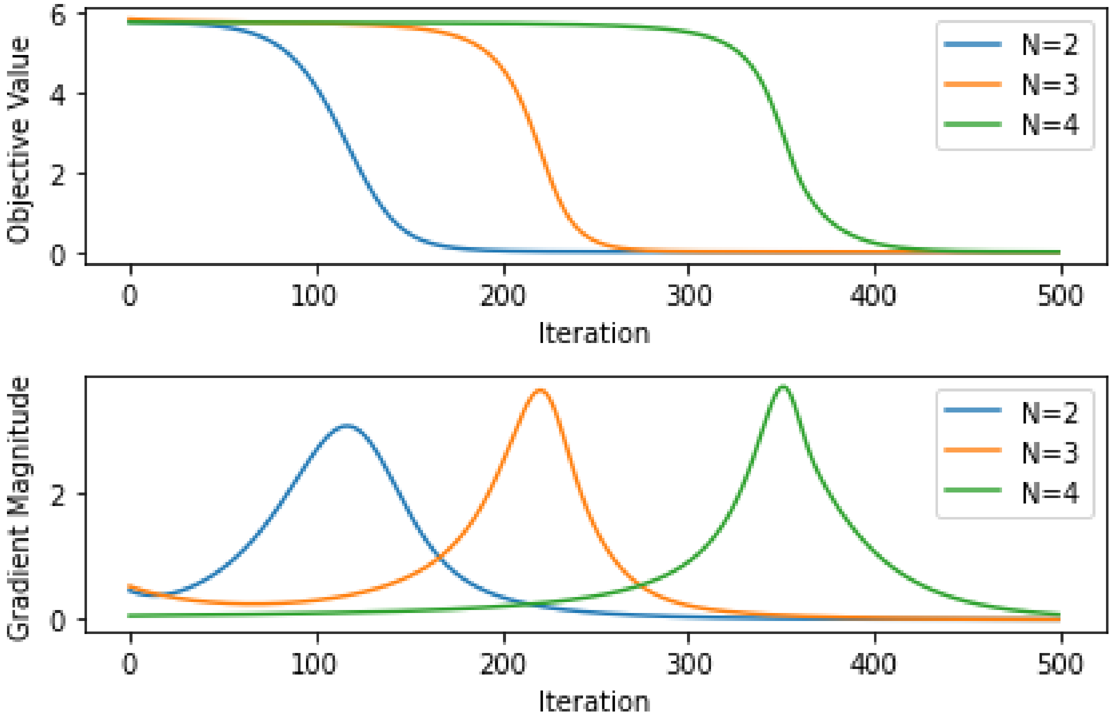
Therefore, we have shown we can keep the upper bounder with the same value, and increase the lower bounder by a factor of α. This implies that we can take $\alpha \to \infty$ and receive a contradiction to the argument in $(*)$.
The above completes the proof. □

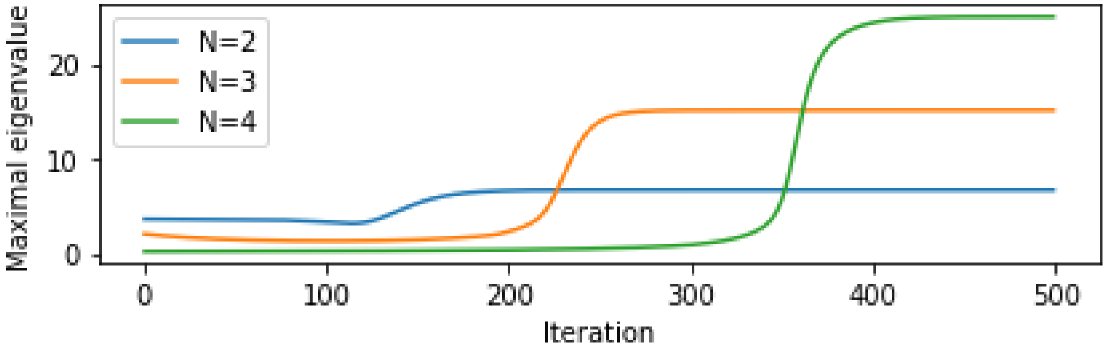
2.4 Question 4

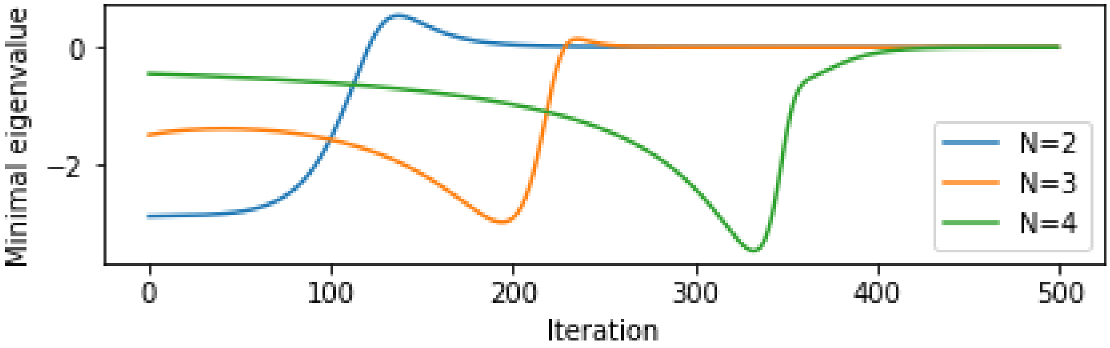
We trained our LNN over a random sample of size 100 that took the following form

$$\forall i \in [100].y_i = w^\top sin(x_i) + \epsilon_i$$

Where $x_i \sim N(0_{10}, I_{10}), \epsilon_i \sim N(0, 0.1^2)$, $w \sim N(0_{10}, I_{10})$, and the sine operates elementwise. Our LNNs had a hidden width of 1. The following are our results after 500 iterations:







As expected, the model was able to reach an almost perfect fit (the LNN can actually realize linear functions).
The gradients at convergence also reach a very small norm - as expected when reaching the optimum.
We can also see that the minimal eigenvalues of the hessians are very close to 0, and the maximal eigenvalues are positive. This indicates that the hessians are PSDs - which in turn will imply that we reached a global minima.

3 Part 3 - Linear Neural Networks

3.1 Question 1

Continuing the notation established in the proof sketch seen in class, we have the following for $t \in \mathbb{R}_{\geq 0}$ and $j \in [N]$:

$$W_{1:j}(t) = W_j(t)...W_1(t) = U_j\Sigma_j V_j^\top ...U_1\Sigma_1 V_1^\top =$$

$$= V_{j+1}D_j\Sigma V_j^\top V_j D_{j-1}\Sigma V_{j-1}^\top ...V_2 D_1 \Sigma V_1^\top = (*)$$

Since the V matrices are orthogonal, and since the D matrices and Σ are diagonal we have that

$$(*) = V_{j+1}\Sigma^j D_j ... D_1 V_1^\top$$

Therefore,

$$W_{1:j}(t)^\top W_{1:j}(t) = (V_{j+1}\Sigma^j D_j ... D_1 V_1^\top)^\top V_{j+1}\Sigma^j D_j ... D_1 V_1^\top =$$

$$= V_1 D_1 .. D_j \Sigma^j V_{j+1}^\top V_{j+1}\Sigma^j D_j ... D_1 V_1^\top = (**)$$

Since the inverse of the D matrices are themselves we have

$$(**) = V_1 \Sigma^{2j} V_1^\top = (V_1 \Sigma V_1^\top)^{2j} = (***)$$

In particular, since $W_{1:N}(t)^\top W_{1:N}(t) = (V_1 \Sigma V_1^\top)^{2N}$, we have that

$$(***) = (W_{1:N}(t)^\top W_{1:N}(t))^{\frac{j}{N}}$$

as desired. □

3.2 Question 2

We'll plug the definition of W, derive and plug the definition of $\dot{U(t)}$ to receive the following dynamics:

$$\dot{W(t)} = \overset{\bullet}{(U(t)U(t)^\top)} = \dot{U(t)}U(t)^\top + U(t)\dot{U(t)}^\top =$$

$$= -\nabla\phi(U(t))U(t)^\top - U(t)\nabla\phi(U(t))^\top = (*)$$

We continue to find an expression for $\nabla\phi(U(t))$ using $\nabla l(U(t)U(t)^\top)$.
Applying the overparameterization of l using ϕ, we get by the first order taylor expansion of l around $U(t)U(t)^\top$ that the following holds:

$$\forall \Delta \in \mathbb{R}^{d\times d}.\ \phi(U(t){+}\Delta) = l((U(t){+}\Delta)(U(t){+}\Delta)^\top) = l(U(t)U(t)^\top{+}U(t)\Delta^\top{+}\Delta U(t)^\top{+}\Delta\Delta^\top) =$$

$$= l(U(t)U(t)^\top) + \langle \nabla l(U(t)U(t)^\top), U(t)\Delta^\top + \Delta U(t)^\top \rangle + o(||\Delta||_{Fro}) =$$

$$= l(U(t)U(t)^\top){+}\langle \nabla l(U(t)U(t)^\top), U(t)\Delta^\top \rangle{+}\langle \nabla l(U(t)U(t)^\top), \Delta U(t)^\top \rangle{+}o(||\Delta||_{Fro})$$

The above is a first order taylor expansion of ϕ around $U(t)$, and thus we get the following equalities:

$$\phi(U(t)) = l(U(t)U(t)^\top)$$

$$\langle \nabla\phi(U(t)), \Delta \rangle = \langle \nabla l(U(t)U(t)^\top), U(t)\Delta^\top \rangle + \langle \nabla l(U(t)U(t)^\top), \Delta U(t)^\top \rangle = (**)$$

Using the cyclic property of the trace, as-well as the relation between inner product of matrices and the trace, we have the following:

$$(**) = tr(\nabla l(U(t)U(t)^\top)(U(t)\Delta^\top)^\top) + tr(\nabla l(U(t)U(t)^\top)(\Delta U(t)^\top)^\top) =$$

$$= tr(\nabla l(U(t)U(t)^\top)\Delta U(t)^\top) + tr(\nabla l(U(t)U(t)^\top)U(t)\Delta^\top) =$$

$$= tr(U(t)^\top \nabla l(U(t)U(t)^\top)\Delta) + \langle \nabla l(U(t)U(t)^\top)U(t), \Delta \rangle =$$

$$= \langle (U(t)^\top \nabla l(U(t)U(t)^\top))^\top, \Delta \rangle + \langle \nabla l(U(t)U(t)^\top)U(t), \Delta \rangle =$$

$$= \langle \nabla l(U(t)U(t)^\top)^\top U(t), \Delta\rangle + \langle \nabla l(U(t)U(t)^\top)U(t), \Delta\rangle =$$
$$= \langle \nabla l(U(t)U(t)^\top)^\top U(t) + \nabla l(U(t)U(t)^\top)U(t), \Delta\rangle$$

This implies the following

$$\nabla\phi(U(t)) = \nabla l(U(t)U(t)^\top)^\top U(t) + \nabla l(U(t)U(t)^\top)U(t)$$

Thus, the dynamics $(*)$ are as follows:

$$-(\nabla l(U(t)U(t)^\top)^\top U(t)+\nabla l(U(t)U(t)^\top)U(t))U(t)^\top - U(t)(\nabla l(U(t)U(t)^\top)^\top U(t)+\nabla l(U(t)U(t)^\top)U(t))^\top =$$
$$= -(\nabla l(U(t)U(t)^\top)^\top + \nabla l(U(t)U(t)^\top))U(t)U(t)^\top - U(t)U(t)^\top(\nabla l(U(t)U(t)^\top)^\top + \nabla l(U(t)U(t)^\top))^\top$$

which is the desired expression. □

3.3 Question 3

In the case where $D_N = 1$, the end to end matrix $W_{1:N}(t)$ is a row vector. We saw in class the following dynamics of the end to end matrix:

$$\dot{W_{1:N}}(t) = -\sum_{j=1}^{N}(W_{1:N}(t)W_{1:N}(t)^\top)^{\frac{j-1}{N}}\nabla l(W_{1:N}(t))(W_{1:N}(t)^\top W_{1:N}(t))^{\frac{N-j}{N}} =$$

$$= -\sum_{j=1}^{N}||W_{1:N}(t)||^{\frac{2j-2}{N}}\nabla l(W_{1:N}(t))(W_{1:N}(t)^\top W_{1:N}(t))^{\frac{N-j}{N}} = (*)$$

The symmetric matrix $W_{1:N}(t)^\top W_{1:N}(t)$ is of rank 1, and thus it has a single non-zero eigenvalue. The following also holds:

$$W_{1:N}(t)^\top W_{1:N}(t)W_{1:N}(t)^\top = ||W_{1:N}(t)||^2 W_{1:N}(t)^\top$$

Therefore, the only non-zero eigenvalue is $||W_{1:N}(t)||^2$, and its appropriate normalized eigenvector is $\frac{W_{1:N}(t)^\top}{||W_{1:N}(t)||}$. We'll complete the eigenvalue decomposition with the following arbitrary eigenvectors $v_2, .., v_{d_0} \in \mathbb{R}^{d_0}$ (for instance, using the Gram-Schmidt procedure). Thus, we have that

$$W_{1:N}(t)^\top W_{1:N}(t) = (\frac{W_{1:N}(t)^\top}{||W_{1:N}(t)||}, v_2, .., v_{d_0})diag(||W_{1:N}(t)||^2, 0, .., 0)(\frac{W_{1:N}(t)^\top}{||W_{1:N}(t)||}, v_2, .., v_{d_0})^\top$$

This implies that for any $p \in \mathbb{R}_{>0}$,

$$(W_{1:N}(t)^\top W_{1:N}(t))^p = ||W_{1:N}(t)||^{2p}\frac{W_{1:N}(t)^\top W_{1:N}(t)}{||W_{1:N}(t)||^2} = ||W_{1:N}(t)||^{2p-2}W_{1:N}(t)^\top W_{1:N}(t)$$

Therefore,

$$(*) = -||W_{1:N}(t)||^{\frac{2N-2}{N}}\nabla l(W_{1:N}(t)) - \sum_{j=1}^{N-1}||W_{1:N}(t)||^{\frac{2j-2}{N}}\nabla l(W_{1:N}(t))||W_{1:N}(t)||^{2\frac{N-j}{N}-2}W_{1:N}(t)^\top W_{1:N}(t) =$$

$$= -||W_{1:N}(t)||^{\frac{2N-2}{N}} \nabla l(W_{1:N}(t)) - (N-1)||W_{1:N}(t)||^{\frac{-2}{N}} \nabla l(W_{1:N}(t)) W_{1:N}(t)^\top W_{1:N}(t) =$$

$$= -||W_{1:N}(t)||^{\frac{2N-2}{N}} \nabla l(W_{1:N}(t)) - (N-1)||W_{1:N}(t)||^{\frac{-2}{N}} \langle \nabla l(W_{1:N}(t)), W_{1:N}(t) \rangle W_{1:N}(t)$$

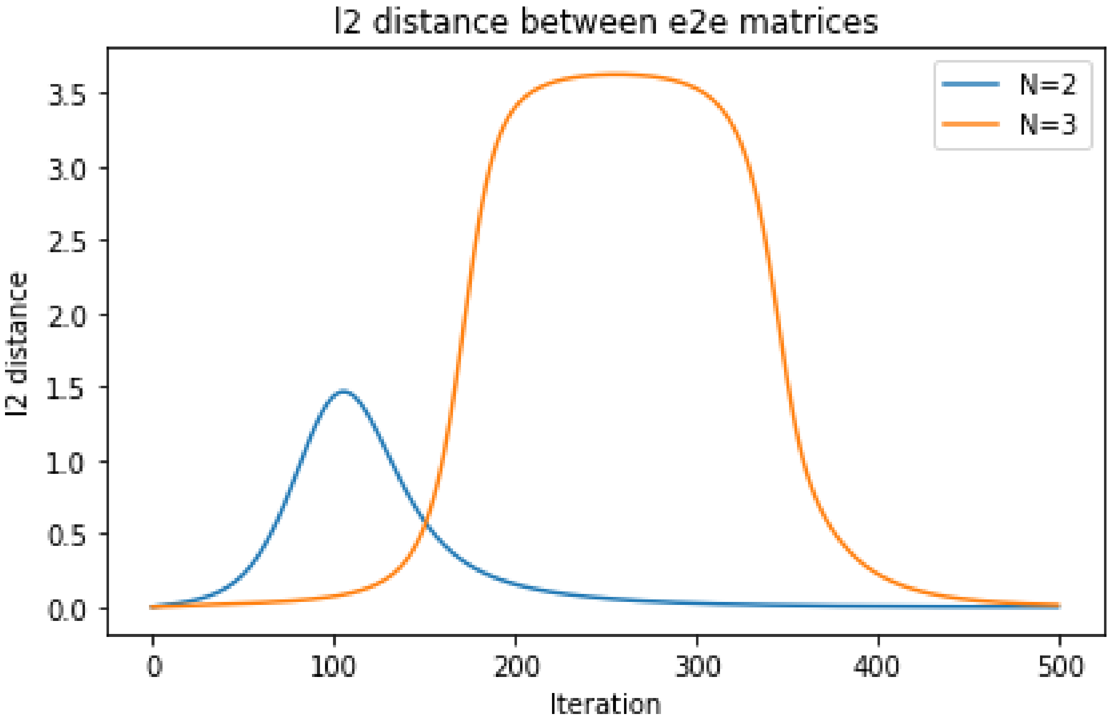
These simplified dynamics resonate with the notion of “promoting movement in direction already taken” since the dynamics factor the current state into the movement of the weights (and not just the loss itself).

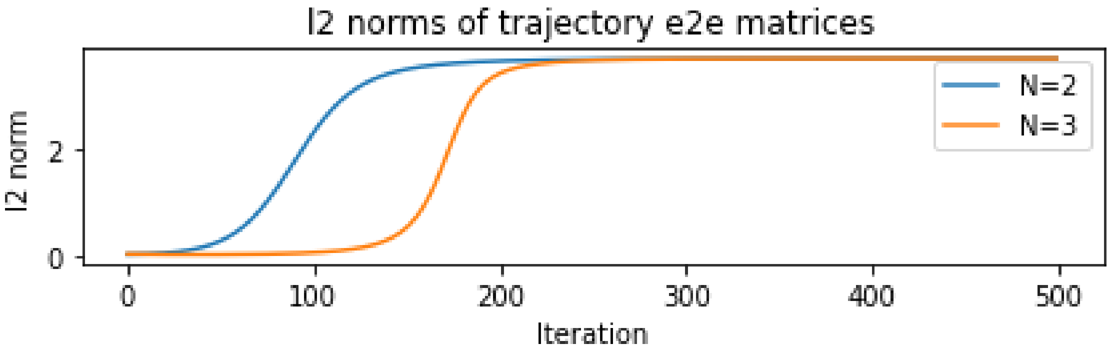
3.4 Question 4

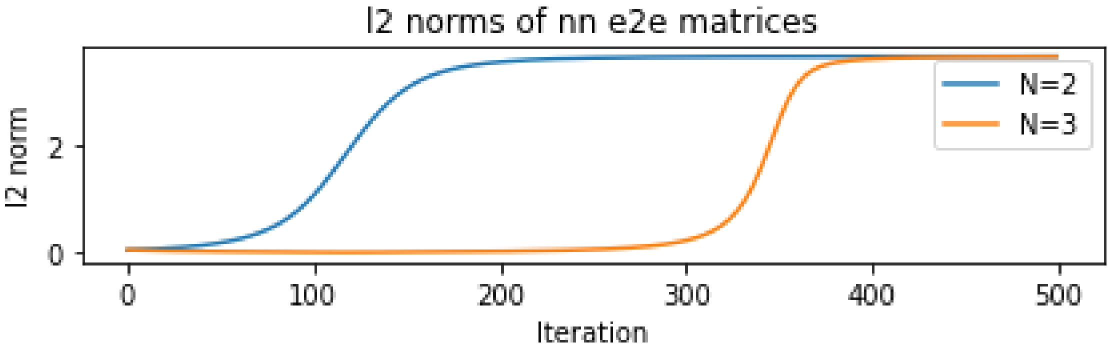
We trained our LNN over a random sample of size 100 that took the following form

$$\forall i \in [100]. y_i = w^\top sin(x_i) + \epsilon_i$$

Where $x_i \sim N(0_{10}, I_{10}), \epsilon_i \sim N(0, 0.1^2)$, $w \sim N(0_{10}, I_{10})$, and the sine operates elementwise. Our LNNs had a hidden width of 1. The following are our results after 500 iterations:







We compared the $e2e$ matrices using the $l2$ distance between them. As expected, the matrices seem to converge, with the deeper network taking more iterations to do so.
This could be explained by the fact that the deeper network has more parameters that need to be optimized.

4 Part 3 - Ultra Wide Neural Networks

4.1 Question 1

The following holds:

$$\frac{d}{dt}||u(t)-y||^2 = \frac{d}{dt}(u(t)-y)^\top(u(t)-y) = \frac{d}{dt}(u(t)^\top u(t) - 2y^\top u(t) + y^\top y) =$$

$$= 2\dot{u(t)}^\top(u(t)-y) = -2(u(t)-y)^\top H^{*\top}(u(t)-y) = (*)$$

From a theorem in the recitation, we have that

$$(*) \leq -2\lambda_{min}||u(t)-y||^2$$

The term $||u(t)-y||^2$ is non-negative, and thus we have shown above that its derivative is non-positive.

We'll assume WLOG that $||u(t) - y||^2 > 0$ for any t, otherwise, after reaching 0 at some point, the argument we have shown for the derivative implies that we stay at 0 for any consecutive t (and thus the convergence is faster than exponential).
Therefore, the following holds from the fundamental theorem of calculus:

$$\frac{\frac{d}{dt}||u(t) - y||^2}{||u(t) - y||^2} \leq -2\lambda_{min} \rightarrow \int_0^t \frac{\frac{d}{d\tau}||u(\tau) - y||^2}{||u(\tau) - y||^2} d\tau \leq \int_0^t -2\lambda_{min} d\tau \rightarrow$$

$$ln(||u(\tau) - y||^2)|_{\tau=0}^t \leq -2\lambda_{min}\tau|_{\tau=0}^t \rightarrow$$

$$ln(||u(t) - y||^2) - ln(||u(0) - y||^2) \leq -2\lambda_{min} t \rightarrow$$

$$||u(t) - y||^2 \leq ||u(0) - y||^2 exp(-2\lambda_{min} t)$$

The above implies that $||u(t)-y||^2 \rightarrow 0$ exponentially fast, which in turn implies that $u(t) \rightarrow y$ exponentially fast. □

4.2 Question 2

We'll denote the maximal and minimal eigenvalues of H^* to be λ^*_{max} and λ^*_{min}.
Similarly, we'll denote for $t \in \mathbb{R}_{\geq 0}$ the eigenvalues of $H(t)$ to be $\lambda_{max}(t)$ and $\lambda_{min}(t)$.
Both H^* and $H(t)$ are PSDs and thus $\lambda^*_{max}, \lambda^*_{min}, \lambda_{max}(t), \lambda_{min}(t) \geq 0$.
We'll denote the network predictions under the ideal dynamics to be $u^*(t)$, and the network predictions under the dynamics of this section to be $u(t)$.
As we've shown in the previous section, $u^*(t)$ adheres to the following inequality:

$$||u^*(t) - y||^2 \leq ||u^*(0) - y||^2 exp(-2\lambda^*_{min} t)$$

We can apply the same analysis in section 1 to $u(t)$ since in the analysis the gram matrix isn't derived, which means it being time-dependent in this case won't affect the analysis. Thus,

$$||u(t) - y||^2 \leq ||u(0) - y||^2 exp(-2\lambda_{min}(t) t)$$

Therefore, applying the triangle inequality, we have the following:

$$||u^*(t) - u(t)||^2 \leq ||u^*(0) - y||^2 exp(-2\lambda^*_{min} t) + ||u(0) - y||^2 exp(-2\lambda_{min}(t) t) = (*)$$

The initial state is the same, implying that

$$(*) = ||u(0) - y||^2 (exp(-2\lambda^*_{min} t) + exp(-2\lambda_{min}(t) t)) = (**)$$

It holds for any $x \geq 0$ that $1 + x \geq exp(-x)$, implying that

$$(**) \leq ||u(0) - y||^2 (1 + 2\lambda^*_{min} t + 1 + 2\lambda_{min}(t) t) = 2||u(0) - y||^2 (1 + t(\lambda^*_{min} + \lambda_{min}(t))) = (***)$$

By definition of the spectral norm, for any matrix with an orthogonal eigenvalue decomposition $A = U\Lambda U^\top$, the following holds:

$$||A||_{spectral} = \sqrt{\lambda_{max}(A^\top A)} = \sqrt{\lambda_{max}(U\Lambda U^\top U \Lambda U^\top)} = \sqrt{\lambda_{max}(U\Lambda^2 U^\top)} = max\{|\lambda_{max}(A)|, |\lambda_{min}(A)|\}$$

The matrix $H(t) - H^*$ is symmetric and thus the above claim holds for it too. Namely, the following holds after applying the assumption

$$\epsilon \geq ||H(t) - H^*||_{spectral} = max\{|\lambda_{max}(H(t) - H^*)|, |\lambda_{min}(H(t) - H^*)|\}$$

By definition, the eigenvalues of $-H^*$ are the opposite sign eigenvalues of H^*. Applying Weyl's theorem and the above claim, we have the following inequality:

$$\lambda_{max}(t) - \lambda^*_{max} \leq \lambda_{max}(H(t) - H^*) \leq \epsilon$$

This implies the following

$$\lambda^*_{min} + \lambda_{min}(t) \leq \lambda^*_{min} + \lambda_{max}(t) + \lambda^*_{max} - \lambda^*_{max} \leq \lambda^*_{min} + \lambda^*_{max} + \epsilon$$

Therefore, we have that

$$(***) \leq 2||u(0) - y||^2(1 + t(\lambda^*_{min} + \lambda^*_{max} + \epsilon)) = O(t\epsilon)$$

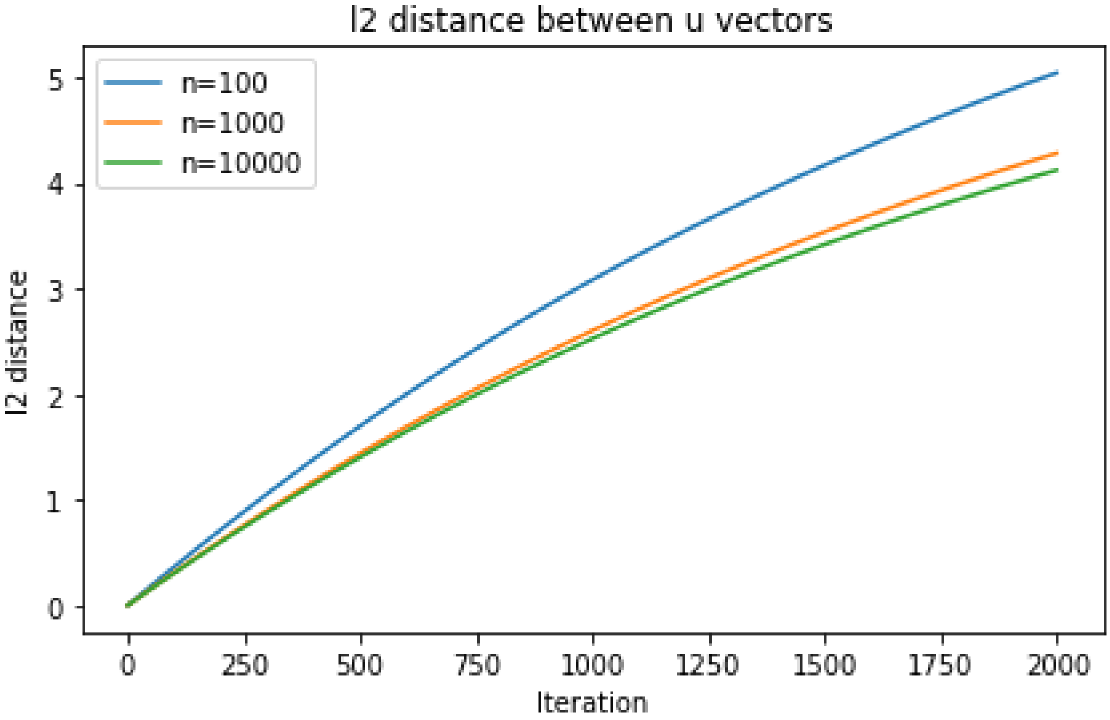
Taking square roots over the leftmost and rightmost sides of the inequality (which are non-negative) gives the desired result. □

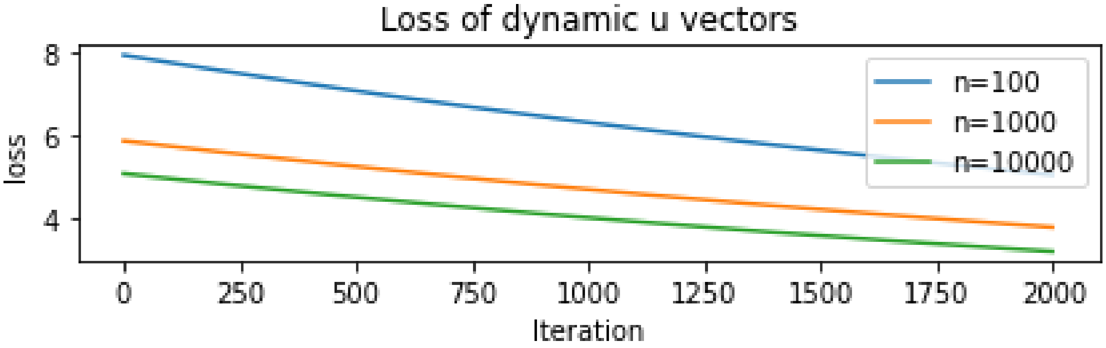
4.3 Question 3

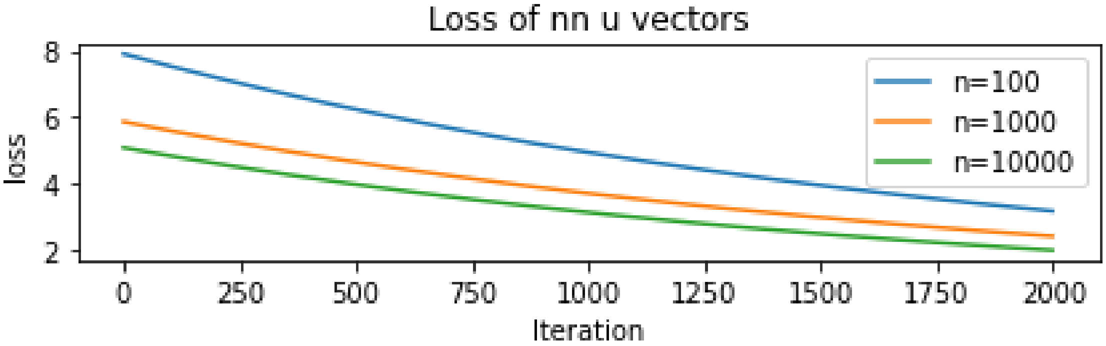
We trained our LNN over a random sample of size 100 that took the following form

$$\forall i \in [100].y_i = w^\top sin(x_i) + \epsilon_i$$

Where $x_i \sim N(0_{10}, I_{10}), \epsilon_i \sim N(0, 0.1^2)$, $w \sim N(0_{10}, I_{10})$, and the sine operates elementwise. Our LNNs had hidden widths of $100, 1000, 10000$. We used a learning rate of 0.001 for the neural net, and a learning rate of 0.00001 for the discrete implementation. The following are our results after 2000 iterations:


l2 distance between u vectors
n=100
n=1000
n=10000
l2 distance
0
1
2
3
4
5
Iteration
0
250
500
750
1000
1250
1500
1750
2000





We compared the different types of u vectors using the $l2$ distance between them, as well as comparing their $l2$ loss over the training set.
We do see an improvement as the hidden width grows, but not a significant one.