Question 1

$\mathcal{H} = \{ x \to y = W_N \sigma(W_{N-1} \cdots W_2 \sigma(W_1 x) \cdots);\ W_n \in \mathbb{R}^{d_n, d_{n-1}},\ n \in [N] \}$

Let $L(W_1, \ldots, W_N)$ be a continuously differentiable loss function that depends on $W_1, \ldots, W_N$ only through the input-output mapping of the network.

Assume that the global minimum of $L(\cdot)$ is not attained at $W_1 = 0, W_2 = 0, \ldots, W_N = 0$.

Assume also that $\sigma(\cdot)$ is continuously differentiable and $\sigma(0) = 0$. Then, $L(\cdot)$ is non-convex.

Proof

Assume by contradiction that L is convex.

Since $L(\cdot)$ is convex, it enough to show that $\nabla L(0, \ldots, 0) = 0$, because if it's true than we can yield that $(W_1 = 0, \ldots W_N = 0)$ is a global min of $L(\cdot)$ in contradiction.

Proof of: $\nabla L(\vec{0}) = 0 \Rightarrow \vec{0}$ is a global min of $L(\cdot)$

Since L is convex we know from calculus 1 that

$$\forall x \quad L(x) \geq L(\vec{0}) + \nabla L(\vec{0})^T (x - \vec{0})$$
$$= L(\vec{0}) + 0 \, (x - \vec{0})$$
$$= L(\vec{0})$$

$\Rightarrow \vec{0} = (0, \ldots, 0)$ is a global min of $L(\cdot)$

Lets assume we calculate the loss L on m training points $x_1, \dots, x_m$ with respectively vector labels $\hat{y}_1, \dots, \hat{y}_m$.

Since L depends on $w_1 \dots w_N$ only through the input-output mapping of the network we can write L as:

$$L(w_1 \dots, w_N) = L\Big(W_N \circ W_{N-1} \cdots W_1 x_1, \hat{y}_1,$$
$$W_N \circ W_{N-1} \cdots W_1 x_2, \hat{y}_2 \qquad (*)$$
$$\vdots$$
$$W_N \circ W_{N-1} \cdots W_1 x_m, \hat{y}_m \Big)$$

[L depends on $w_1 \dots w_N$ only through the network outputs on the inputs $x_1 \dots x_m$].

We would like to prove that $\nabla L(0, \dots 0) = 0$ in order to get the required. Thus, we'll prove that for every $j \in [N]$: $\frac{\partial L}{\partial w_j}(0, \dots 0) = 0$:

Notice that: from $(*)$ and the chain rule (+ definitions of h_i below), we get:

$$\frac{\partial L}{\partial w_j}(0, \dots 0) = \frac{\partial L}{\partial h_1}\left(h_1(0, \dots 0), \hat{y}_1, \dots, h_m(0, \dots 0), \hat{y}_m\right) \cdot \frac{\partial h_1}{\partial w_j}(0, \dots 0)$$
$$+ \frac{\partial L}{\partial h_2}\left(h_1(0, \dots 0), \hat{y}_1, \dots, h_m(0, \dots 0), \hat{y}_m\right) \cdot \frac{\partial h_2}{\partial w_j}(0, \dots 0)$$
$$+ \vdots$$
$$+ \frac{\partial L}{\partial h_m}\left(h_1(0, \dots 0), \hat{y}_1, \dots, h_m(0, \dots 0) \hat{y}_m\right) \cdot \frac{\partial h_m}{\partial w_j}(0, \dots 0)$$

Where $\forall 1 \le i \le m \quad h_i(w_1 \dots w_N) = W_N \circ W_{N-1} \cdots W_1 x_i$ [by definition].

Thus, in order to show that $\frac{\partial L}{\partial w_j}(0, \dots 0) = 0$, we can show that $\frac{\partial h_i}{\partial w_j}(0, \dots 0) = 0 \quad \forall 1 \le i \le m$ and to finish.

As desired $h_i(W_1 \dots W_N) = W_N \sigma W_{N-1} \cdots W_1 X_i$

$= W_N \sigma \cdots W_j \sigma \cdots W_1 X_i$

We'll define: $f_i(W_1 \dots W_j) = W_j \sigma \cdots W_1 X_i$

$g(V, W_{j+1} \dots W_N) = W_N \sigma W_{N-1} \cdots W_{j+1} \sigma(V)$

Therefore by these definitions we yield

$h_i(W_1 \dots W_N) = g(f_i(W_1 \dots W_j), W_{j+1} \dots W_N)$

$$\Rightarrow \frac{\partial h_i}{\partial W_j}(0, \dots 0) = \frac{\partial g}{\partial f_i}(0 \dots 0) \cdot \frac{\partial f_i}{\partial W_j}(0, \dots 0)$$

Thus we can show that $\frac{\partial f_i}{\partial W_j}(0, \dots 0) = 0$ and finish.

$f_i(W_1 \dots W_j) = W_j \sigma \underbrace{W_{j-1} \cdots W_1 X_i}_{(**)}$

$\Rightarrow$ When we calculate the derivative of f_i by W_j we can treat $(**)$ as a const vector $y = \sigma W_{j-1} \cdots W_1 X_i$:

$$\frac{\partial f_i}{\partial W_j}(W_1 \dots W_j) = \frac{\partial (W_j y)}{\partial W_j}(W_1 \dots W_j), \quad W_j \in \mathbb{R}^{d_j, d_{j-1}}, \; y \in \mathbb{R}^{d_{j-1}} = \begin{pmatrix} y_1 \\ \vdots \\ y_{d_{j-1}} \end{pmatrix}$$

Let's sign $d_j = m$, $d_{j-1} = k$ for simplicity. Then:

$$W_j y = \begin{pmatrix} (W_j)_{11} y_1 + \dots + (W_j)_{1k} y_k \\ \vdots \\ (W_j)_{m1} y_1 + \dots + (W_j)_{mk} y_k \end{pmatrix}$$

Thus $\forall \; 1 \le i \le m, \; 1 \le \tilde{j} \le k$:

$$\frac{\partial W_j y}{\partial (W_j)_{i\tilde{j}}} = \begin{pmatrix} 0 \\ \vdots \\ y_{\tilde{j}} \\ \vdots \\ 0 \end{pmatrix} \leftarrow i\text{-th row}$$

As we proved $\forall\ 1 \le i \le m \quad 1 \le \tilde{j} \le k$:

$$\frac{\partial(W_j y)}{\partial (W_j)_{i\tilde{j}}} = \begin{pmatrix} 0 \\ \vdots \\ y_{\tilde{j}} \\ \vdots \\ 0 \end{pmatrix} \leftarrow i$$

$$\Rightarrow \frac{\partial(W_j y)}{\partial (W_j)_{i\tilde{j}}}(0,\dots 0) = \begin{pmatrix} 0 \\ \vdots \\ (y(0\dots 0))_{\tilde{j}} \\ \vdots \\ 0 \end{pmatrix} \leftarrow i$$

Remember $y = \sigma W_{j-1} \cdots W_1 X_i$

$$\Rightarrow y(0,\dots 0) = \sigma 0 \cdots x_1 = \sigma(0) = 0$$

Therefore we've got $\frac{\partial(W_j y)}{\partial (W_j)_{i\tilde{j}}}(0,\dots 0) = 0$

$$\Rightarrow \frac{\partial(W_j y)}{\partial W_j}(0,\dots 0) = 0$$

$$\Rightarrow \frac{\partial f_i}{\partial W_j}(0,\dots 0) = \frac{\partial(W_j y)}{\partial W_j}(0,\dots 0) = 0$$

as required.

□

Foundations of Deep Learning – Homework Assignment #3

Adi Album & Tomer Epshtein

Part 2: (1)

Question:

In this question we extend the convergence of stationary point result delivered in class from gradient descent to stochastic gradient descent.

Let $f: \mathbb{R}^d \to \mathbb{R}$ be a twice continuously differentiable and β-smooth function that attains its global minimum $f^* = \min_{\vec{w} \in \mathbb{R}^d} f(\vec{w})$. Suppose we run stochastic gradient descent over $f(\cdot)$:

$$\overrightarrow{w_{t+1}} \leftarrow \overrightarrow{w_t} - \eta_t(\nabla f(\overrightarrow{w_t}) + \xi_t), t = 1,2,3,\dots$$

Where η_t is a (possibly time varying) learning rate, and ξ_t stands for a time-independent noise with zero mean ($\mathbb{E}[\xi_t] = 0$) and σ^2 variance ($\mathbb{E}[\|\xi_t\|^2] = \sigma^2$). Let $\epsilon > 0$.

Assume $\eta_t = \frac{1}{\beta}$, and derive an upper bound on the number of iterations needed for reaching an expected $(\epsilon + \sigma)$-stationary point, i.e. on $T \in \mathbb{N}$ that will ensure:

$$\min_{t \in [T]} \mathbb{E}[\|\nabla f(\overrightarrow{w_t})\|] \leq \epsilon + \sigma$$

Proof:

Let $f: \mathbb{R}^d \to \mathbb{R}$ be a twice continuously differentiable and β-smooth function that attains its global minimum $f^* = \min_{\vec{w} \in \mathbb{R}^d} f(\vec{w})$.

Reminder: In class we proved the following lemma:

Lemma:

Let $f: \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable and β-smooth. Then:

$$\forall \overrightarrow{w_1}, \overrightarrow{w_2} \in \mathbb{R}^d: \left| f(\overrightarrow{w_2}) - \underbrace{(f(\overrightarrow{w_1}) + < \nabla f(\overrightarrow{w_1}), \overrightarrow{w_2} - \overrightarrow{w_1} >)}_{1^{st}\ order\ approx\ around\ \overrightarrow{w_1}} \right| \leq \frac{\beta}{2} \|\overrightarrow{w_1} - \overrightarrow{w_2}\|^2$$

So, for any $t \geq 0$:

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

We'll take expected value over $\xi_1, \xi_2, \dots, \xi_t$:

Denote- $E_{t+1} := \mathbb{E}_{\xi_1, \xi_2, \dots, \xi_t}[f(\overrightarrow{w_{t+1}})]$, so

$$E_{t+1} = \mathbb{E}_{\xi_1, \xi_2, \dots, \xi_t}[f(\overrightarrow{w_{t+1}})] \leq \mathbb{E}_{\xi_1, \xi_2, \dots, \xi_t}\left[f(\overrightarrow{w_t}) + < \nabla f(\overrightarrow{w_t}), \overrightarrow{w_{t+1}} - \overrightarrow{w_t} > + \frac{\beta}{2} \|w_{t+1} - w_t\|^2\right] =$$

$$= \mathbb{E}_{\xi_1, \xi_2, \dots, \xi_t}[f(\overrightarrow{w_t})] + \mathbb{E}_{\xi_1, \xi_2, \dots, \xi_t}[< \nabla f(\overrightarrow{w_t}), \overrightarrow{w_{t+1}} - \overrightarrow{w_t} >] + \frac{\beta}{2} \mathbb{E}_{\xi_1, \xi_2, \dots, \xi_t}[\|w_{t+1} - w_t\|^2] =$$

w_t is independent of ξ_t:

$$= \mathbb{E}_{\xi_1, \xi_2, \dots, \xi_{t-1}}[f(\overrightarrow{w_t})] + \mathbb{E}_{\xi_1, \xi_2, \dots, \xi_t}[< \nabla f(\overrightarrow{w_t}), \overrightarrow{w_{t+1}} - \overrightarrow{w_t} >] + \frac{\beta}{2} \mathbb{E}_{\xi_1, \xi_2, \dots, \xi_t}[\|w_{t+1} - w_t\|^2] =$$

$$= E_t + \mathbb{E}_{\xi_1, \xi_2, \dots, \xi_t}[< \nabla f(\overrightarrow{w_t}), \overrightarrow{w_{t+1}} - \overrightarrow{w_t} >] + \frac{\beta}{2} \mathbb{E}_{\xi_1, \xi_2, \dots, \xi_t}[\|w_{t+1} - w_t\|^2] \ (*) \quad =$$

Plugging in the SGD step definition

$$\overrightarrow{w_{t+1}} \leftarrow \overrightarrow{w_t} - \eta_t(\nabla f(\overrightarrow{w_t}) + \xi_t), t = 1,2,3,\dots$$

We achieve:

$$E_{t+1} \le E_t + \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}[< \nabla f(\overrightarrow{w_t}), \overrightarrow{w_{t+1}} - \overrightarrow{w_t} >] + \frac{\beta}{2}\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}[\|w_{t+1} - w_t\|^2] =$$

$$= E_t + \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}[< \nabla f(\overrightarrow{w_t}), -\eta_t(\nabla f(\overrightarrow{w_t}) + \xi_t) >] + \frac{\beta}{2}\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}[\|-\eta_t(\nabla f(\overrightarrow{w_t}) + \xi_t)\|^2] \overset{\eta_t \equiv const}{=}$$

$$= E_t - \eta_t \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}[\|\nabla f(\overrightarrow{w_t})\|^2] - \eta_t \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}[< \nabla f(\overrightarrow{w_t}), \xi_t >] + \frac{\beta}{2}\eta_t^2 \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}[\|\nabla f(\overrightarrow{w_t}) + \xi_t\|^2] =$$

$$= E_t - \eta_t \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_{t-1}}[\|\nabla f(\overrightarrow{w_t})\|^2] - \eta_t \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}[< \nabla f(\overrightarrow{w_t}), \xi_t >] + \frac{\beta}{2}\eta_t^2 \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}[\|\nabla f(\overrightarrow{w_t}) + \xi_t\|^2] =$$

$\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}[< \nabla f(\overrightarrow{w_t}), \xi_t >] = 0$, because $\mathbb{E}[\xi_t] = 0$, and linearity of expected value with inner product.

$$= E_t - \eta_t \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_{t-1}}[\|\nabla f(\overrightarrow{w_t})\|^2] + \frac{\beta}{2}\eta_t^2 \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}[\|\nabla f(\overrightarrow{w_t}) + \xi_t\|^2]$$

Now, let's focus on $\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}[\|\nabla f(\overrightarrow{w_t}) + \xi_t\|^2]$:

$$\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}[\|\nabla f(\overrightarrow{w_t}) + \xi_t\|^2] = \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}[\|\nabla f(\overrightarrow{w_t})\|^2 + \|\xi_t\|^2 + 2 < \nabla f(\overrightarrow{w_t}), \xi_t >] =$$

$$= \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}[\|\nabla f(\overrightarrow{w_t})\|^2] + \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}[\|\xi_t\|^2] + 2\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}[< \nabla f(\overrightarrow{w_t}), \xi_t >] =$$

Here, once again, $\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}[< \nabla f(\overrightarrow{w_t}), \xi_t >] = 0$. And $\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}[\|\xi_t\|^2] = \mathbb{E}_{\xi_t}[\|\xi_t\|^2] = \sigma^2$ by ξ_t's distribution's definition.

$$= \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_{t-1}}[\|\nabla f(\overrightarrow{w_t})\|^2] + \sigma^2$$

So we have:

$$\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}[\|\nabla f(\overrightarrow{w_t}) + \xi_t\|^2] = \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_{t-1}}[\|\nabla f(\overrightarrow{w_t})\|^2] + \sigma^2$$

Plugging in the result:

$$E_{t+1} \le E_t - \eta_t \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_{t-1}}[\|\nabla f(\overrightarrow{w_t})\|^2] + \frac{\beta}{2}\eta_t^2 \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}[\|\nabla f(\overrightarrow{w_t}) + \xi_t\|^2] =$$

$$= E_t - \eta_t \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_{t-1}}[\|\nabla f(\overrightarrow{w_t})\|^2] + \frac{\beta}{2}\eta_t^2 \left(\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_{t-1}}[\|\nabla f(\overrightarrow{w_t})\|^2] + \sigma^2\right)$$

Plugging in $\eta_t = \frac{1}{\beta}$:

$$E_{t+1} \le E_t - \frac{1}{\beta}\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_{t-1}}[\|\nabla f(\overrightarrow{w_t})\|^2] + \frac{1}{2\beta}\left(\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_{t-1}}[\|\nabla f(\overrightarrow{w_t})\|^2] + \sigma^2\right) = E_t - \frac{1}{2\beta}\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_{t-1}}[\|\nabla f(\overrightarrow{w_t})\|^2] + \frac{\sigma^2}{2\beta}$$

Assume that for steps $t = 0,1,\dots,T-1$ we didn't achieve an expected $(\epsilon + \sigma)$-stationary point.
I.e. $\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_{t-1}}[\|\nabla f(\overrightarrow{w_t})\|] > \epsilon + \sigma$ for all $t = 0,1,\dots,T-1$.

For any random variable X: $\mathbb{E}[X^2] = \mathbb{E}[X]^2 + \underset{\ge 0}{Var(X)} \ge \mathbb{E}[X]^2$.
So, for steps $t = 0,1,\dots,T-1$:

$$\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_{t-1}}[\|\nabla f(\overrightarrow{w_t})\|^2] \ge \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_{t-1}}[\|\nabla f(\overrightarrow{w_t})\|]^2 > (\epsilon + \sigma)^2$$

So:

$$E_{t+1} \le E_t - \frac{1}{2\beta}\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_{t-1}}[\|\nabla f(\overrightarrow{w_t})\|^2] + \frac{\sigma^2}{2\beta} < E_t - \frac{1}{2\beta}(\epsilon + \sigma)^2 + \frac{\sigma^2}{2\beta} = E_t - \frac{\epsilon}{\beta}\left(\frac{\epsilon}{2} + \sigma\right)$$

This inequality holds for all $t = 0,1,\dots,T-1$. So:

$$E_T < E_0 - T\frac{\epsilon}{\beta}\left(\frac{\epsilon}{2} + \sigma\right)$$

Since f^* is f's global minimum, $f^* \le E_T \left(= \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_{T-1}}[f(\overrightarrow{w_T})]\right)$

So we have

$$f^* < E_0 - T\frac{\epsilon}{\beta}\left(\frac{\epsilon}{2} + \sigma\right)$$

Where $E_0 = f(\overrightarrow{w_0})$, so:

$$f^* < f(\overrightarrow{w_0}) - T\frac{\epsilon}{\beta}\left(\frac{\epsilon}{2} + \sigma\right) \Rightarrow T < \frac{2\beta(f(\overrightarrow{w_0}) - f^*)}{\epsilon(\epsilon + 2\sigma)}$$

So, for any $\epsilon > 0$ an expected $(\epsilon + \sigma)$-stationary point will be reached within a number of steps at most

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

Foundations of Deep Learning – Homework Assignment #3

Adi Album & Tomer Epshtein

Part 2: (2)

Question: (Bonus)

Show that by decaying learning rate it is possible to guarantee convergence to expected ϵ-stationary point. In particular, define a scheme for η_t and a respective upper bound on $T \in \mathbb{N}$ that will ensure:

$$\min_{t\in[T]} \mathbb{E}[\|\nabla f(\overrightarrow{w_t})\|] \leq \epsilon$$

Proof:

Let $f: \mathbb{R}^d \to \mathbb{R}$ be a twice continuously differentiable and β-smooth function that attains its global minimum $f^* = \min_{\vec{w}\in\mathbb{R}^d} f(\vec{w})$.

We define an update scheme for $\eta_t := \frac{\|\nabla f(\overrightarrow{w_t})\|^2}{\beta(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)}$

In part(2) 1 we derived

$$E_{t+1} \leq E_t + \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}[< \nabla f(\overrightarrow{w_t}), \overrightarrow{w_{t+1}} - \overrightarrow{w_t} >] + \frac{\beta}{2}\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}[\|w_{t+1} - w_t\|^2] \qquad (*)$$

Plugging in SGD's update scheme $\overrightarrow{w_{t+1}} \leftarrow \overrightarrow{w_t} - \eta_t(\nabla f(\overrightarrow{w_t}) + \xi_t)$ where $\eta_t = \frac{\|\nabla f(\overrightarrow{w_t})\|^2}{\beta(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)}$, we achieve:

$$E_{t+1} \leq E_t + \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}\left[< \nabla f(\overrightarrow{w_t}), \overrightarrow{w_t} - \frac{\|\nabla f(\overrightarrow{w_t})\|^2}{\beta(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)}(\nabla f(\overrightarrow{w_t}) + \xi_t) - \overrightarrow{w_t} >\right]$$
$$+\frac{\beta}{2}\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}\left[\left\|\overrightarrow{w_t} - \frac{\|\nabla f(\overrightarrow{w_t})\|^2}{\beta(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)}(\nabla f(\overrightarrow{w_t}) + \xi_t) - \overrightarrow{w_t}\right\|^2\right] =$$

$$= E_t + \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}\left[< \nabla f(\overrightarrow{w_t}), -\frac{\|\nabla f(\overrightarrow{w_t})\|^2}{\beta(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)}(\nabla f(\overrightarrow{w_t}) + \xi_t) >\right]$$
$$+\frac{\beta}{2}\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}[\left\|-\frac{\|\nabla f(\overrightarrow{w_t})\|^2}{\beta(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)}(\nabla f(\overrightarrow{w_t}) + \xi_t)\right\|^2] =$$

$$= E_t + \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}\left[-\frac{\|\nabla f(\overrightarrow{w_t})\|^2}{\beta(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)} < \nabla f(\overrightarrow{w_t}), \nabla f(\overrightarrow{w_t}) + \xi_t >\right]$$
$$+\frac{\beta}{2}\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}\left[\frac{\|\nabla f(\overrightarrow{w_t})\|^4}{\beta^2(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)^2}\|\nabla f(\overrightarrow{w_t}) + \xi_t\|^2\right] =$$

$$= E_t + \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}\left[-\frac{\|\nabla f(\overrightarrow{w_t})\|^4}{\beta(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)} - \frac{\|\nabla f(\overrightarrow{w_t})\|^2}{\beta(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)} < \nabla f(\overrightarrow{w_t}), \xi_t >\right]$$
$$+\frac{\beta}{2}\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}\left[\frac{\|\nabla f(\overrightarrow{w_t})\|^4}{\beta^2(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)^2}(\|\nabla f(\overrightarrow{w_t})\|^2 + 2 < \nabla f(\overrightarrow{w_t}), \xi_t > +\|\xi_t\|^2)\right] =$$

$$= E_t \underbrace{- \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}\left[\frac{\|\nabla f(\overrightarrow{w_t})\|^4}{\beta(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)}\right]}_{(I)} \underbrace{- \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}\left[\frac{\|\nabla f(\overrightarrow{w_t})\|^2}{\beta(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)} < \nabla f(\overrightarrow{w_t}), \xi_t >\right]}_{(II)}$$
$$\underbrace{+\frac{\beta}{2}\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}\left[\frac{\|\nabla f(\overrightarrow{w_t})\|^6}{\beta^2(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)^2}\right]}_{(III)} \underbrace{+\beta\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}\left[\frac{\|\nabla f(\overrightarrow{w_t})\|^4}{\beta^2(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)^2} < \nabla f(\overrightarrow{w_t}), \xi_t >\right]}_{(IV)}$$
$$\underbrace{+\frac{\beta}{2}\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}\left[\frac{\|\nabla f(\overrightarrow{w_t})\|^4}{\beta^2(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)^2}\|\xi_t\|^2\right]}_{(V)}$$

We'll analyze each $(I), (II), (III), (IV), (V)$ separately.

- $(I) = \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}\left[\frac{\|\nabla f(\overrightarrow{w_t})\|^4}{\beta(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)}\right] = \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_{t-1}}\left[\frac{\|\nabla f(\overrightarrow{w_t})\|^4}{\beta(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)}\right]$
  Because $\overrightarrow{w_t}$ is independent of ξ_t
- $(II) = \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}\left[\frac{\|\nabla f(\overrightarrow{w_t})\|^2}{\beta(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)} < \nabla f(\overrightarrow{w_t}), \xi_t >\right] = \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}\left[< \frac{\|\nabla f(\overrightarrow{w_t})\|^2}{\beta(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)} \nabla f(\overrightarrow{w_t}), \xi_t >\right] = 0$
  Because $\overrightarrow{w_t}$ is independent of ξ_t and $\mathbb{E}[\xi_t] = 0$, and from linearity of expected value with inner product, LHS is independent of ξ_t.
- $(III) = \frac{\beta}{2}\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}\left[\frac{\|\nabla f(\overrightarrow{w_t})\|^6}{\beta^2(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)^2}\right] = \frac{\beta}{2}\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_{t-1}}\left[\frac{\|\nabla f(\overrightarrow{w_t})\|^6}{\beta^2(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)^2}\right]$
  Because $\overrightarrow{w_t}$ is independent of ξ_t
- $(IV) = \beta\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}\left[\frac{\|\nabla f(\overrightarrow{w_t})\|^4}{\beta^2(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)^2} < \nabla f(\overrightarrow{w_t}), \xi_t >\right] = \beta\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}\left[< \frac{\|\nabla f(\overrightarrow{w_t})\|^4}{\beta^2(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)^2} \nabla f(\overrightarrow{w_t}), \xi_t >\right] = 0$
  Because $\overrightarrow{w_t}$ is independent of ξ_t and $\mathbb{E}[\xi_t] = 0$, and from linearity of expected value with inner product, LHS is independent of ξ_t.
- $(V) = \frac{\beta}{2}\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_t}\left[\frac{\|\nabla f(\overrightarrow{w_t})\|^4}{\beta^2(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)^2}\|\xi_t\|^2\right] = \frac{\beta}{2}\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_{t-1}}\left[\frac{\|\nabla f(\overrightarrow{w_t})\|^4}{\beta^2(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)^2}\right]\sigma^2$
  Because $\overrightarrow{w_t}$ is independent of ξ_t and $\mathbb{E}[\|\xi_t\|^2] = \sigma^2$

Bringing it all together we achieve

$$E_{t+1} \leq$$

$$E_t - \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_{t-1}}\left[\frac{\|\nabla f(\overrightarrow{w_t})\|^4}{\beta(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)}\right] + \frac{\beta}{2}\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_{t-1}}\left[\frac{\|\nabla f(\overrightarrow{w_t})\|^6}{\beta^2(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)^2}\right] + \frac{\beta}{2}\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_{t-1}}\left[\frac{\|\nabla f(\overrightarrow{w_t})\|^4}{\beta^2(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)^2}\right]\sigma^2$$

$$= E_t + \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_{t-1}}\left[-\frac{\|\nabla f(\overrightarrow{w_t})\|^4}{\beta(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)} + \frac{\|\nabla f(\overrightarrow{w_t})\|^6}{2\beta(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)^2} + \frac{\|\nabla f(\overrightarrow{w_t})\|^4}{2\beta(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)^2}\sigma^2\right] =$$

$$= E_t + \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_{t-1}}\left[-\frac{\|\nabla f(\overrightarrow{w_t})\|^4}{\beta(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)} + \frac{\|\nabla f(\overrightarrow{w_t})\|^4}{2\beta(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)^2}\|\nabla f(\overrightarrow{w_t})\|^2 + \frac{\|\nabla f(\overrightarrow{w_t})\|^4}{2\beta(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)^2}\sigma^2\right] =$$

$$= E_t + \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_{t-1}}\left[-\frac{\|\nabla f(\overrightarrow{w_t})\|^4}{\beta(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)} + \frac{\|\nabla f(\overrightarrow{w_t})\|^4}{2\beta(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)^2}(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)\right] =$$

$$= E_t + \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_{t-1}}\left[-\frac{\|\nabla f(\overrightarrow{w_t})\|^4}{\beta(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)} + \frac{\|\nabla f(\overrightarrow{w_t})\|^4}{2\beta(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)}\right] =$$

$$= E_t - \frac{1}{2\beta}\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_{t-1}}\left[\frac{\|\nabla \mathrm{f}(\overrightarrow{\mathrm{w_t}})\|^4}{\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2}\right]$$

So

$$E_{t+1} \leq E_t - \frac{1}{2\beta}\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_{t-1}}\left[\frac{\|\nabla \mathrm{f}(\overrightarrow{\mathrm{w_t}})\|^4}{\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2}\right]$$

Assume that for steps $t = 0,1,\dots T-1$ we didn't achieve an expected ϵ-stationary point.
I.e. $\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_{t-1}}[\|\nabla f(\overrightarrow{w_t})\|] > \epsilon$ for all $t = 0,1,\dots,T-1$.

Define $g\colon [0,\infty) \to \mathbb{R}$ by $g(x) = \frac{x^4}{x^2+\sigma^2}$. g is convex monotonically increasing.
So, for all $x > \epsilon$, $g(x) > \frac{\epsilon^4}{\epsilon^2+\sigma^2}$.

g is convex, so by Jensen's inequality:

$$\mathbb{E}[X] > \epsilon \Longrightarrow \mathbb{E}[g(X)] > g(\epsilon)$$

I.e.

$$\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_{t-1}}\left[\frac{\|\nabla \mathrm{f}(\overrightarrow{\mathrm{w_t}})\|^4}{\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2}\right] = \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_{t-1}}[g(\|\nabla f(\overrightarrow{w_t})\|)] > g(\epsilon) = \frac{\epsilon^4}{\epsilon^2+\sigma^2}$$

So, for steps $t = 0,1,\dots,T-1$:

$$\mathbb{E}_{\xi_1,\xi_2,\dots,\xi_{t-1}}\left[\frac{\|\nabla \mathrm{f}(\overrightarrow{\mathrm{w_t}})\|^4}{\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2}\right] > \frac{\epsilon^4}{\epsilon^2+\sigma^2}$$

So:

$$E_{t+1} < E_t - \frac{1}{2\beta}\frac{\epsilon^4}{\epsilon^2+\sigma^2}$$

This inequality holds for all $t = 0,1,\dots,T-1$:

$$E_T < E_0 - T\frac{1}{2\beta}\frac{\epsilon^4}{\epsilon^2+\sigma^2}$$

Since f^* is f's global minimum, $f^* \le E_T\left(= \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_{T-1}}[f(\overrightarrow{w_T})]\right)$
So,

$$f^* < E_0 - T\frac{1}{2\beta}\frac{\epsilon^4}{\epsilon^2+\sigma^2}$$

Where $E_0 = f(\overrightarrow{w_0})$, so:

$$f^* < f(\overrightarrow{w_0}) - T\frac{\epsilon^4}{2\beta(\epsilon^2+\sigma^2)}$$

$\Rightarrow$

$$T < \frac{2\beta(\epsilon^2+\sigma^2)(f(\overrightarrow{w_0}) - f^*)}{\epsilon^4}$$

Side note:

Where did the update rule for η_t, $\eta_t = \frac{\|\nabla f(\overrightarrow{w_t})\|^2}{\beta(\|\nabla f(\overrightarrow{w_t})\|^2+\sigma^2)}$, come from?

Denote $g_t := \mathbb{E}_{\xi_1,\xi_2,\dots,\xi_{t-1}}[\|\nabla f(\overrightarrow{w_t})\|^2]$, In part 2(1) we showed:

$$E_{t+1} \le E_t + \left(-\eta_t g_t + \frac{\beta}{2}\eta_t^2(g_t+\sigma^2)\right)$$

We would like to select an update scheme for η_t such that the above bound $(-\eta_t g_t + \frac{\beta}{2}\eta_t^2(g_t+\sigma^2))$ is minimal.

Let $t \in \mathbb{N}, d$efine $h_t: \mathbb{R} \to \mathbb{R}, h_t(\eta_t) := \frac{\beta}{2}(g_t+\sigma^2)\eta_t^2 - g_t\eta_t$.
h_t is a parabola that attains its minimum at $\eta_t = \frac{g_t}{\beta(g_t+\sigma^2)}$
(This can easily be seen noticing h_t is a parabola of the form $h_t(x) = ax^2 + bx$ with $a > 0$ so its minimum is achieved at $x = -\frac{b}{2a}$)

Selecting this update scheme gives us the desired bound

Foundations of Deep Learning – Homework Assignment #3

Adi Album & Tomer Epshtein

Part 2: (3)

Question:

Let $\ell: \mathbb{R}^{d,d} \to \mathbb{R}$ be a twice continuously differentiable convex loss, overparameterized by a depth N linear neural network with hidden widths d:

$$\phi: \mathbb{R}^{d,d} \times \mathbb{R}^{d,d} \times \cdots \times \mathbb{R}^{d,d} \to \mathbb{R} \quad , \quad \phi(W_1, W_2, \dots, W_N) = \ell(W_N W_{N-1} \cdots W_1)$$

Prove the following claims:

A. If we restrict the domain of $\phi(\cdot)$ to $B^{d,d} \times B^{d,d} \times \cdots \times B^{d,d}$, where $B^{d,d} \coloneqq \{x \in \mathbb{R}^{d,d} : \|x\|_F \leq 1\}$, then we obtain a smooth (Lipschitz gradient) function.
B. If we do not restrict the domain of $\phi(\cdot)$, the function is smooth if and only if at least one of the following conditions hold:
    a. $\ell(\cdot)$ is constant
    b. $\ell(\cdot)$ is affine and the depth is $N = 2$.

Proof (A):

Let $(W_1, W_2, \dots, W_N), (\widetilde{W_1}, \widetilde{W_2}, \dots, \widetilde{W_N}) \in B^{d,d} \times B^{d,d} \times \cdots \times B^{d,d}$.

Denote by $W_{j:j'} = W_{j'} \cdots W_j$ for $1 \leq j \leq j' \leq N$ and Id matrix otherwise.

Let's take a look at the expression:

$$\nabla\phi(\widetilde{W_1}, \widetilde{W_2}, \dots, \widetilde{W_N}) - \nabla\phi(W_1, W_2, \dots, W_N) \in \underbrace{\mathbb{R}^{d,d} \times \mathbb{R}^{d,d} \times \cdots \times \mathbb{R}^{d,d}}_{N\ times}$$

We saw in class:

For $i \in [N]$, let's look at $\left(\nabla\phi(W_1, W_2, \dots, W_N)\right)_i \in \mathbb{R}^{d,d}$

$$\left(\nabla\phi(W_1, W_2, \dots, W_N)\right)_i = \frac{\partial}{\partial W_i}\phi(W_1, W_2, \dots, W_N) = W_{i+1:N}^T \nabla l(W_{1:N}) W_{1:i-1}^T$$

Yielding

$$\begin{aligned}
&\left(\nabla\phi(\widetilde{W_1}, \widetilde{W_2}, \dots, \widetilde{W_N}) - \nabla\phi(W_1, W_2, \dots, W_N)\right)_i = \\
&= \frac{\partial}{\partial \widetilde{W_i}}\phi(\widetilde{W_1}, \widetilde{W_2}, \dots, \widetilde{W_N}) - \frac{\partial}{\partial W_i}\phi(W_1, W_2, \dots, W_N) = \\
&= \widetilde{W}_{i+1:N}^T \nabla l(\widetilde{W}_{1:N}) \widetilde{W}_{1:i-1}^T - W_{i+1:N}^T \nabla l(W_{1:N}) W_{1:i-1}^T
\end{aligned}$$

We'll add and subtract terms…

$$= \underbrace{\widetilde{W}_{i+1:N}^T \nabla l(\widetilde{W}_{1:N}) \widetilde{W}_{1:i-1}^T - W_{i+1:N}^T \nabla l(\widetilde{W}_{1:N}) \widetilde{W}_{1:i-1}^T}_{(A)} +$$

$$+ \underbrace{W_{i+1:N}^T \nabla l(\widetilde{W}_{1:N}) \widetilde{W}_{1:i-1}^T - W_{i+1:N}^T \nabla l(W_{1:N}) \widetilde{W}_{1:i-1}^T}_{(B)} +$$

$$+ \underbrace{W_{i+1:N}^T \nabla l(W_{1:N}) \widetilde{W}_{1:i-1}^T - W_{i+1:N}^T \nabla l(W_{1:N}) W_{1:i-1}^T}_{(C)} =$$

$$= \underbrace{(\widetilde{W}_{i+1:N}^T - W_{i+1:N}^T) \nabla l(\widetilde{W}_{1:N}) \widetilde{W}_{1:i-1}^T}_{(A)} + \underbrace{W_{i+1:N}^T \left(\nabla l(\widetilde{W}_{1:N}) - \nabla l(W_{1:N})\right) \widetilde{W}_{1:i-1}^T}_{(B)} + \underbrace{W_{i+1:N}^T \nabla l(W_{1:N}) (\widetilde{W}_{1:i-1}^T - W_{1:i-1}^T)}_{(C)}$$

So:

$$\left\|\nabla\phi(\widetilde{W_1},\widetilde{W_2},\dots,\widetilde{W_N})-\nabla\phi(W_1,W_2,\dots,W_N)\right\|_2^2=\sum_{i=1}^{N}\left\|\left(\nabla\phi(\widetilde{W_1},\widetilde{W_2},\dots,\widetilde{W_N})-\nabla\phi(W_1,W_2,\dots,W_N)\right)_i\right\|_F^2\leq$$

Denote $j\coloneqq argmax_{i\in[N]}\left(\left\|\left(\nabla\phi(\widetilde{W_1},\widetilde{W_2},\dots,\widetilde{W_N})-\nabla\phi(W_1,W_2,\dots,W_N)\right)_i\right\|_F^2\right)$, so

$$\leq\sum_{i=1}^{N}\left\|\left(\nabla\phi(\widetilde{W_1},\widetilde{W_2},\dots,\widetilde{W_N})-\nabla\phi(W_1,W_2,\dots,W_N)\right)_j\right\|_F^2=N\cdot\left\|\left(\nabla\phi(\widetilde{W_1},\widetilde{W_2},\dots,\widetilde{W_N})-\nabla\phi(W_1,W_2,\dots,W_N)\right)_j\right\|_F^2=$$

$$=N\cdot\left\|\underbrace{(\widetilde{W}_{j+1:N}^T-W_{j+1:N}^T)\nabla l(\widetilde{W}_{1:N})\widetilde{W}_{1:j-1}^T}_{(A)}+\underbrace{W_{j+1:N}^T\left(\nabla l(\widetilde{W}_{1:N})-\nabla l(W_{1:N})\right)\widetilde{W}_{1:j-1}^T}_{(B)}+\underbrace{W_{j+1:N}^T\nabla l(W_{1:N})(\widetilde{W}_{1:j-1}^T-W_{1:j-1}^T)}_{(C)}\right\|_F^2\overset{\Delta}{\leq}$$

$$N\cdot\left(\underbrace{\left\|(\widetilde{W}_{j+1:N}^T-W_{j+1:N}^T)\nabla l(\widetilde{W}_{1:N})\widetilde{W}_{1:j-1}^T\right\|_F^2}_{(A)}+\underbrace{\left\|W_{j+1:N}^T\left(\nabla l(\widetilde{W}_{1:N})-\nabla l(W_{1:N})\right)\widetilde{W}_{1:j-1}^T\right\|_F^2}_{(B)}\right.$$

$$\left.+\underbrace{\left\|W_{j+1:N}^T\nabla l(W_{1:N})(\widetilde{W}_{1:j-1}^T-W_{1:j-1}^T)\right\|_F^2}_{(C)}\right)$$

We will bound each $(A),(B)$ & (C) separately.

<u>Propositions:</u>

1. $\forall k\in\mathbb{N}$, and $A_1,\dots,A_k,\tilde{A}_1,\dots,\tilde{A}_k\in B^{d,d}$:

   $$\left\|\tilde{A}_{1:k}-A_{1:k}\right\|_F^2\leq\sum_{i=1}^{k}\left\|\tilde{A}_i-A_i\right\|_F^2$$

   Notice that if we denote $(A_1,\dots,A_k),(\tilde{A}_1,\dots,\tilde{A}_k)\in B^{d,d}\times B^{d,d}\times\cdots\times B^{d,d}$ we obtain:

   $$\left\|\tilde{A}_{1:k}-A_{1:k}\right\|_F^2\leq\left\|(\tilde{A}_1,\dots,\tilde{A}_k)-(A_1,\dots,A_k)\right\|_2^2$$

2. $\forall k\in\mathbb{N}$, and $A_1,\dots,A_k\in B^{d,d}$:

   $$A_1\cdots A_k\in B^{d,d}$$

We will use these propositions to complete our proof. We'll prove them later.

<u>Note:</u>

i. $\ell\colon\mathbb{R}^{d,d}\to\mathbb{R}$ is a twice continuously differentiable convex function, $B^{d,d}$ is compact and convex and therefore $\ell|_{B^{d,d}}\colon B^{d,d}\to\mathbb{R}$ is Lipschitz gradient. I.e there exists $\beta>0$, such that $\forall A,B\in B^{d,d}$:

   $$\|\nabla\ell(B)-\nabla\ell(A)\|_F\leq\beta\|B-A\|_F$$

   (This is true by claim in course's recap notes on convexity and Lipschitzness)

ii. $\ell\colon\mathbb{R}^{d,d}\to\mathbb{R}$ is a twice continuously differentiable function, $B^{d,d}$ is compact, therefor $\nabla\ell|_{B^{d,d}}\colon B^{d,d}\to\mathbb{R}$ is bounded on $B^{d,d}$. I.e. there exists $M>0$ such that $\forall A\in B^{d,d}$:

   $$\|\nabla\ell(A)\|_F\leq M$$

Bound on (A):

$$(A) = \left\|(\widetilde{W}_{j+1:N}^T - W_{j+1:N}^T)\nabla l(\widetilde{W}_{1:N})\widetilde{W}_{1:j-1}^T\right\|_F^2 \leq \left\|\widetilde{W}_{j+1:N}^T - W_{j+1:N}^T\right\|_F^2 \left\|\nabla l(\widetilde{W}_{1:N})\right\|_F^2 \left\|\widetilde{W}_{1:j-1}^T\right\|_F^2$$

- By proposition 1:
$$\left\|\widetilde{W}_{j+1:N}^T - W_{j+1:N}^T\right\|_F^2 \leq \sum_{i=j+1}^{N} \left\|\widetilde{W}_i^T - W_i^T\right\|_F^2$$
- $\widetilde{W}_{1:N} \in B^{d,d}$ because $\widetilde{W}_{1:N} = \widetilde{W}_N \cdots \widetilde{W}_1$ where $\widetilde{W}_i \in B^{d,d}\ \forall i \in [N]$ (from proposition 2), so, by note (ii)
$$\left\|\nabla l(\widetilde{W}_{1:N})\right\|_F^2 \leq M^2$$
- $\widetilde{W}_{1:N}^T \in B^{d,d}$ because $\widetilde{W}_{1:N}^T = (W_N \cdots W_1)^T = W_1^T \cdots W_N^T$ where $W_i^T \in B^{d,d}\ \forall i \in [N]$ (from proposition 2), so
$$\left\|\widetilde{W}_{1:N}^T\right\|_F \leq 1$$

Bringing it together:

$$(A) \leq \left\|\widetilde{W}_{j+1:N}^T - W_{j+1:N}^T\right\|_F^2 \left\|\nabla l(\widetilde{W}_{1:N})\right\|_F^2 \left\|\widetilde{W}_{1:j-1}^T\right\|_F^2 \leq M^2 \cdot \sum_{i=j+1}^{N} \left\|\widetilde{W}_i^T - W_i^T\right\|_F^2 \leq$$

$$\leq M^2 \cdot \sum_{i=1}^{N} \left\|\widetilde{W}_i^T - W_i^T\right\|_F^2 = M^2 \cdot \left\|(\widetilde{W}_1^T, \dots, \widetilde{W}_N^T) - (W_1^T, \dots, W_N^T)\right\|_2^2 =$$

$$= M^2 \cdot \left\|(\widetilde{W}_1, \dots, \widetilde{W}_N) - (W_1, \dots, W_N)\right\|_2^2$$

Bound on (B):

$$(B) = \left\| W_{j+1:N}^T \left(\nabla l(\widetilde{W}_{1:N}) - \nabla l(W_{1:N})\right) \widetilde{W}_{1:j-1}^T \right\|_F^2 \leq \left\| W_{j+1:N}^T \right\|_F^2 \left\|\nabla l(\widetilde{W}_{1:N}) - \nabla l(W_{1:N})\right\|_F^2 \left\|\widetilde{W}_{1:j-1}^T\right\|_F^2$$

- $\widetilde{W}_{j+1:N}^T \in B^{d,d}$ because $\widetilde{W}_{j+1:N}^T = \left(W_N \cdots W_{j+1}\right)^T = W_{j+1}^T \cdots W_N^T$ where $W_i^T \in B^{d,d}\ \forall i \in [N]$ (from proposition 2), so
$$\left\|\widetilde{W}_{j+1:N}^T\right\|_F^2 \leq 1$$
- by note (i):
$$\left\|\nabla l(\widetilde{W}_{1:N}) - \nabla l(W_{1:N})\right\|_F^2 \leq \beta^2 \cdot \left\|\widetilde{W}_{1:N} - W_{1:N}\right\|_F^2 \leq$$
Now, by proposition 1:
$$\leq \beta^2 \cdot \sum_{i=1}^{N} \left\|\widetilde{W}_i^T - W_i^T\right\|_F^2 = \beta^2 \cdot \left\|(\widetilde{W}_1, \dots, \widetilde{W}_N) - (W_1, \dots, W_N)\right\|_2^2$$
- $\widetilde{W}_{1:j-1}^T \in B^{d,d}$ because $\widetilde{W}_{1:j-1}^T = \left(W_{j-1} \cdots W_1\right)^T = W_1^T \cdots W_{j-1}^T$ where $W_i^T \in B^{d,d}\ \forall i \in [N]$ (from proposition 2), so
$$\left\|\widetilde{W}_{1:j-1}^T\right\|_F \leq 1$$

Bringing it together:

$$(B) \leq \left\| W_{j+1:N}^T \right\|_F^2 \left\|\nabla l(\widetilde{W}_{1:N}) - \nabla l(W_{1:N})\right\|_F^2 \left\|\widetilde{W}_{1:j-1}^T\right\|_F^2 \leq \beta^2 \cdot \left\|(\widetilde{W}_1, \dots, \widetilde{W}_N) - (W_1, \dots, W_N)\right\|_2^2$$

Bound on (C):

$$(C) = \left\|W_{j+1:N}^T \nabla l(W_{1:N})\left(\widetilde{W}_{1:j-1}^T - W_{1:j-1}^T\right)\right\|_F^2 \leq \left\|W_{j+1:N}^T\right\|_F^2 \|\nabla l(W_{1:N})\|_F^2 \left\|\widetilde{W}_{1:j-1}^T - W_{1:j-1}^T\right\|_F^2$$

- $\widetilde{W}_{j+1:N}^T \in B^{d,d}$ because $\widetilde{W}_{j+1:N}^T = \left(W_N \cdots W_{j+1}\right)^T = W_{j+1}^T \cdots W_N^T$ where $W_i^T \in B^{d,d}\ \forall i \in [N]$ (from proposition 2), so
$$\left\|\widetilde{W}_{j+1:N}^T\right\|_F \leq 1$$
- $W_{1:N} \in B^{d,d}$ because $W_{1:N} = W_N \cdots W_1$ where $W_i \in B^{d,d}\ \forall i \in [N]$ (from proposition 2), so, by note (ii)
$$\|\nabla l(W_{1:N})\|_F^2 \leq M^2$$
- By proposition 1:
$$\left\|\widetilde{W}_{1:j-1}^T - W_{1:j-1}^T\right\|_F^2 \leq \sum_{i=1}^{j-1} \left\|\widetilde{W}_i^T - W_i^T\right\|_F^2$$

Bringing it together:

$$(C) \leq \left\|W_{j+1:N}^T\right\|_F^2 \|\nabla l(W_{1:N})\|_F^2 \left\|\widetilde{W}_{1:j-1}^T - W_{1:j-1}^T\right\|_F^2 \leq M^2 \cdot \sum_{i=1}^{j-1} \left\|\widetilde{W}_i^T - W_i^T\right\|_F^2 \leq$$

$$\leq M^2 \cdot \sum_{i=1}^{N} \left\|\widetilde{W}_i^T - W_i^T\right\|_F^2 = M^2 \cdot \left\|\left(\widetilde{W}_1, \dots, \widetilde{W}_N\right) - (W_1, \dots, W_N)\right\|_2^2$$

So:

$$\left\|\nabla\phi\left(\widetilde{W}_1, \widetilde{W}_2, \dots, \widetilde{W}_N\right) - \nabla\phi(W_1, W_2, \dots, W_N)\right\|_2^2 \leq$$

$$\leq N \cdot \left((A) + (B) + (C)\right) \leq$$

$$\leq N \cdot \Big(M^2 \cdot \left\|\left(\widetilde{W}_1, \dots, \widetilde{W}_N\right) - (W_1, \dots, W_N)\right\|_2^2 + \beta^2 \cdot \left\|\left(\widetilde{W}_1, \dots, \widetilde{W}_N\right) - (W_1, \dots, W_N)\right\|_2^2 + M^2 \cdot \left\|\left(\widetilde{W}_1, \dots, \widetilde{W}_N\right) - (W_1, \dots, W_N)\right\|_2^2\Big) =$$

$$= \left(2 \cdot N \cdot M^2 + N \cdot \beta^2\right) \cdot \left\|\left(\widetilde{W}_1, \dots, \widetilde{W}_N\right) - (W_1, \dots, W_N)\right\|_2^2$$

Overall:

$$\left\|\nabla\phi\left(\widetilde{W}_1, \widetilde{W}_2, \dots, \widetilde{W}_N\right) - \nabla\phi(W_1, W_2, \dots, W_N)\right\|_2^2 \leq \left(2 \cdot N \cdot M^2 + N \cdot \beta^2\right) \cdot \left\|\left(\widetilde{W}_1, \dots, \widetilde{W}_N\right) - (W_1, \dots, W_N)\right\|_2^2$$

So ϕ: $B^{d,d} \times B^{d,d} \times \cdots \times B^{d,d} \to \mathbb{R}$ is a smooth ρ-Lipschitz-gradient function with $\rho = \sqrt{2 \cdot N \cdot M^2 + N \cdot \beta^2}$.

■

We will now prove the two remaining propositions:

1. $\forall k \in \mathbb{N}$, and $A_1, \dots, A_k, \tilde{A}_1, \dots, \tilde{A}_k \in B^{d,d}$:
$$\left\|\tilde{A}_{1:k} - A_{1:k}\right\|_F^2 \leq \sum_{i=1}^{k} \left\|\tilde{A}_i - A_i\right\|_F^2$$
Notice that if we denote $(A_1, \dots, A_k), \left(\tilde{A}_1, \dots, \tilde{A}_k\right) \in B^{d,d} \times B^{d,d} \times \cdots \times B^{d,d}$ we obtain:
$$\left\|\tilde{A}_{1:k} - A_{1:k}\right\|_F^2 \leq \left\|\left(\tilde{A}_1, \dots, \tilde{A}_k\right) - (A_1, \dots, A_k)\right\|_2^2$$

2. $\forall k \in \mathbb{N}$, and $A_1, \dots, A_k \in B^{d,d}$:
$$A_1 \cdots A_k \in B^{d,d}$$

Proof of proposition 1:

Proof by induction:

- $\underline{k = 1}$:
  Immediate.

$$\left\|\tilde{A}_{1:1} - A_{1:1}\right\|_F^2 = \left\|\tilde{A}_1 - A_1\right\|_F^2 = \sum_{i=1}^{1}\left\|\tilde{A}_i - A_i\right\|_F^2$$

- $\underline{(k-1) \to k}$:
  Let $A_1, \dots, A_k, \tilde{A}_1, \dots, \tilde{A}_k \in B^{d,d}$:
  $\left\|\tilde{A}_{1:k} - A_{1:k}\right\|_F^2 =$

$$= \left\|\tilde{A}_{1:k-1}\tilde{A}_k - A_{1:k-1}A_k\right\|_F^2 = \left\|\tilde{A}_{1:k-1}\tilde{A}_k - \tilde{A}_{1:k-1}A_k + \tilde{A}_{1:k-1}A_k - A_{1:k-1}A_k\right\|_F^2 =$$
$$= \left\|\widetilde{A}_{1:k-1}(\widetilde{A}_k - A_k) + (\widetilde{A}_{1:k-1} - A_{1:k-1})A_k\right\|_F^2 \overset{\Delta}{\leq} \left\|\tilde{A}_{1:k-1}(\tilde{A}_k - A_k)\right\|_F^2 + \left\|(\tilde{A}_{1:k-1} - A_{1:k-1})A_k\right\|_F^2 \leq$$
$$\leq \underbrace{\left\|\tilde{A}_{1:k-1}\right\|_F^2}_{\leq 1;prop.2}\left\|\tilde{A}_k - A_k\right\|_F^2 + \underbrace{\left\|\tilde{A}_{1:k-1} - A_{1:k-1}\right\|_F^2}_{(*)-induction}\underbrace{\|A_k\|_F^2}_{\leq 1} \leq \left\|\tilde{A}_k - A_k\right\|_F^2 + \sum_{i=1}^{k-1}\left\|\tilde{A}_i - A_i\right\|_F^2 =$$
$$= \sum_{i=1}^{k}\|\widetilde{A}_i - A_i\|_F^2$$

  Overall:

$$\left\|\tilde{A}_{1:k} - A_{1:k}\right\|_F^2 \leq \sum_{i=1}^{k}\left\|\tilde{A}_i - A_i\right\|_F^2$$

Proof of proposition 2:

Proof by induction

- $\underline{k = 1}$:
  Immediate.

- $\underline{(k-1) \to k}$:
  Let $A_1, \dots, A_k \in B^{d,d}$:

$$\|A_1 \cdots A_{k-1}A_k\|_F \leq \underbrace{\|A_1 \cdots A_{k-1}\|_F}_{\leq 1\ ;\ induction}\underbrace{\|A_k\|_F}_{\leq 1} \leq 1$$

  So $A_1 \cdots A_k \in B^{d,d}$

Proof (B):

Assume that $\phi(\cdot)$ is Lipschitz smooth on $\mathbb{R}^{d,d} \times \cdots \times \mathbb{R}^{d,d}$, I.e.
$\forall (W_1, \ldots, W_N), (\widetilde{W}_1, \ldots, \widetilde{W}_N) \in \mathbb{R}^{d,d} \times \cdots \times \mathbb{R}^{d,d}$:

$$\left\|\nabla\phi(\widetilde{W}_1, \ldots, \widetilde{W}_N) - \nabla\phi(W_1, \ldots, W_N)\right\|_2 \leq \rho\left\|(\widetilde{W}_1, \ldots, \widetilde{W}_N) - (W_1, \ldots, W_N)\right\|_2$$

Let $(W_1, \ldots, W_N), (\widetilde{W}_1, \ldots, \widetilde{W}_N) \in \mathbb{R}^{d,d} \times \cdots \times \mathbb{R}^{d,d}$, so:

$$\rho\left\|(\widetilde{W}_1, \ldots, \widetilde{W}_N) - (W_1, \ldots, W_N)\right\|_2 \geq$$

$$\geq \left\|\nabla\phi(\widetilde{W}_1, \ldots, \widetilde{W}_N) - \nabla\phi(W_1, \ldots, W_N)\right\|_2 =$$

$$= \sqrt{\sum_{i=1}^{N} \left\|\left(\nabla\phi(\widetilde{W}_1, \widetilde{W}_2, \ldots, \widetilde{W}_N) - \nabla\phi(W_1, W_2, \ldots, W_N)\right)_i\right\|_F^2} \geq$$

$$\geq \left\|\left(\nabla\phi(\widetilde{W}_1, \widetilde{W}_2, \ldots, \widetilde{W}_N) - \nabla\phi(W_1, W_2, \ldots, W_N)\right)_j\right\|_F ; \ \forall j \in [N]$$

So, $\forall j \in [N]$

$$\rho\left\|(\widetilde{W}_1, \ldots, \widetilde{W}_N) - (W_1, \ldots, W_N)\right\|_2 \geq$$

$$\geq \left\|\left(\nabla\phi(\widetilde{W}_1, \widetilde{W}_2, \ldots, \widetilde{W}_N) - \nabla\phi(W_1, W_2, \ldots, W_N)\right)_j\right\|_F \geq$$

$$\geq \left\|\widetilde{W}_{j+1:N}^T \nabla l(\widetilde{W}_{1:N}) \widetilde{W}_{1:j-1}^T - W_{j+1:N}^T \nabla l(W_{1:N}) W_{1:-1}^T\right\|_F =$$

We'll add and subtract terms:

$$= \left\| \underbrace{\left(\widetilde{W}_{j+1:N}^T - W_{j+1:N}^T\right) \nabla l(\widetilde{W}_{1:N}) \widetilde{W}_{1:j-1}^T}_{(A)} + \underbrace{W_{j+1:N}^T \left(\nabla l(\widetilde{W}_{1:N}) - \nabla l(W_{1:N})\right) \widetilde{W}_{1:j-1}^T}_{(B)} + \underbrace{W_{j+1:N}^T \nabla l(W_{1:N}) \left(\widetilde{W}_{1:j-1}^T - W_{1:j-1}^T\right)}_{(C)} \right\|_F$$

Overall:

$\forall (W_1, \ldots, W_N), (\widetilde{W}_1, \ldots, \widetilde{W}_N) \in \mathbb{R}^{d} \times \cdots \times \mathbb{R}^{d}$:

$$\rho\left\|(\widetilde{W}_1, \ldots, \widetilde{W}_N) - (W_1, \ldots, W_N)\right\|_2 \overset{(*)}{\geq}$$

$$\geq \left\| \underbrace{\left(\widetilde{W}_{j+1:N}^T - W_{j+1:N}^T\right) \nabla l(\widetilde{W}_{1:N}) \widetilde{W}_{1:j-1}^T}_{(A)} + \underbrace{W_{j+1:N}^T \left(\nabla l(\widetilde{W}_{1:N}) - \nabla l(W_{1:N})\right) \widetilde{W}_{1:j-1}^T}_{(B)} + \underbrace{W_{j+1:N}^T \nabla l(W_{1:N}) \left(\widetilde{W}_{1:j-1}^T - W_{1:j-1}^T\right)}_{(C)} \right\|_F$$

First, we assume $N = 2$. We would like to prove $l(\cdot)$ is affine.
So $j \in \{1,2\}$. Choose $j = 1$

- Define a series of matrices $k \in \mathbb{N}$:
$$W_1^{(k)} \coloneqq \frac{1}{k} W_1, \widetilde{W_1}^{(k)} \coloneqq \frac{1}{k} \widetilde{W_1}$$
Where $W_1, \widetilde{W_1} \in \mathbb{R}^{d,d}$ are arbitrary matrices (const w.r.t k)
- Define a series of matrices $k \in \mathbb{N}$:
$$W_2^{(k)} \coloneqq k \cdot I_d, \widetilde{W_2}^{(k)} \coloneqq k \cdot I_d$$
Where $I_d \in \mathbb{R}^{d,d}$ is the identity matrix.

For every $k \in \mathbb{N}$, inequality $(*)$ holds for $\left(W_1^{(k)}, W_2^{(k)}\right), \left(\widetilde{W_1}^{(k)}, \widetilde{W_2}^{(k)}\right)$
So:

- $LHS = \rho \left\| \left(\widetilde{W_1}^{(k)}, \widetilde{W_2}^{(k)}\right) - \left(W_1^{(k)}, W_2^{(k)}\right) \right\|_2 = \rho \left\| \widetilde{W_1}^{(k)} - W_1^{(k)} \right\|_F = \frac{\rho}{k} \left\| \widetilde{W_1} - W_1 \right\|_F$
- $(A) = \left(\widetilde{W}_2{}^{(k)^{\mathrm{T}}} - W_2^{(k)^T} \right) \left(\nabla l \left(\widetilde{W}_2^{(k)} \widetilde{W_1}^{(k)} \right) \right) = (kI_d - kI_d)\left(\nabla l(\widetilde{W}_2 \widetilde{W_1})\right) = 0$
- $(C) = W_2^{(k)^T} \nabla l\left(W_2^{(k)} W_1^{(k)}\right)(I_d - I_d) = 0$
- $(B) = W_2^{(k)^T} \left(\nabla l \left(\widetilde{W_2}^{(k)} \widetilde{W_1}^{(k)} \right) - \nabla l\left(W_2^{(k)} W_1^{(k)}\right) \right) = k \cdot \left(\nabla l \left(k \cdot \frac{1}{k} \cdot \widetilde{W_1} \right) - \nabla l \left(k \cdot \frac{1}{k} \cdot W_1 \right) \right)$
$= k \cdot \left(\nabla l(\widetilde{W_1}) - \nabla l(W_1) \right)$

So:

$$\underbrace{\frac{\rho}{k} \left\| \widetilde{W_1} - W_1 \right\|_F}_{\xrightarrow[k \to \infty]{} 0} \geq \underbrace{k}_{\xrightarrow[k \to \infty]{} \infty} \left\| \nabla l(\widetilde{W_1}) - \nabla l(W_1) \right\|_F$$

$\Longrightarrow \left\| \nabla l\left(\widetilde{W_1}\right) - \nabla l(W_1) \right\|_F \xrightarrow[k \to \infty]{} 0$, but $\left\| \nabla l\left(\widetilde{W_1}\right) - \nabla l(W_1) \right\|_F$ is constant w.r.t k
so $\left\| \nabla l\left(\widetilde{W_1}\right) - \nabla l(W_1) \right\|_F = 0 \ \forall \widetilde{W_1}, W_1 \in \mathbb{R}^{d,d}$
$\Longrightarrow \nabla l(\cdot) \equiv const \Longrightarrow l(\cdot)$ is affine.

Now, we assume $N > 2$. We would like to prove $\ell(\cdot)$ is constant.

$j \in [N]$. Choose $j = 2$

$$\rho\left\|\left(\widetilde{W}_1, \dots, \widetilde{W}_N\right) - (W_1, \dots, W_N)\right\|_2 \overset{(*)}{\geq}$$

$$\geq \left\| \underbrace{\left(\widetilde{W}_{j+1:N}^T - W_{j+1:N}^T\right)\nabla l(\widetilde{W}_{1:N})\widetilde{W}_{1:j-1}^T}_{(A)} + \underbrace{W_{j+1:N}^T\left(\nabla l(\widetilde{W}_{1:N}) - \nabla l(W_{1:N})\right)\widetilde{W}_{1:j-1}^T}_{(B)} + \underbrace{W_{j+1:N}^T\nabla l(W_{1:N})\left(\widetilde{W}_{1:j-1}^T - W_{1:j-1}^T\right)}_{(C)} \right\|_F =$$

$$= \left\| \underbrace{\left(\widetilde{W}_{3:N}^T - W_{3:N}^T\right)\nabla l(\widetilde{W}_{1:N})\widetilde{W}_1^T}_{(A)} + \underbrace{W_{3:N}^T\left(\nabla l(\widetilde{W}_{1:N}) - \nabla l(W_{1:N})\right)\widetilde{W}_1^T}_{(B)} + \underbrace{W_{3:N}^T\nabla l(W_{1:N})\left(\widetilde{W}_1^T - W_1^T\right)}_{(C)} \right\|_F$$

- Define a series of matrices $k \in \mathbb{N}$:
  $$\widetilde{W}_1^{(k)} = k^{N-2}I_d \;\; ; \;\; W_1^{(k)} = k^{N-2}I_d$$
  Where I_d is the $d \times d$ identity matrix
- Define a series of matrices $k \in \mathbb{N}$:
  $$\widetilde{W}_2^{(k)} = W_2 \;\; ; \;\; W_2^{(k)} = W_2$$
  Where $W_2 \in \mathbb{R}^{d,d}$ is an arbitrary matrix
- For all $3 \leq i \leq N$, define a series of matrices $k \in \mathbb{N}$:
  $$\widetilde{W}_i^{(k)} = \frac{1}{k}I_d \;\; ; \;\; W_i^{(k)} = 0$$

For every $k \in \mathbb{N}$, inequality $(*)$ holds for $\left(W_1^{(k)}, \dots, W_N^{(k)}\right), \left(\widetilde{W}_1^{(k)}, \dots, \widetilde{W}_N^{(k)}\right)$

So:

- $LHS = \rho\left\|\left(\widetilde{W}_1^{(k)}, \dots, \widetilde{W}_N^{(k)}\right) - \left(W_1^{(k)}, \dots, W_N^{(k)}\right)\right\|_2 =$
  $= \rho\left\|\left(0,0,\widetilde{W}_3^{(k)} - W_3^{(k)}, \dots, \widetilde{W}_N^{(k)} - \widetilde{W}_N^{(k)}\right)\right\|_2 = \rho\left\|\left(0,0,\frac{1}{k}I_d, \dots, \frac{1}{k}I_d\right)\right\|_2 = \frac{\rho}{k}(N-2)\|I_d\|_F =$
  $= \frac{\rho(N-2)d}{k}$
- $(A) = \left(\widetilde{W_{3:N}}^{(k)^T} - W_{3:N}^{(k)^T}\right)\nabla l\left(\widetilde{W}_{1:N}^{(k)}\right)\widetilde{W}_1^{(k)^T} = \left(\frac{1}{k^{N-2}}I_d - 0\right)\nabla l\left(\widetilde{W_{3:N}}^{(k)}\widetilde{W}_2^{(k)}\widetilde{W}_1^{(k)}\right)k^{N-2}I_d =$
  $= \nabla l\left(\frac{1}{k^{N-2}}I_d W_2 k^{N-2}I_d\right) = \nabla l(W_2)$
- $(B) = W_{3:N}^T\left(\nabla l(\widetilde{W}_{1:N}) - \nabla l(W_{1:N})\right)\widetilde{W}_1^T = 0\left(\nabla l(\widetilde{W}_{1:N}) - \nabla l(W_{1:N})\right)\widetilde{W}_1^T = 0$
- $(C) = W_{3:N}^T\nabla l(W_{1:N})\left(\widetilde{W}_1^T - W_1^T\right) = 0\nabla l(W_{1:N})\left(\widetilde{W}_1^T - W_1^T\right) = 0$

So,

$$\underbrace{\frac{\rho(N-2)d}{k}}_{\xrightarrow[k\to\infty]{}0} \geq \|\nabla l(W_2)\|_F$$

$\Longrightarrow \|\nabla l(W_2)\|_F \xrightarrow[k\to\infty]{} 0$, but $\|\nabla l(W_2)\|_F$ is constant w.r.t k so $\|\nabla l(W_2)\|_F = 0 \; \forall W_2 \in \mathbb{R}^{d,d}$

$\Longrightarrow \|\nabla l(\cdot)\|_F \equiv 0 \Longrightarrow l(\cdot)$ is const.

It remains to prove the opposite direction:

(1) $l(\cdot)$ is constant $\Longrightarrow \phi(\cdot)$ is smooth
(2) $l(\cdot)$ is affine and $N = 2 \Longrightarrow \phi(\cdot)$ is smooth

Proof:

(1) $l(\cdot) = const \Longrightarrow \nabla l(\cdot) \equiv 0$
(2) $l(\cdot)$ is affine $\Longrightarrow \nabla l(\cdot) \equiv 0$

In either case $\nabla l(\cdot) \equiv 0$. This will be the only assumption we need.

We saw in "Proof (A)":

$\forall (W_1, \dots, W_N), (\widetilde{W_1}, \dots, \widetilde{W_N}) \in \mathbb{R}^{d,d} \times \cdots \times \mathbb{R}^{d,d}$:

$$\left\|\nabla\phi(\widetilde{W_1}, \widetilde{W_2}, \dots, \widetilde{W_N}) - \nabla\phi(W_1, W_2, \dots, W_N)\right\|_2^2 \leq$$

$$\leq N \cdot \left(\underbrace{\left\|(\widetilde{W}_{j+1:N}^T - W_{j+1:N}^T)\nabla l(\widetilde{W}_{1:N})\widetilde{W}_{1:j-1}^T\right\|_F^2}_{(A)} + \underbrace{\left\|W_{j+1:N}^T\left(\nabla l(\widetilde{W}_{1:N}) - \nabla l(W_{1:N})\right)\widetilde{W}_{1:j-1}^T\right\|_F^2}_{(B)} \right.$$

$$\left. + \underbrace{\left\|W_{j+1:N}^T \nabla l(W_{1:N})(\widetilde{W}_{1:j-1}^T - W_{1:j-1}^T)\right\|_F^2}_{(C)} \right) =$$

$$= N \cdot (0 + 0 + 0) = 0$$

So in particular, selecting $\rho = 1$:

$\forall (W_1, \dots, W_N), (\widetilde{W_1}, \dots, \widetilde{W_N}) \in \mathbb{R}^{d,d} \times \cdots \times \mathbb{R}^{d,d}$:

$$(0 =)\left\|\nabla\phi(\widetilde{W_1}, \widetilde{W_2}, \dots, \widetilde{W_N}) - \nabla\phi(W_1, W_2, \dots, W_N)\right\|_2^2 \leq \rho^2 \left\|(\widetilde{W_1}, \widetilde{W_2}, \dots, \widetilde{W_N}) - (W_1, W_2, \dots, W_N)\right\|_2^2$$

$\Longrightarrow \phi(\cdot)$ is ρ-smooth on $\mathbb{R}^{d,d} \times \cdots \times \mathbb{R}^{d,d}$.

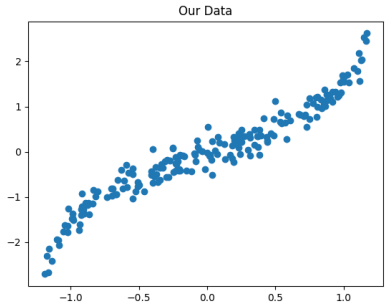
■

Question 4:

We used in our experiment a LNN network with the following properties:

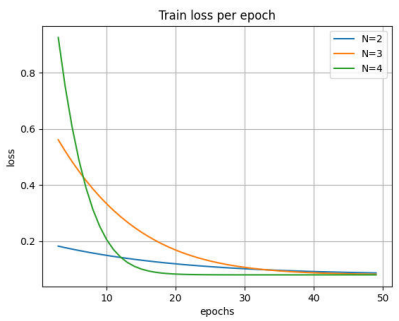
1. Input dimension = 1
2. Output dimensions = 1
3. Width of hidden layers = 3
4. Depth of the LNN: N=2,3,4

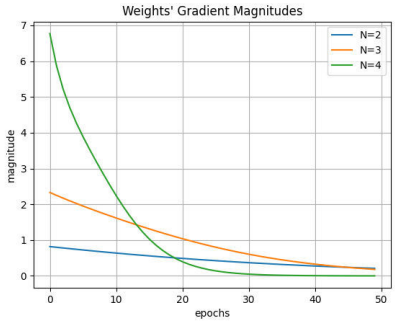
We ran our experiments on the following dataset:

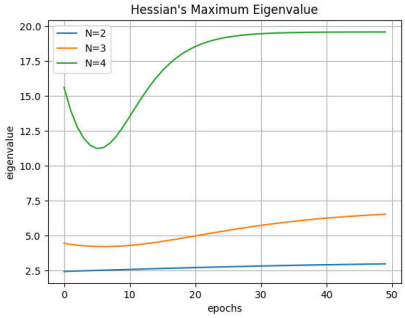


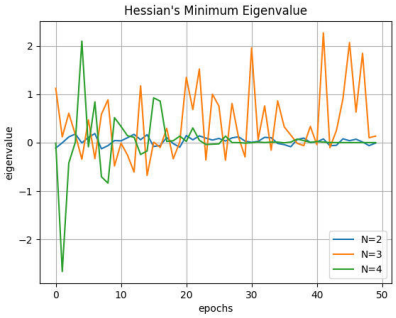
We chose uniformly 200 random points between -1.2 and 1.2, we calculate for each point the function $tan(x)$ and added noise which was chosen normally around 0 with standard deviation of 0.2.

The results of our experiment are the followings:









Those are the results in each epoch while running of train loss, magnitude of weights' gradient, maximal value of Hessian of the weights and minimum value of Hessian of the weights.

Explanation of the results:

1. No bad local minima – as we proved in the lecture of optimization 1, for LNN networks without a "bottleneck" layers (which is the case in our case) every minima is a global one for the objective function which induced by the loss of the network.
   We can clearly see that result by the figure of train loss per epoch as all the values of the loss (for N=2, 3, 4 networks) converge to the same value (and not stay stuck in a local minima which is not global).

2. From the proposition in class where we vectorized the end-to-end matrix and wrote the dynamics with the preconditioned PSD matrix $P_{W_{1:N}(t)}$, we concluded that the end-to-end dynamics are in fact an instance of preconditioned GF, Thus overparameterizing the loss with a LNN "promotes movement in directions already taken" – which means stretches the gradient in singular directions of large singular values and attenuates it in directions of small singular values.

   a. We can clearly see the results we derived in class here in our results - Note that in the figure of maximal eigenvalue of the Hessian, as N is bigger; the maximal value of the Hessian is bigger. Thus, it implies on the movement of the gradient such that as N is bigger the weights converge **faster** to the global min, as we can see in the figure of the train loss (as N is bigger, it converge faster). In addition the magnitude of the gradient decreases faster as N is bigger as we can see in our results.

   b. In addition we can see (in the figure with the maximal and minimal eigenvalues of the Hessian) that in convergence the minimal eigenvalue is bigger than 0 (very close but still bigger), and thus we converge to a Hessian which is PSD so we really get closer to a local minima point (also a global as we mentioned in 1) as the point is a stationary point.

3. From the lecture of optimization 1, we proved that for LNN with $depth = 2$ any stationary point which is not a global minimizer - a strict saddle point.
   In addition we proved that for LNNs with $depth \geq 3$ the objective function doesn't have any non-strict saddle, and that GD escapes a strict saddle point if it arrives to it.
   We clearly can see from the figure of the training loss that since all the networks converge to the same value - it strengthens the proof of escaping strict saddle points for the networks such that they will arrive to a global min.

Part 3 - Trajectory approach

Question 1

We have to show that under the conditions of the respective theorem:

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

$$\forall\, t \in \mathbb{R}_{\geq 0}\,, \quad j \in [N]$$

Proof

By definition: $W_{1:j}(t) = W_j(t) \cdots W_1(t)$

As we proved in lecture of optimization 2:

$\forall\, j \in [N-1]$ $\exists\, V_j \in \mathbb{R}^{d \times d}$ orthogonal and $\exists\, D_j \in \mathbb{R}^{d \times d}$ diagonal matrix whose diagonal values are in $\{+1, -1\}$

Such that $W_j(t) = V_{j+1} D_j \Sigma V_j^T$

For fixed $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_d) \in \mathbb{R}^{d \times d}$, $\forall i\ \sigma_i \geq 0$

Therefore:

$$W_{1:j}(t) = W_j(t) \cdots W_1(t) = \underline{V_{j+1} D_j \Sigma V_j^T}\ \underline{V_j D_{j-1} \Sigma V_{j-1}} \cdots \underline{V_2 D_1 \Sigma V_1^T}$$

$\Rightarrow$

$$W_{1:j}(t)^T W_{1:j}(t) = (V_{j+1} D_j \Sigma V_j^T \cdots V_2 D_1 \Sigma V_1^T)^T (V_{j+1} D_j \Sigma V_j^T \cdots V_2 D_1 \Sigma V_1^T)$$

$$= (V_1 \Sigma^T D_1^T V_2^T \cdots V_j \Sigma^T D_j^T V_{j+1}^T)(V_{j+1} D_j \Sigma V_j^T \cdots V_2 D_1 \Sigma V_1^T)$$

Since the matrices $\Sigma, D_1, \dots, D_N$ are diagonal, it means

$\forall j \quad D_j^T = D_j$, $\Sigma^T = \Sigma$.

$$\Rightarrow \quad W_{1:j}(t)^T W_{1:j}(t) =$$

$$(V_1 \Sigma D_1 \underbrace{V_2^T V_2} \Sigma D_2 \underbrace{V_3^T V_3} \cdots \underbrace{V_j^T V_j} \Sigma D_j \underbrace{V_{j+1}^T)(V_{j+1}} D_j \Sigma \underbrace{V_j^T} \cdots \underbrace{V_2^T V_2} D_1 \Sigma V_1^T)$$

$$= (V_1 \Sigma D_1 \Sigma D_2 \cdots \Sigma D_j)(D_j \Sigma D_{j-1} \cdots D_1 \Sigma V_1^T)$$

$\forall j \quad V_j^T V_j = I$

$$= V_1 \Sigma^{2j} D_1^2 \cdots D_j^2 V_1^T$$

$D_1, \dots, D_j, \Sigma$ are diagonal and therefore comutative w.r.t matrix multiplication

$$\Rightarrow \quad W_{1:j}(t)^T W_{1:j}(t) = V_1 \Sigma^{2j} D_1^2 \cdots D_j^2 V_1^T = V_1 \Sigma^{2j} V_1^T$$

(since $D_j^2 = I \ \forall j$ because its diagonal matrix with values ± 1 on diagonal)

In particular $W_{1:N}(t)^T W_{1:N}(t) = V_1 \Sigma^{2N} V_1^T$

$$\Rightarrow \quad \forall t \in \mathbb{R}_{\geq 0} \ , \ j \in [N] :$$

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

□

Part 3 - question 2

Let $\ell: \mathbb{R}^{d\times d} \to \mathbb{R}$ be a continously differentiable loss parameterized by: $\phi: \mathbb{R}^{d,d} \to \mathbb{R}$

$$\phi(U) = \ell(UU^T)$$

$W(t) := U(t)U(t)^T$.

Develop an expression for the dynamics of $W(t)$ that are induced by gradient flow over $\phi(\cdot)$:

$$\dot{U}(t) = \frac{d}{dt}U(t) = -\nabla\phi(U(t))$$

Solution

We'll calculate the dynamics for $W(t)$:

$$\dot{W}(t) = \frac{d}{dt}W(t) = \frac{d}{dt}(U(t)U(t)^T) = \dot{U}(t)U(t)^T + U(t)\dot{(U(t)}^T)$$

$$= \dot{U}(t)U(t)^T + U(t)\dot{U}(t)^T =$$

$$= -\nabla\phi(U(t))U(t)^T + U(t)\cdot(-\nabla\phi(U(t))^T$$

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

Now, we would like to continue develop the expression by develop the term $\nabla\phi(U(t))$ using the first order Taylor expansions of ϕ and ℓ.

As we know: $\phi(U) = \ell(UU^T)$

First order Taylor expansion of ϕ around U:

$$\phi(U+\Delta) = \phi(U) + \langle \nabla\phi(U), \Delta \rangle + o(\|\Delta\|_{Fro})$$

We also know that:

$$\phi(U+\Delta) = \ell((U+\Delta)(U+\Delta)^T) =$$
$$= \ell((U+\Delta)(U^T+\Delta^T)) =$$
$$= \ell(UU^T + U\Delta^T + \Delta U^T + \Delta\Delta^T)$$

Develop the first order Taylor expansion of ℓ around UU^T will yield:

$$\ell(UU^T + \underline{U\Delta^T + \Delta U^T + \Delta\Delta^T}) =$$
$$= \ell(UU^T) + \langle \nabla\ell(UU^T), U\Delta^T + \Delta U^T + \Delta\Delta^T \rangle + o(\|\Delta\|_{Fro})$$
$$= \ell(UU^T) + \langle \nabla\ell(UU^T), U\Delta^T + \Delta U^T \rangle$$
$$+ \langle \nabla\ell(UU^T), \Delta\Delta^T \rangle + o(\|\Delta\|_{Fro})$$

We've got: $\phi(U+\Delta) = \phi(U) + \langle \nabla\phi(U), \Delta \rangle + o(\|\Delta\|_{Fro})$

$$\ell(UU^T + U\Delta^T + \Delta U^T + \Delta\Delta^T) =$$
$$= \ell(UU^T) + \langle \nabla\ell(UU^T), U\Delta^T + \Delta U^T \rangle$$
$$+ \langle \nabla\ell(UU^T), \Delta\Delta^T \rangle + o(\|\Delta\|_{Fro})$$

Since $\phi(U) = \ell(UU^T)$, $\phi(U+\Delta) = \ell(UU^T + U\Delta^T + \Delta U^T + \Delta\Delta^T)$ and the uniqueness of first order Δ Taylor expansion we conclude:

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

We've got: $\langle \nabla\phi(U), \Delta \rangle = \langle \nabla e(UU^T), U\Delta^T + \Delta U^T \rangle$ ⊜

$\ominus \langle \nabla e(UU^T), U\Delta^T \rangle + \langle \nabla e(UU^T), \Delta U^T \rangle =$

$= tr(\nabla e(UU^T)^T U\Delta^T) + tr(\nabla e(UU^T)^T \Delta U^T) =$

$tr(AB^T) = tr(A^TB)$

$= tr((\nabla e(UU^T)^T U)^T \Delta) + tr(\nabla e(UU^T)^T \Delta U^T)$

$= tr(U^T \nabla e(UU^T) \Delta) + tr(\nabla e(UU^T)^T \Delta U^T) =$

$tr(ABC) = tr(ACB)$

$= tr(U^T \nabla e(UU^T)\Delta) + tr(\nabla e(UU^T)^T U^T \Delta)$

$= tr(\nabla e(UU^T) U^T \Delta) + tr(\nabla e(UU^T)^T U^T \Delta)$

$= tr((U \nabla e(UU^T)^T)^T \Delta) + tr((U \nabla e(UU^T))^T \Delta)$

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

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

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

$\Rightarrow \quad \nabla\phi(U) = U(\nabla e(UU^T) + \nabla e(UU^T)^T)$

Plugging it in ⊛:

$\dot{W}(t) = -\nabla\phi(U(t))\, U(t)^T - U(t) \cdot \nabla\phi(U(t))^T =$

$= -U[\nabla e(UU^T) + \nabla e(UU^T)^T] U^T - U[U(\nabla e(UU^T) + \nabla e(UU^T)^T)]^T$

[$U = U(t)$ for simplicity]

$= -U[\nabla e(UU^T) + \nabla e(UU^T)^T] U^T - U[\nabla e(UU^T) + \nabla e(UU^T)^T]^T U^T$

$= -U[\nabla e(UU^T) + \nabla e(UU^T)^T] U^T - U[\nabla e(UU^T) + \nabla e(UU^T)^T U^T]$

$= -2 \cdot U[\nabla e(UU^T) + \nabla e(UU^T)^T] U^T$

$\Rightarrow \dot{W}(t) = -2\, U(t)[\nabla e(U(t)U(t)^T) + \nabla e(U(t)U(t)^T)^T] U(t)^T$

□

Question 3

Simlify the end-to-end dynamics delivered in class for the special case of $d_N = 1$. Explain how this resonates with our interpretation of the end-to-end dynamics "promoting movement in direction already taken"

Solution

As we've already showed the end-to-end dynamics are:

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

$t \in \mathbb{R}_{\geq 0}$

By the assumption $d_N = 1$, thus $W_{1:N}(t) = W_N(t) \cdots W_1(t) \in \mathbb{R}^{1, d_0}$

dims: $(1, d_{N-1}), \cdots (d_1, d_0)$

(1) Therefore, for each $j \in [N]$:

$$W_{1:N}(t) W_{1:N}(t)^T = \langle W_{1:N}(t), W_{1:N}(t) \rangle = \|W_{1:N}(t)\|_2^2 \in \mathbb{R}$$

$$[W_{1:N}(t) W_{1:N}(t)^T]^{\frac{j-1}{N}} = \|W_{1:N}(t)\|_2^{2 \cdot \frac{j-1}{N}}$$

(2) On the other hand, $W_{1:N}(t)^T W_{1:N}(t) \in \mathbb{R}^{d,d}$, thus:

- For $j = N$: $[W_{1:N}(t)^T W_{1:N}(t)]^{\frac{N-j}{N}} = [W_{1:N}(t)^T W_{1:N}(t)]^0 = I_{d,d}$

↑ By definition

- Claim

For $j = 1, \ldots, N-1$:

$$[W_{1:N}(t)^T W_{1:N}(t)]^{\frac{N-j}{N}} = \|W_{1:N}(t)\|_2^{2 \cdot \frac{N-j}{N}} \cdot \frac{W_{1:N}(t)^T}{\|W_{1:N}(t)\|_2} \cdot \frac{W_{1:N}(t)}{\|W_{1:N}(t)\|_2}$$

Proof of claim:

First, we know that $W_{1:N}(t)^T W_{1:N}(t)$ is PSD matrix and thus all it's singular values are ≥ 0. Notice that $\text{rank}(W_{1:N}(t)) = 1 \Rightarrow \text{rank}(W_{1:N}(t)^T W_{1:N}(t)) \leq 1$.

In addition $W_{1:N}(t)^T$ is an eigenvector with eigenvalue $\|W_{1:N}(t)\|_2^2$:

$$[W_{1:N}(t)^T W_{1:N}(t)]\, W_{1:N}(t)^T = W_{1:N}(t)^T \cdot \|W_{1:N}(t)\|_2^2 = \|W_{1:N}(t)\|_2^2 \cdot W_{1:N}(t)^T$$

Thus $\sigma_1 := \|W_{1:N}(t)\|_2^2 \neq 0$ is the only non-zero eigenvalue of $W_{1:N}(t)^T W_{1:N}(t)$. [Suppose $\|W_{1:N}(t)\|_2 \neq 0$]

We'll write the SVD composition of $W_{1:N}(t)^T W_{1:N}(t)$:

$$W_{1:N}(t)^T W_{1:N}(t) = U \Sigma V^T = U \begin{bmatrix} \sigma_1 & & \\ & 0 & \\ & & \ddots \end{bmatrix} V^T$$

Thus for every $x \geq 0$:

$$[W_{1:N}(t)^T W_{1:N}(t)]^x = U \Sigma^x V^T = U \begin{pmatrix} \sigma_1^x & & \\ & 0 & \\ & & \ddots \end{pmatrix} V^T$$

$$= \sigma_1^x \cdot U \begin{pmatrix} \sigma_1 & & \\ & 0 & \\ & & \ddots \end{pmatrix} V^T = \|W_{1:N}(t)\|_2^{2x}$$

$$\frac{}{\sqrt{\sigma_1}\ \sqrt{\sigma_1}}$$

$$= \|W_{1:N}(t)\|_2^{2x} \cdot \frac{W_{1:N}(t)^T \cdot W_{1:N}(t)}{\sqrt{\sigma_1} \cdot \sqrt{\sigma_1}} =$$

$$= \|W_{1:N}(t)\|_2^{2x} \cdot \left(\frac{W_{1:N}(t)}{\|W_{1:N}(t)\|_2}\right)^T \cdot \left(\frac{W_{1:N}(t)}{\|W_{1:N}(t)\|_2}\right)$$

We can use $x = \frac{N-j}{N}$ and get the required.

Now, we can use the equations ① and ② into ③ and we'll get:

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

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

$$- \|W_{1:N}(t)\|_2^{2\cdot\frac{N-1}{N}} \nabla \ell(W_{1:N}(t)) \cdot I_{d,d}$$

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

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

$$= -\|W_{1:N}(t)\|_2^{2-\frac{2}{N}} \left[\nabla \ell(W_{1:N}(t)) + (N-1) \cdot \nabla \ell(W_{1:N}(t)) \frac{W_{1:N}(t)^T}{\|W_{1:N}(t)\|_2} \cdot \frac{W_{1:N}(t)}{\|W_{1:N}(t)\|_2} \right]$$

Where we assumed $W_{1:N}(t) \neq \vec{0}$ and $N \geq 2$ on our proof.

We've got: $\dot{W}_{1:N}(t) = -\|W_{1:N}(t)\|_2^{2-\frac{2}{N}} \cdot \left[\nabla \ell(W_{1:N}(t)) + (N-1) \cdot \nabla \ell(W_{1:N}(t)) \frac{W_{1:N}(t)^T}{\|W_{1:N}(t)\|_2} \cdot \frac{W_{1:N}(t)}{\|W_{1:N}(t)\|_2} \right]$

Thus, besides the multiplication by $\|W_{1:N}(t)\|_2^{2-\frac{2}{N}}$, we also add to $\nabla \ell(W_{1:N}(t))$ the projection of itself on $W_{1:N}(t)$.

Recall that we view $W_{1:N}(t)$ not only as the parameters but also as the overall movement of the system.

Thus by adding the projection of the gradient on $W_{1:N}(t)$ we can think above it as promoting movements in directions which already taken.

□

Foundations of Deep Learning – Homework Assignment #3

Adi Album & Tomer Epshtein

Part 3: (4)

Question:

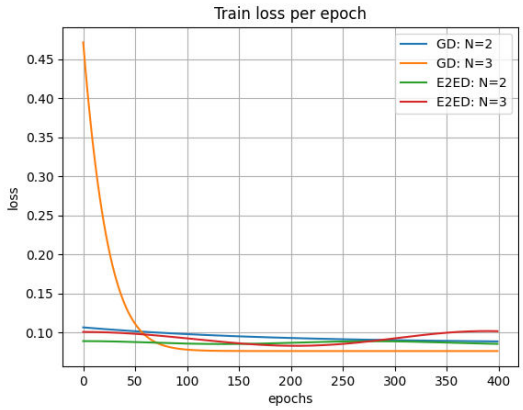
On a scalar regression dataset of your choice, train a depth N linear neural network (with hidden widths no smaller that the minimum between input dimension and output dimension) by minimizing ℓ_2 loss via (full batch) gradient descent with small learning rate and initialization close to zero. Compare the trajectory taken by the end-to-end matrix to that obtained by directly applying the (discrete version of the) end-to-end dynamics to a linear model:

$$W_{t+1} \leftarrow W_t - \eta \sum_{j=1}^{N} [W_t W_t^T]^{\frac{j-1}{N}} \nabla l(W_t) [W_t^T W_t]^{\frac{N-j}{N}} \quad ; \quad t = 1,2,3,\dots$$

Where η is the learning rate used for gradient descent over the linear neural network. Repeat the experiment with depths $N = 2$ and 3

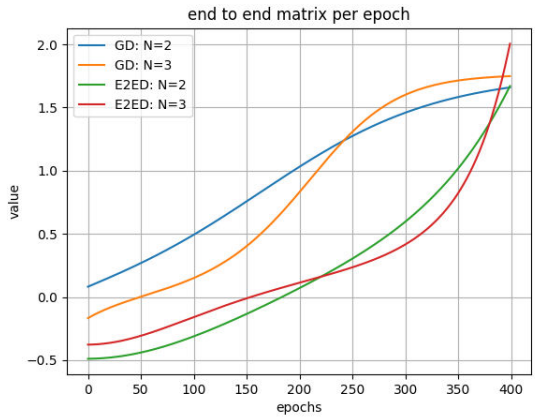
Solution:

The training procedure can be viewed in the following plots:



Convergence of all four models: GD with $N = 2,3$ and E2ED (End-to-end dynamics) with $N = 2,3$ is achieved.

Additionally, because our network is a Linear Neural Network with input dimension 1 and output dimension 1, the end-to-end matrix is a 1×1 matrix. I.e. a single real value. We can see convergence to global minimum, with different learning rates, in the following graph, where all four models converge to and end-to-end matrix $\approx (1.7)_{1\times1}$



Part 4 - Ultra wide neural networks

Question 1

$$\dot{u}(t) = -H^*(u(t)-y) \qquad t \in \mathbb{R}_{\geq 0}$$

Where:

- $u(t) \in \mathbb{R}^m$ holds the neural network predictions at time t.
- $H^* \in \mathbb{R}^{m,m}$ is the Neural Tangent Kernel's Gram matrix PD with eigenvalues $\geq \lambda_{min} > 0$
- $y \in \mathbb{R}^m$ holds training labels.

We have to show that $u(t) \longrightarrow y$ exponentially fast without using the change of variables as done in class.

Solution

We'll develop the expression $\frac{d}{dt}\|u(t)-y\|^2$:

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

$$= 2\cdot(u(t)-y)^T \cdot \dot{u}(t)$$

$$= -2\,(u(t)-y)^T H^*(u(t)-y)$$

[Notice that $\lambda_{min}(H^*) = \min_{x \neq 0 \in \mathbb{R}^m} \frac{x^T H^* x}{\|x\|^2} \Rightarrow \forall x \quad x^T H^* x \geq \lambda_{min} \cdot \|x\|^2$]

$$\Rightarrow \frac{d}{dt}\|u(t)-y\|^2 = -2\,(u(t)-y)^T H^* (u(t)-y) \leq -2\cdot\lambda_{min}\cdot\|u(t)-y\|^2$$

$[x = u(t)-y]$

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

Thus, we can write it as:

$$\frac{d}{dt}\ln(\|u(t)-y\|^2) = \frac{\frac{d}{dt}\|u(t)-y\|^2}{\|u(t)-y\|^2} \leq -2\lambda_{min}$$

$$\Downarrow$$

$$\int_0^t \frac{d}{ds}\ln(\|u(s)-y\|^2) \leq \int_0^t -2\lambda_{min}$$

$$\Downarrow$$

$$\ln(\|u(s)-y\|^2)[s=t] - \ln(\|u(s)-y\|^2)[s=0] \leq -2\lambda_{min}t$$

$$\ln\left(\frac{\|u(t)-y\|^2}{\|u(0)-y\|^2}\right) \leq -2\lambda_{min}t \quad / \; e^{\uparrow} \text{ is monotonic}$$

$$\Downarrow$$

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

$$\Downarrow$$

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

$\Rightarrow$ $u(t) \longrightarrow y$ exponentially fast as required.

□

Question 2

The dynamics in question (1) (Part 4) are acctually an idealization of:

$$\dot{u}(t) = -H(t)\,(u(t)-y) \qquad t \in \mathbb{R}_{\geq 0} \qquad (*)$$

Where $H(t) \in \mathbb{R}^{n \times n}$ satisfies $\|H(t) - H^*\|_{spec} \leq \varepsilon$.

Assuming identical initialization $u(0)$, show that under $(*)$ the value of $u(t)$ is at most $O(\sqrt{t\varepsilon})$ away (in Euclidean distance) from what it would be under question (1) dynamics.

Solution

① Recall that we proved in question (1) that under the dynamics $\dot{u}_1(t) = -H^*(u_1(t)-y)$ we got:

$$\|u_1(t)-y\|^2 \leq \|u_1(0)-y\|^2 \cdot e^{-2\lambda_{min} t}$$

$$\Rightarrow \|u_1(t)-y\|^2 \leq \|u_1(0)-y\|^2 \cdot e^{-2\lambda_{min} t}$$

② We'll prove that under the dynamics $\dot{u}_2(t) = -H(t)(u_2(t)-y)$ we'll get:

$$\|u_2(t)-y\|^2 \leq \|u(0)-y\|^2 \cdot e^{-2(\lambda_{min}+\varepsilon)t}$$

Proof

Note: $\frac{d}{dt}\|u_2(t)-y\|^2 = (u_2(t)-y)^T \frac{d}{dt}(u_2(t)) =$

$$= -(u_2(t)-y)^T H(t)\,(u_2(t)-y)$$

$$\Rightarrow \frac{d}{dt}\|u_2(t)-y\|^2 = -(u_2(t)-y)^T H(t)\,(u_2(t)-y)$$

By the assumption $\forall_t \quad \|H(t) - H^*\|_{spe} \le \varepsilon$.

Since $\|\cdot\|_{spe}$ is a norm then we conclude:

$$\|H(t)\|_{spec} = \|H(t) - H^* + H^*\|_{spec}$$

$$\geq \|H^*\|_{spe} - \|H(t) - H^*\|_{spe} \quad \leftarrow \Delta \text{ inequality}$$

$$\geq \lambda_{min} - \|H(t) - H^*\|_{spe} \quad \leftarrow \text{by definition } H^* \text{ PSD so } \|H^*\|_{spe} \geq \lambda_{min}$$

$$\geq \lambda_{min} - \varepsilon$$

$\Rightarrow$

$$\frac{d}{dt}\|u_2(t) - y\|^2 = -2(u_2(t) - y)^T H(t)(u_2(t) - y) \quad \left[\forall_x \; x^T A x \le \|A\|_{spe} \cdot \|x\|^2\right]$$

$$\le -2\|H(t)\|_{spe} \cdot \|u_2(t) - y\|^2$$

$$\le -2(\lambda_{min} - \varepsilon)\|u_2(t) - y\|^2$$

$$\Rightarrow \quad \frac{\frac{d}{dt}\|u_2(t) - y\|^2}{\|u_2(t) - y\|^2} \le -2(\lambda_{min} - \varepsilon)$$

$\Rightarrow$ With exactly the same steps as in question 1

where we proved $\frac{\frac{d}{dt}\|u(t) - y\|^2}{\|u(t) - y\|^2} \le -2\lambda_{min} \Rightarrow \|u(t) - y\|^2 \le \|u(0) - y\|^2 \cdot e^{-2\lambda_{min} t}$

we'll get the required:

$$\|u_2(t) - y\|^2 \le \|u_2(0) - y\|^2 \cdot e^{-2(\lambda_{min} - \varepsilon)t} = \|u(0) - y\|^2 \cdot e^{-2(\lambda_{min} - \varepsilon)t}$$

as we claimed.

Therefore, from ① and ② we've got:

$$\|u_1(t)-y\|^2 \le \|u(0)-y\|^2 \cdot e^{-2\lambda_{min} t}$$

$$\|u_2(t)-y\|^2 \le \|u(0)-y\|^2 \cdot e^{-2(\lambda_{min}-\varepsilon)t}$$

Now:

$$\|u_1(t)-u_2(t)\|_2^2 = \|u_1(t)-y-(u_2(t)-y)\|^2$$

$$\left[\|u-v\|^2 \le 2\|u\|^2+2\|v\|^2\right] \le 2\cdot\|u_1(t)-y\|^2 + 2\|u_2(t)-y\|^2$$

$$\le 2\cdot\|u(0)-y\|^2\cdot e^{-2\lambda_{min}t} + 2\cdot\|u(0)-y\|^2\cdot e^{-2(\lambda_{min}-\varepsilon)t}$$

$$= 2\|u(0)-y\|^2\cdot\left(e^{-2\lambda_{min}t} + e^{-2(\lambda_{min}-\varepsilon)t}\right)$$

$e^{-2\lambda_{min}t} \le e^0 = 1$ $\forall t \ge 0$ since $\lambda_{min} > 0$

$$\le 2\cdot\|u(0)-y\|^2\cdot\left(1 + e^{-2(\lambda_{min}-\varepsilon)t}\right)$$

$e^{-2(\lambda_{min}-\varepsilon)t} \le 1$ for $\varepsilon < \lambda_{min}$

$$\le 2\cdot\|u(0)-y\|^2\cdot(1+1)$$

$\varepsilon t \ge 1$ for $t \ge \frac{1}{\varepsilon}$

$$\le 2\cdot\|u(0)-y\|^2\cdot(\varepsilon t + \varepsilon t) = 4\cdot\|u(0)-y\|^2\,\varepsilon t$$

$\Rightarrow$ for $\varepsilon < \lambda_{min}(H^*)$, $t \ge \frac{1}{\varepsilon} = t_0$

$$\|u_1(t)-u_2(t)\| \le \sqrt{4\cdot\|u(0)-y\|^2\cdot\varepsilon t} = 2\cdot\|u(0)-y\|\cdot\sqrt{\varepsilon t}$$

$\Rightarrow$ $\|u_1(t)-u_2(t)\| \le O(\sqrt{\varepsilon t})$

as required.

□