Assignment 3: Optimization

Dana Arad 311183321
Amit Elyasi 316291434

Part 1: Optimization in deep learning is non-convex

1. Let us use the notation from the backpropagation recitation:

$$J_{f_{t+1}}(h_t)\begin{cases} W_{t+1} & for\ linear\ layer \\ diag\left(\sigma'(h_t)\right) & for\ activation\ layer \end{cases}$$

With:

$$h_{t+1} = f_{t+1}(h_t) = W_{t+1}h_t$$

for t+1 being a linear layer, and:

$$h_{t+1} = f_{t+1}(h_t) = \sigma(h_t)$$

for t+1 being an activation layer, and

$$f(x) = f_N \circ f_{N-1} \circ \ldots \circ f_1(x).$$

Now to find the gradient of L we use the chain rule:

$$J_{L(W_1,\ldots,W_N)}(W_t) = J_{f_t}(h_t)J_{W\mapsto Wh_{t-1}}(W_t) = J_{f_t}(h_t)^T h_{t-1}^T$$

And denote by $\nabla_{W_t} L(W_1, \ldots, W_N)$ the arrangement of $J_{L(W_1,\ldots,W_N)}(W_t)$ as a matrix (as we did in class).

Note that $\sigma(0) = 0 \Rightarrow f(x) \equiv 0$ where f is the network with $W_1 = W_2 = \cdots = W_N = 0$, so we'll get

$$\forall t, \quad J_{L(W_1,\ldots,W_N)}(W_t) = J_{L(W_1,\ldots,W_N)}(0) = J_{f_t}(h_t)^T h_{t-1}^T = J_{f_t}(h_t)^T \cdot 0 = 0$$

Now we'll assume by a way of contradiction that L is convex, so by the fact that every point W where $J_{L(W_1,\ldots,W_N)}(W_t) = 0$ is a global minimum we'll get that $(0,0,\ldots,0)$ is a global minimum of L, contradiction.

Part 2: Landscape approach

1. $w_{t+1} = w_t - \eta_t(\nabla f(w_t) + \xi_t)$
   f is β-smooth and twice continuously differentiable. From the lemma we saw in class:

$$\forall w_1, w_2: |f(w_2) - (f(w_1) + \langle \nabla f(w_1), w_2 - w_1 \rangle)| \leq \frac{\beta}{2}\|w_2 - w_1\|^2$$

   Plug in $w_{t+1}, and\ w_t$ we get:

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

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

$$and$$

$$\begin{aligned} f(w_{t+1}) - f(w_t) - \langle \nabla f(w_t), \eta_t(\nabla f(w_t) + \xi_t) \rangle \\ = f(w_{t+1}) - f(w_t) - \eta_t\|\nabla f(w_t)\|^2 - \eta_t \nabla f(w_t)^T \xi_t \\ = f(w_{t+1}) - f(w_t) - \eta_t\|\nabla f(w_t)\|^2 - \eta_t \sum_{i=1}^{d} \nabla f(w_t)_i (\xi_t)_i \end{aligned}$$

$$and$$

$$\frac{\beta}{2}\|\eta_t(\nabla f(w_t) + \xi_t)\|^2 = \frac{\beta\eta_t^2}{2}\|(\nabla f(w_t) + \xi_t)\|^2 = \frac{\eta_t}{2}\|(\nabla f(w_t) + \xi_t)\|^2$$

$$\Downarrow$$

$$-\frac{\eta_t}{2}\|(\nabla f(w_t) + \xi_t)\|^2 \leq f(w_{t+1}) - f(w_t) - \eta_t\|\nabla f(w_t)\|^2 - \eta_t \sum_{i=1}^{d} \nabla f(w_t)_i (\xi_t)_i$$

Now assume $\|(\nabla f(w_t) + \xi_t)\|^2 < M^2$ for all t (can choose M to be the maximum of it over t) and take expected value on both sides:

$$\begin{aligned} -\mathbb{E}\left[\frac{\eta_t M^2}{2}\right] = -\frac{\eta_t M^2}{2} \\ \leq \mathbb{E}[f(w_{t+1}) - f(w_t)] - \eta_t \mathbb{E}[\|\nabla f(w_t)\|^2] \\ - \eta_t \sum_{i=1}^{d} \mathbb{E}[\nabla f(w_t)_i (\xi_t)_i] \end{aligned}$$

Note that $\nabla f(w_t)\ and\ \xi_t$ are independent for all t, so $\mathbb{E}[\nabla f(w_t)_i (\xi_t)_i] = \mathbb{E}[\nabla f(w_t)_i]\mathbb{E}[(\xi_t)_i]$, and by the assumption that $\mathbb{E}[\xi_t] = 0$ we will get:

$$-\eta_t \sum_{\mathrm{i}=1}^{\mathrm{d}} \mathbb{E}[\nabla f(w_t)_i (\xi_t)_i] = -\eta_t \sum_{\mathrm{i}=1}^{\mathrm{d}} \mathbb{E}[\nabla f(w_t)_i] \mathbb{E}[(\xi_t)_i]$$

$$= -\eta_t \sum_{\mathrm{i}=1}^{\mathrm{d}} \mathbb{E}[\nabla f(w_t)_i] \cdot 0 = 0$$

$$\Downarrow$$

$$-\frac{\eta_t M^2}{2} \leq \mathbb{E}[f(w_{t+1}) - f(w_t)] - \eta_t \mathbb{E}[\|\nabla f(w_t)\|^2]$$

Rearrange and divide by $-\eta_t$ (note that $\eta_t = \frac{1}{\beta} > 0$)

$$\mathbb{E}[\|\nabla f(w_t)\|^2] \leq \frac{M^2}{2} + \frac{\mathbb{E}[f(w_{t+1}) - f(w_t)]}{\eta_t} = \frac{M^2}{2} + \beta \mathbb{E}[f(w_{t+1}) - f(w_t)]$$

Now we'll apply this equation for all t=1,2,…,T and sum them all up, also note that we can assume $f^* = f(w_T)$ because f is smooth, which means the actual minimum will be close to the output of SGD:

$$\sum_{t=1}^{T} \mathbb{E}[\|\nabla f(w_t)\|^2]$$

$$\leq \sum_{t=1}^{T} \frac{M^2}{2} + \beta \mathbb{E}[f(w_{t+1}) - f(w_t)]$$

$$= \frac{TM^2}{2} + \beta \sum_{t=1}^{T} \mathbb{E}[f(w_{t+1})] - \mathbb{E}[f(w_t)]$$

$$= \frac{TM^2}{2} + \beta \mathbb{E}[f^*] - \beta \mathbb{E}[f(w_1)] = \frac{TM^2}{2} + \beta(f^* - f(w_1))$$

Divide both sides by T:

$$\min_t \mathbb{E}[\|\nabla f(w_t)\|^2] \leq \frac{\sum_{t=1}^{T} \mathbb{E}[\|\nabla f(w_t)\|^2]}{T} \leq \frac{M^2}{2} + \frac{\beta}{T}(f^* - f(w_1)) \overset{?}{\leq} \epsilon + \sigma$$

We want to "force" the right most inequality and Isolate T:

$$\frac{M^2}{2} + \frac{\beta}{T}(f^* - f(w_1)) \leq \epsilon + \sigma$$

$$\Updownarrow$$

$$T\left(\frac{M^2}{2} - \epsilon - \sigma\right) = T\frac{\left(M^2 - 2(\epsilon + \sigma)\right)}{2} \leq -\beta(f^* - f(w_1)) = \beta(f(w_1) - f^*)$$

$$\Updownarrow$$

$$T \leq \frac{2\beta(f(w_1) - f^*)}{M^2 - 2(\epsilon + \sigma)}$$

■

3.

- (*) Corollary from calculus 3: let $f: U \subseteq \mathbb{R}^n \to \mathbb{R}^m$ be a continuously differentiable function, let $x, y \in U$ such that $[x;y] \subseteq U$, then $|f(y) - f(x)| \leq \max_{a \in [x;y]} \left\| D_f(a) \right\| |x - y|$.

  Denote:

  $$\phi_B(\cdot) = \phi|_{B^{d,d} \times \ldots \times B^{d,d}}(\cdot)$$

  Let $W = (W_1, \ldots, W_N)$, $Q = (Q_1, \ldots, Q_N) \in B^{d,d} \times B^{d,d} \times \cdots \times B^{d,d} = B$.

  Note that in our case B is connected space, so for any $W, Q \in B$ we get $[W;Q] \subseteq B$, and so the corollary holds for any $W, Q \in B$.

  Now since $\nabla\phi_B$ is continuous and its domain is bounded, $\nabla\phi_B$ is bounded,

  denote $M \coloneqq \max_{W \in B} \|\nabla\phi_B(W)\|$.

  put it all together:

$$\|\nabla\phi_B(W) - \nabla\phi_B(Q)\| \overset{(*)}{\leq} \max_{A \in [W;Q]} \left\| D_f(A) \right\| |W - Q| \overset{[W;Q] \subseteq B}{\leq} \max_{W \in B} \|\nabla\phi_B(W)\| \, \|W - Q\| = M\|W - Q\|$$

■

- (i) $l(\cdot) \equiv C \in \mathbb{R} \Leftrightarrow \phi(\cdot) \equiv C \Leftrightarrow \nabla\phi(\cdot) \equiv C_2 (= 0) \Leftrightarrow \forall W, Q \in \mathbb{R}^{d,d} \times \mathbb{R}^{d,d} \times \ldots \times \mathbb{R}^{d,d}, \|\nabla\phi_B(W) - \nabla\phi_B(Q)\| = 0 \leq \|W - Q\|$

  (ii) $\|\nabla\phi_B(W) - \nabla\phi_B(Q)\| \leq C\|W - Q\|$

  $$\phi(W_1, W_2, \ldots, W_N) = l(W_1 W_2 \ldots W_N) = C W_1 W_2 \ldots W_N + A, \quad C, A \in \mathbb{R}^{d,d} \times \mathbb{R}^{d,d} \times \ldots \times \mathbb{R}^{d,d}$$

4. Each row displays the relevant plots for every value of N (2,3,4). In each row you can find the loss, magnitude of it's gradient and minimum and maximum eigenvalues of it's hessian, for every epoch. we used a very small learning rate (0.00000001) and initialized the network weights with Xavier initialization.

we see that for all depths the loss function decreases over the number of epochs, and also the magnitude of the gradient – meaning that we are indeed moving toward the global minimum. The minimum and maximum eigenvalues of the hessian do not seem to change over the course of the training.

Part 3: Trajectory approach

Linear neural networks

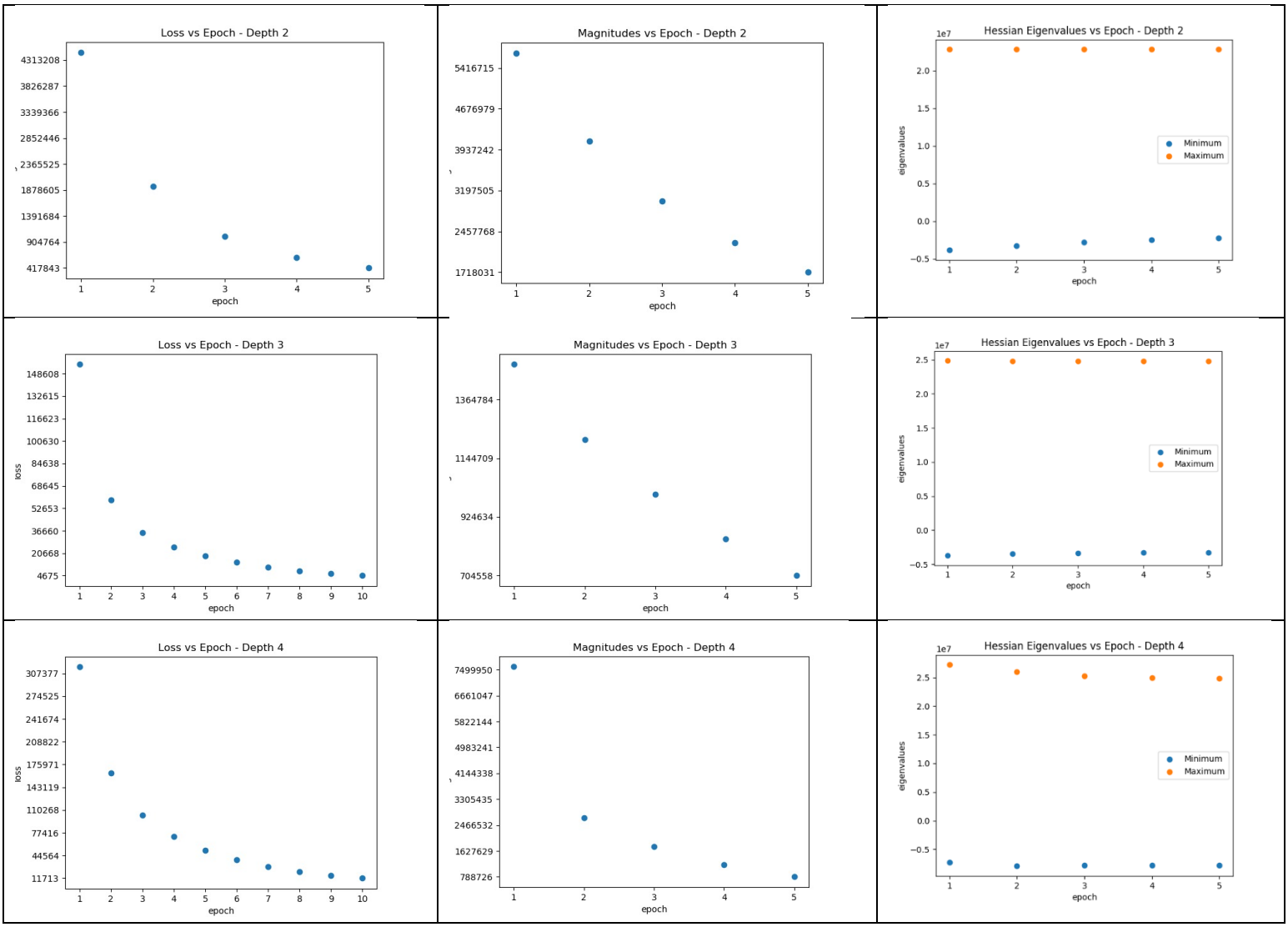
1. We've seen in class that $\forall t, j, \quad W_{j+1}^T(t)W_{j+1}(t) = W_j(t)W_j^T(t) \quad (*)$

$$W_{1:j}(t)^T W_{1:j}(t) \overset{by\ def}{=} W_j^T(t)W_{j-1}^T(t)\cdots W_2^T(t)W_1^T(t)W_1(t)W_2(t)\cdots W_{j-1}(t)W_j(t)$$
$$\overset{(*)}{=} W_j^T(t)W_{j-1}^T(t)\cdots W_2^T(t)W_2(t)W_2^T(t)W_2(t)\cdots W_{j-1}(t)W_j(t)$$
$$\overset{(*)}{=} W_j^T(t)W_{j-1}^T(t)\cdots [W_3^T(t)W_3(t)]^3\cdots W_{j-1}(t)W_j(t) \overset{(*)}{=} \cdots \overset{(*)}{=} \left[W_j^T(t)W_j(t)\right]^j$$

Now apply the same equality on $j = N$:

$$W_{1:N}(t)^T W_{1:N}(t) = [W_N^T(t)W_N(t)]^N \Rightarrow [W_{1:N}(t)^T W_{1:N}(t)]^{\frac{1}{N}} = W_N^T(t)W_N(t)$$

And use (*) to get:



$$\left[W_j^T(t)W_j(t)\right]^j \overset{(*)}{=} \left[W_{j+1}(t)W_{j+1}^T(t)\right]^j \overset{(*)}{=} \cdots \overset{(*)}{=} [W_N^T(t)W_N(t)]^j$$

Put it all together and we get:

$$W_{1:j}(t)^T W_{1:j}(t) = \left[W_j^T(t)W_j(t)\right]^j = [W_N^T(t)W_N(t)]^j = \left[\left(W_{1:N}(t)^T W_{1:N}(t)\right)^{\frac{1}{N}}\right]^j$$
$$= [W_{1:N}(t)^T W_{1:N}(t)]^{\frac{j}{N}}$$

■

2. First, we'll find $\nabla\phi$

   We have seen in class that $\frac{\partial\phi(W_1,\ldots,W_N)}{\partial W_j} = W_{j+1:N}^T \nabla l(W_{1:N}) W_{1:j-1}^T$

   So in are case we get:

$$\nabla\phi(U) = U^T \nabla l(UU^T) U^T$$

$\phi(U) = l(UU^T) \Rightarrow \phi(U+\Delta) = l\big((U+\Delta)(U^T+\Delta^T)\big) = l(UU^T + U\Delta^T + \Delta U^T + \Delta\Delta^T) \overset{tylor}{=} l(UU^T) + \langle \nabla l(UU^T), \Delta U^T + U\Delta^T + \Delta\Delta^T\rangle + o(\|\Delta\|_F^2) = l(UU^T) + \langle \nabla l(UU^T), \Delta U^T + U\Delta^T\rangle + o(\|\Delta\|_F^2)$

$$\Downarrow$$

$$\langle \nabla\phi(U), \Delta\rangle = \langle \nabla l(UU^T), \Delta\rangle = \langle \nabla l(UU^T), \Delta U^T + U\Delta^T\rangle$$

And we also have:
$\langle \nabla l(UU^T), \Delta U^T + U\Delta^T\rangle = \langle \nabla l(UU^T), \Delta U^T\rangle = \langle \nabla l(UU^T), U\Delta^T\rangle = Tr(\nabla l(UU^T)^T \Delta U^T) + Tr(\nabla l(UU^T)^T U\Delta^T) = Tr(\nabla l(UU^T)^T \Delta U^T) + Tr(\nabla l(UU^T)^T U\Delta^T) = Tr(\mathrm{U}^{\mathrm{T}} \nabla l(UU^T)^T \Delta) + Tr(\mathrm{U}^{\mathrm{T}} \nabla l(UU^T)\Delta) = \langle \nabla l(UU^T), \Delta\rangle U + \langle \nabla l(UU^T)^T U, \Delta\rangle = \langle \nabla l(UU^T)U + \nabla l(UU^T)^T U, \Delta\rangle$

So over all we get:

$$\nabla\phi(U) = \nabla l(UU^T)U + \nabla l(UU^T)^T U$$

Now denote $W \coloneqq UU^T$ and find $\frac{\partial W(t)}{\partial t}$:
$\dot{W} \coloneqq \frac{\partial W(t)}{\partial t} = \frac{\partial}{\partial t}(U(t)U(t)^T) = \dot{U}(t)U(t)^T + U(t)\dot{U}(t)^T = -\nabla\phi\big(U(t)\big)U(t)^T - U(t)\nabla\phi\big(U(t)\big)^T = -(\nabla l(UU^T)U + \nabla l(UU^T)^T U)U^T - U(\nabla l(UU^T)U + \nabla l(UU^T)^T U)^T = -[\nabla l(UU^T) + \nabla l(UU^T)^T]UU^T - UU^T[\nabla l(UU^T)^T + \nabla l(UU^T)]$

■

3. Assume $d_N = 1$,
   Recall:

$$\dot{W}_{1:N} = -\sum_{j-1}^{N} [W_{1:N}(t)W_{1:N}(t)^T]^{\frac{j-1}{N}} \nabla l(W_{1:N}(t)) \, [W_{1:N}(t)^T W_{1:N}(t)]^{\frac{N-j}{N}}$$

   Note:

   (1) $j = N \Rightarrow [W_{1:N}(t)^T W_{1:N}(t)]^{\frac{N-j}{N}} = I$

   (2) $\forall j \in [1,2,\dots,N]: W_{1:N}(t)W_{1:N}(t)^T \in \mathbb{R} \Rightarrow [W_{1:N}(t)W_{1:N}(t)^T]^{\frac{j-1}{N}} = \|W_{1:N}(t)\|_2^{2\frac{(j-1)}{N}}$

   (3) $\forall j \in [1,2,\dots,N-1]: \; [W_{1:N}(t)W_{1:N}(t)^T]^{\frac{j-1}{N}} = \|W_{1:N}(t)\|_2^{2\frac{(j-1)}{N}} \left(\frac{W_{1:N}}{\|W_{1:N}\|_2}\right)^T \frac{W_{1:N}}{\|W_{1:N}\|_2}$

   Put it all together:

$$\dot{W}_{1:N} = -\sum_{j=1}^{N} [W_{1:N}(t)W_{1:N}(t)^T]^{\frac{j-1}{N}} \nabla l(W_{1:N}(t)) \, [W_{1:N}(t)^T W_{1:N}(t)]^{\frac{N-j}{N}}$$

$$= -\sum_{j=1}^{N-1} \|W_{1:N}(t)\|_2^{2\frac{j-1}{N}} \nabla l(W_{1:N}(t)\|W_{1:N}(t)\|_2^{2\frac{N-j}{N}} \left(\frac{W_{1:N}}{\|W_{1:N}\|_2}\right)^T \frac{W_{1:N}}{\|W_{1:N}\|_2}$$

$$- \|W_{1:N}(t)\|_2^{2\frac{N-j}{N}} \nabla l(W_{1:N}(t)) =$$

$$-(N-1)\|W_{1:N}(t)\|_2^{2\frac{N-1}{N}} \nabla l(W_{1:N}(t) \left(\frac{W_{1:N}}{\|W_{1:N}\|_2}\right)^T \frac{W_{1:N}}{\|W_{1:N}\|_2}$$

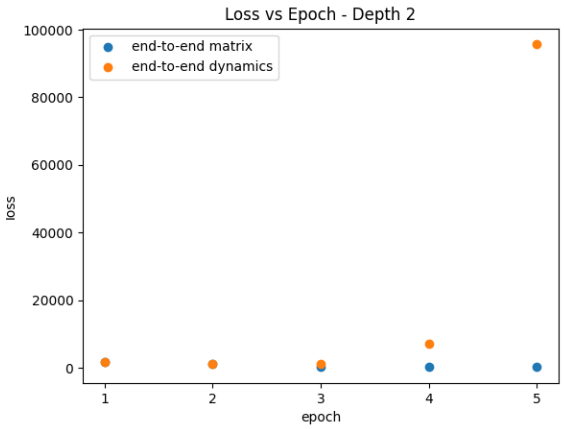
$$- \|W_{1:N}(t)\|_2^{2\frac{N-j}{N}} \nabla l(W_{1:N}(t))$$

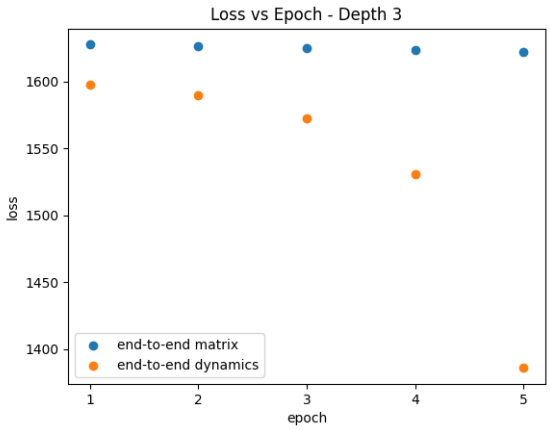
$$= -\|W_{1:N}(t)\|_2^{2\frac{N-1}{N}} \left[(N-1)\nabla l(W_{1:N}(t) \left(\frac{W_{1:N}}{\|W_{1:N}\|_2}\right)^T \frac{W_{1:N}}{\|W_{1:N}\|_2}\right.$$

$$\left. + \nabla l(W_{1:N}(t)\right]$$

   We saw in class that to get balanceness we assume that $W_{1:N}(0)$ is very close to 0, so the learning rate is increasing with the direction that is already taken.

4. For N=2 we see that for the first three epochs, the end-to-end matrix advances similarly to the end-to-end dynamics, but for epochs 4 and 5, the end-to-end dynamics misses the global minimum, and the loss increases while the end-to-end matrix keeps decreasing.
   For N=3 we see that the end-to-end dynamics outperforms the end-to-end matrix , and the trajectory leads to a much smaller loss that keeps decreasing as the epochs progress. For both experiments we ran for 5 epochs with a learning rate of 0.00001 and Xavier initialization.





Ultra wide neural networks

1.

$$\frac{d}{dt}\|u(t)-y\|^2 = 2*(u(t)-y)*\dot{u}(t) = -2H^**(u(t)-y)^2$$
$$= -2H^*\|u(t)-y\|^2$$
$$\rightarrow \|u(t)-y\|^2 = 2l\big(w(t)\big) = H^*e^{-2t}$$

Meaning that again the training loss converges to global minimum exponentially fast.

3. We can see by inspecting the kernel dynamics that the kernel preforms much better than the shallow network: