Foundation of Deep Leaning - Ex. 3

Nadav Gat & Yoav Nagel

IDs: 207291683 & 205947336

a short README.md file is hereby attached to explain the handout format, please view it.

Part III

Trajectory Approach

Linear Neural Networks

Exercise 1. As we saw in class, we can write $\forall j \in [N-1] : W_j(t) = V_{j+1} D_j \Sigma V_j^T$ for some $V_j \in \mathbb{R}^{d \times d}$ orthogonal matrices, $D_j = \operatorname{diag}(a_{j,1}, ..., a_{j,d})$ where $\forall i \in [d] : a_i \in \{-1, 1\}$ and $\Sigma = \operatorname{diag}(\sigma_1, ..., \sigma_d)$ with $\sigma_1 \geq ... \geq \sigma_d \geq 0$. Thus

$$\begin{aligned} W_{1:j}(t) &:= W_j(t) W_{j-1}(t) ... W_1(t) = V_{j+1} D_j \Sigma \underbrace{V_j^T V_j}_{I_d} D_{j-1} \Sigma V_{j-1}^T ... V_2 D_1 \Sigma V_1^T \\ &= V_{j+1} D_j \Sigma D_{j-1} \Sigma ... D_1 \Sigma V_1^T = V_{j+1} \Sigma^j D_j D_{j-1} ... D_1 V_1^T \end{aligned}$$

where we used orthogonality properties, and the fact that diagonal matrices commute in multiplication. Therefore

$$\begin{aligned} W_{1:j}(t)^T W_{1:j}(t) &= \left(V_{j+1} \Sigma^j D_j D_{j-1} ... D_1 V_1^T\right)^T V_{j+1} \Sigma^j D_j D_{j-1} ... D_1 V_1^T \\ &= V_1 D_1^T ... D_j^T \left(\Sigma^j\right)^T \underbrace{V_{j+1}^T V_{j+1}} \Sigma^j D_j D_{j-1} ... D_1 V_1^T \\ &= V_1 \Sigma^{2j} \underbrace{D_1 D_1} ... D_j D_j V_1^T \\ &= V_1 \Sigma^{2j} V_1^T \end{aligned}$$

where used the above mentioned properties again, that diagonal matrices are symmetrical and $D_i D_i = I_d$. In particular, we have

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

And combining these expression we get the desired

$$W_{1:j}(t)^T W_{1:j}(t) = V_1 \Sigma^{2j} V_1^T = \left(V_1 \Sigma^{2N} V_1^T\right)^{\frac{j}{N}} = \left[W_{1:N}(t)^T W_{1:N}(t)\right]^{\frac{j}{N}}$$

Exercise 2. First, we saw in class that GF results in

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

where under our settings, we define $\phi(U) = \ell\left(UU^T\right)$. Let us first define $f : \mathbb{R}^{d,d} \to \mathbb{R}^{d,d}$ as $f(X) = XX^T$, and using chain rule we have

$$\nabla\phi(U) = \left\langle \nabla\ell\left(UU^T\right), \nabla f(U)\right\rangle \tag{0.1}$$

Notice that for any $i, j \in [d]$ that $\left(UU^T\right)_{i,j} = \sum_{k=1}^{d} (U)_{ik}\left(U^T\right)_{kj} = \sum_{k=1}^{d} u_{ik}u_{jk}$ by matrix multiplication. Thus we can deduce

$$\frac{\partial f_{i,j}}{\partial u_{s,t}} = \begin{cases} 2u_{s,t} & i = j = s \\ u_{j,t} & i \neq j \wedge i = s \\ u_{i,t} & i \neq j \wedge j = s \\ 0 & else \end{cases}$$

Plugging this into 0.1 , we can write

$$\begin{aligned}
(\nabla\phi(U))_{s,t} &= \sum_{i,j}\left(\nabla\ell\left(UU^T\right)\right)_{i,j}\frac{\partial f_{i,j}}{\partial u_{s,t}} \\
&= 2u_{s,t}\left(\nabla\ell\left(UU^T\right)\right)_{s,t} + \sum_{i=1,i\neq j}^{d} u_{i,t}\left(\nabla\ell\left(UU^T\right)\right)_{i,s} + \sum_{j=1,i\neq j}^{d} u_{j,t}\left(\nabla\ell\left(UU^T\right)\right)_{j,s} \\
&= 2\sum_{i=1}^{d} u_{i,t}\left(\nabla\ell\left(UU^T\right)\right)_{i,s} \\
&= 2\sum_{i=1}^{d}\left(\nabla\ell\left(UU^T\right)^T\right)_{s,i} u_{i,t}
\end{aligned}$$

and thus we get $\nabla\phi(U) = 2\nabla\ell\left(UU^T\right)^T U$, and

$$\dot{U_t} = -2\nabla\ell\left(U(t)\,U(t)^T\right)^T U(t)$$

With that conclusion, we can proceed to understand the dynamics of W:

$$\begin{aligned}
\dot{W_t} &= \frac{d}{dt}\left(U(t)\,U(t)^T\right) = \left(\frac{d}{dt}U(t)\right)U(t)^T + U(t)\left(\frac{d}{dt}U(t)^T\right) \\
&= -2\nabla\ell\left(U(t)\,U(t)^T\right)^T U(t)\,U(t)^T - 2U(t)\,U(t)^T\,\nabla\ell\left(U(t)\,U(t)^T\right) \\
&= -2\nabla\ell(W(t))^T\,W(t) - 2W(t)\,\ell(W(t)) \\
&= -2\left(X(t) + X(t)^T\right)
\end{aligned}$$

for $X(t) := \nabla\ell(W(t))^T\,W(t)$

Exercise 3. Let us notice that we have $W_{1:N} \in \mathbb{R}^{1,d_0}$ in our special case. Therefore,

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

where we write $||\cdot||$ for the ℓ_2 norm. Denote $W_{1:N} = \Sigma V^T$ the SVD decomposition of the e2e matrix. This is written not as the usual $U\Sigma V^T$ form since in our case $U \in \mathbb{R}$ and makes an orthonormal basis, meaning $= 1$. As we saw in class, we have for the end-to-end dynamics

$$\begin{aligned}
\dot{W}_{1:N}(t) &= -\sum_{i=1}^{N}\left[W_{1:N}(t)\,W_{1:N}^T(t)\right]^{\frac{i-1}{N}}\nabla\ell\left(W_{1:N}(t)\right)\left[W_{1:N}^T(t)\,W_{1:N}(t)\right]^{\frac{N-i}{N}} \\
&= -||W_{1:N}(t)||^{2\cdot\frac{N-1}{N}}\nabla\ell\left(W_{1:N}(t)\right) - \sum_{i=1}^{N-1}||W_{1:N}(t)||^{2\cdot\frac{i-1}{N}}\nabla\ell\left(W_{1:N}(t)\right)\left[V\Sigma\Sigma V^T\right]^{\frac{N-i}{N}} \\
&= -||W_{1:N}(t)||^{2\cdot\frac{N-1}{N}}\nabla\ell\left(W_{1:N}(t)\right) - \sum_{i=1}^{N-1}||W_{1:N}(t)||^{2\cdot\frac{i-1}{N}}\nabla\ell\left(W_{1:N}(t)\right)V\Sigma^{2\cdot\frac{N-i}{N}}V^T
\end{aligned}$$

We shall explore this decomposition in our case, noticing $V\Sigma^2V^T$ is an eigendecomposition of $W_{1:N}^T(t)\,W_{1:N}(t)$. Since $\text{rank}\,(W_{1:N}) = 1$, we can say $\text{rank}\left(W_{1:N}^T(t)\,W_{1:N}(t)\right) \leq 1$ and therefore $\Sigma^2 = \text{diag}\,(\sigma, 0, ..., 0)$ for some $\sigma \geq 0$. Moreover,

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

which implies that regardless of its norm value, we have an eigenvalue equal to the norm of the e2e matrix, therefore $\sigma = ||W_{1:N}(t)||^2$. This allows us to proceed with

$$\begin{aligned}\Sigma^{2\cdot\frac{N-i}{N}} &= \left(\Sigma^2\right)^{\frac{N-i}{N}} = \text{diag}\Big(||W_{1:N}(t)||^{2\cdot\frac{N-i}{N}}, 0, .., 0\Big) = \\ &= ||W_{1:N}(t)||^{2\cdot\frac{N-i}{N}} \cdot \text{diag}\,(1, 0, ..., 0)\end{aligned}$$

Plugging this in the dynamics, we get

$$\begin{aligned}\dot{W}_{1:N}(t) &= -||W_{1:N}(t)||^{2\cdot\frac{N-1}{N}} \nabla\ell\,(W_{1:N}(t)) - \sum_{i=1}^{N-1} ||W_{1:N}(t)||^{2\cdot\frac{i-1}{N}} \nabla\ell\,(W_{1:N}(t))\,V\left(||W_{1:N}(t)||^{2\cdot\frac{N-i}{N}} \cdot \text{diag}\,(1, 0, ..., 0)\right)V^T \\ &= -||W_{1:N}(t)||^{2\cdot\frac{N-1}{N}} \left(\nabla\ell\,(W_{1:N}(t)) + (N-1)\,\nabla\ell\,(W_{1:N}(t))\,V\text{diag}\,(1, 0, ..., 0)\,V^T\right)\end{aligned}$$

which is a desired expression since we can give it our interpretation. Ignoring the multiplicative factor and its possible meanings, we can see that in addition to proceeding with the direction determined by the gradient, we also have the gradient multiplied by a PSD matrix that promotes merely its principal component, i.e its largest eigenvector, all multiplied by $N-1$. This implies another amplification of the step towards the most dominant direction that was taken to get the current e2e matrix, which suits our proposition.

Ultra Wide Neural Networks

Exercise 1. Following the given hint, we can get

$$\begin{aligned}\frac{d}{dt}||u(t) - y||^2 &= 2\,(u(t) - y)^T \cdot \frac{d}{dt}u(t) \\ &= 2\,(u(t) - y)^T\,\dot{u}(t) \\ &= -2\,(u(t) - y)^T\,H^*\,(u(t) - y)\end{aligned}$$

Notice that H^* is a PSD matrix, and in particular it is symmetric.

Claim. Given a symmetric matrix $A \in \mathbb{R}^{m\times m}$ and $f : \mathbb{R}^m \setminus \{0\} \to \mathbb{R}$ defined $f(x) = \frac{x^T Ax}{||x||^2}$, we have

$$\min_{x\in\mathbb{R}^m} f(x) = \lambda_{min}(A)$$
$$\max_{x\in\mathbb{R}^m} f(x) = \lambda_{max}(A)$$

Proof. Recall that we can write $A = V\Lambda V^T$ for some orthogonal V, where Λ is a diagonal matrix holding $\{\lambda_i\}_{i=1}^m$ the eigenvalues corresponding to the columns of V, which are the eigenvectors of A. Since V's columns make an orthonormal basis, we can write every $x \in \mathbb{R}^m$ as $\sum_{i=1}^m \alpha_i v_i$ where v_i is the i-th column of V. Thus

$$\begin{aligned} x^T Ax &= \left(\sum_{i=1}^m \alpha_i v_i^T\right) A \sum_{i=1}^m \alpha_i v_i \\ &= \sum_{i=1}^m \alpha_i v_i^T \cdot \sum_{i=1}^m \lambda_i \alpha_i v_i \\ &= \sum_{i=1}^m \alpha_i^2 \lambda_i \end{aligned}$$

where we used orthonormality on the last equation. Additionally, $||x||^2 = \sum_{i=1}^m \alpha_i v_i^T \cdot \sum_{i=1}^m \alpha_i v_i = \sum_{i=1}^m \alpha_i^2$ for similar reasoning. This gives us

$$g(\alpha) = f(x) = \frac{\sum_{i=1}^m \alpha_i^2 \lambda_i}{\sum_{i=1}^m \alpha_i^2}$$

where $g(\alpha) := f(\alpha_i v_i)$. Let us show what are the max and min values of g, and since any x can be represented as this linear combination it would also be max and min of f respectively.

$$\begin{aligned} \frac{\sum_{i=1}^m \alpha_i^2 \lambda_i}{\sum_{i=1}^m \alpha_i^2} &\le \frac{\sum_{i=1}^m \alpha_i^2 \lambda_{max}}{\sum_{i=1}^m \alpha_i^2} = \frac{\lambda_{max} \cdot \sum_{i=1}^m \alpha_i^2}{\sum_{i=1}^m \alpha_i^2} = \lambda_{max} \\ \frac{\sum_{i=1}^m \alpha_i^2 \lambda_i}{\sum_{i=1}^m \alpha_i^2} &\ge \frac{\sum_{i=1}^m \alpha_i^2 \lambda_{min}}{\sum_{i=1}^m \alpha_i^2} = \frac{\lambda_{min} \cdot \sum_{i=1}^m \alpha_i^2}{\sum_{i=1}^m \alpha_i^2} = \lambda_{min} \end{aligned}$$

which can be obtained with the corresponding one-hot vectors of α, and this proves the claim above. ■

Using the proven claim we can deduce

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

and this gives us

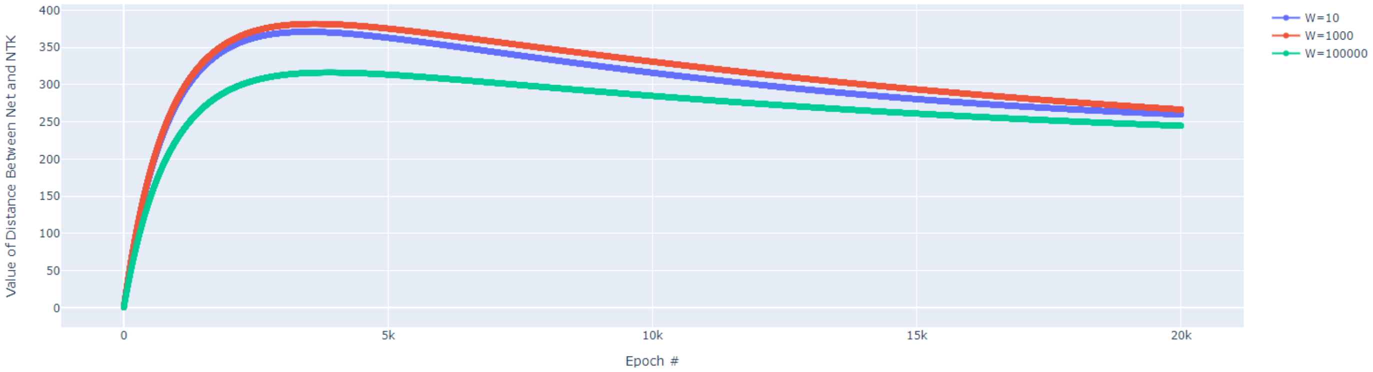
$$-2\lambda_{min}(H^*) \ge \frac{\frac{d}{dt}||u(t) - y||^2}{||u(t) - y||^2} = \frac{d}{dt}\ln\left(||u(t) - y||^2\right)$$

Taking integral from both sides

$$\begin{aligned} \int_0^t -2\lambda_{min}(H^*) &\ge \int_0^t \frac{d}{dx}\ln\left(||u(x) - y||^2\right) \\ \Rightarrow -2\lambda_{min}(H^*)\, t &\ge \ln||u(t) - y||^2 - \ln||u(0) - y||^2 = \ln\left(\frac{||u(t) - y||^2}{||u(0) - y||^2}\right) \\ \Rightarrow &\frac{||u(t) - y||^2}{||u(0) - y||^2} \le e^{-2\lambda_{min}(H^*)t} \\ \Rightarrow &||u(t) - y||^2 \le e^{-2\lambda_{min}(H^*)t}||u(0) - y||^2 \end{aligned}$$

Thus $u(t) \to y$ exponentially fast.

Exercise 3. The experiment suggests that after around 3000 epochs, the output vector of the network u and the output vector of the NTK u' start to converge and get closer, which was measured by norm of $u - u'$. First, the norm goes higher, which implies that initially the dynamics of those algorithm are quite different. However, from a certain point the norm declines, which can be interpreted as a convergence in mapping over those outputs. As epochs are running, the algorithms start to suggest similar solutions.
One can observe that the ultra-wide network has a lower peak, but accordingly the narrower networks show faster pace of convergence. Still, the graph suggests a slight advantage to wider networks in minimizing the difference, meaning those are closer to NTK.


W=10
W=1000
W=100000
Value of Distance Between Net and NTK
5k
10k
15k
20k
Epoch #