Foundations of Deep Learning **Fall 2024**

Exercise 5: Optimization Landscape of Neural Networks

Lecturer: Aurelien Lucchi

Problem 1 (Matrix Completion):

We consider the problem of matrix sensing where we have a model parametrized by $\mathbf{W} \in \mathbb{R}^{d\times d}$. We observe a set of linear measurements of the form $\langle \mathbf{W}, \mathbf{A}_n \rangle_F$ where $[\mathbf{A}_n]_{ij} \sim \mathcal{N}(0,1)$. We will further assume that the data is labeled by a matrix sensing model parameterized by $\mathbf{W}^* \in \mathbb{R}^{d\times d}$ (this is sometimes called "planted model" in the literature).
We will study the dynamics of this model trained with gradient flow on a squared loss. As we will soon see, this setting is related to the problem of training a deep linear network. In order to simulate depth, we will consider the matrix $\mathbf{W}$ as the product of a set of square matrices $\mathbf{W}_i \in \mathbb{R}^{d\times d}$, i.e. $\mathbf{W} = \mathbf{W}_L \dots \mathbf{W}_1$.
The objective function is given by

$$\mathcal{L}(\mathbf{W}) = \frac{1}{2L}\mathbb{E}_{\mathbf{A}}\langle \mathbf{W} - \mathbf{W}^*, \mathbf{A}\rangle_F^2 = \frac{1}{2L}\|\mathbf{W}_L \dots \mathbf{W}_1 - \mathbf{W}^*\|_F^2.$$

We will use gradient flow to optimize the parameters and denote by $\mathbf{W}_k(t)$ the matrices $\mathbf{W}_k$ at time t.

a) Denote $\mathbf{W}_{j:k}^\top = \prod_{i=j}^k \mathbf{W}_i^\top = \mathbf{W}_j^\top \mathbf{W}_{j+1}^\top \cdots \mathbf{W}_k^\top$. The partial gradient of $\mathcal{L}$ with respect to $\mathbf{W}_k$ (where $k = 1, \dots, L$) is

$$\frac{\partial \mathcal{L}(\mathbf{W})}{\partial \mathbf{W}_k} = \frac{1}{L}\mathbf{W}_{k+1:L}^\top(\mathbf{W} - \mathbf{W}^*)\mathbf{W}_{1:k-1}^\top.$$

What is the gradient flow equation for the matrix $\mathbf{W}_k$?

b) Prove that for all $t \geq 0$ and $k = 1, ..., L-1$:

$$\mathbf{W}_{k+1}^\top(t)\dot{\mathbf{W}}_{k+1}(t) = \dot{\mathbf{W}}_k(t)\mathbf{W}_k^\top(t). \tag{1}$$

And hence

$$\mathbf{W}_{k+1}^\top(t)\mathbf{W}_{k+1}(t) = \mathbf{W}_k(t)\mathbf{W}_k^\top(t). \tag{2}$$

c) For any $t \geq 0$ and $k = 1, ..., L$, assume the singular values of $\mathbf{W}_k$ are all distinct and indexed in strictly decreasing order: $\sigma_1 > ... > \sigma_d > 0$, and let $\mathbf{W}_k = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^\top$ be its (unique) singular value decomposition (SVD). Show that

$$\mathbf{\Sigma}_{k+1} = \mathbf{\Sigma}_k, \qquad \mathbf{U}_k = \mathbf{V}_{k+1},$$

for any $t \geq 0$ and $k = 1, ..., L$.

d) Denote $\mathbf{\Sigma} = \mathbf{\Sigma}_1 = \cdots = \mathbf{\Sigma}_L \in \mathbb{R}^{d\times d}$. Now prove that the gradient flow equation can be written as:

$$\dot{\mathbf{W}}_k = \frac{1}{L}\mathbf{V}_{k+1}\mathbf{\Sigma}^{L-k}\mathbf{U}_L^\top\left(\mathbf{W}^* - \mathbf{U}_L\mathbf{\Sigma}^L\mathbf{V}_1^\top\right)\mathbf{V}_1\mathbf{\Sigma}^{k-1}\mathbf{V}_{k-1}^\top,$$

for all $k = 1, .., L$.

e) By the product rule $\dot{\mathbf{W}} = \sum_{k=1}^L \mathbf{W}_{L:k+1}\dot{\mathbf{W}}_k\mathbf{W}_{k-1:1}$, show that

$$\dot{\mathbf{W}} = \frac{1}{L}\sum_{k=1}^L \mathbf{U}_L\mathbf{\Sigma}^{2L-2k}\mathbf{U}_L^\top\left(\mathbf{W}^* - \mathbf{U}_L\mathbf{\Sigma}^L\mathbf{V}_1^\top\right)\mathbf{V}_1\mathbf{\Sigma}^{2k-2}\mathbf{V}_1^\top.$$

f) Alternatively, show that the gradient flow can be expressed solely in terms of $\mathbf{W}$:

$$\dot{\mathbf{W}} = \frac{1}{L}\sum_{k=1}^L \left[\mathbf{W}\mathbf{W}^\top\right]^{\frac{L-k}{L}}(\mathbf{W}^* - \mathbf{W})\left[\mathbf{W}^\top\mathbf{W}\right]^{\frac{k-1}{L}}.$$

Problem 2 (Network near initialization):
Given an input vector $\mathbf{x} \in \mathbb{R}^d$, consider a shallow neural network defined by

$$f(\mathbf{W}) := \sum_{j=1}^{m} a_j \sigma(\mathbf{w}_j^\top \mathbf{x}), \qquad \mathbf{W} := \begin{bmatrix} \leftarrow \mathbf{w}_1^\top \rightarrow \\ \vdots \\ \leftarrow \mathbf{w}_m^\top \rightarrow \end{bmatrix} \in \mathbb{R}^{m \times d},$$

where σ is a β-smooth activation, $\mathbf{a} \in \mathbb{R}^m$ is a fixed (not trainable) vector of weights initialized such that $a_j \overset{\text{i.i.d.}}{\sim} \text{Unif}\left(\pm\frac{1}{\sqrt{m}}\right)$ [1], and $\mathbf{W} \in \mathbb{R}^{m \times d}$ are trainable weights.

We will consider the linearization of the function f around some initial weights $\mathbf{W}_0$ defined by

$$f_0(\mathbf{W}) = f(\mathbf{W}_0) + \langle \nabla f(\mathbf{W}_0), \mathbf{W} - \mathbf{W}_0 \rangle_F \,, \tag{3}$$

where $\langle \mathbf{A}, \mathbf{B} \rangle_F = \text{tr}(\mathbf{A}^\top \mathbf{B})$ is the Frobenius inner product.

a) Show that

$$\nabla f(\mathbf{W}) = \mathbf{D}\mathbf{a}\mathbf{x}^\top, \text{ where } \mathbf{D} = \text{diag}(\sigma'(\mathbf{w}_i^\top \mathbf{x})) = \begin{pmatrix} \sigma'(\mathbf{w}_1^\top \mathbf{x}) & & & \\ & \sigma'(\mathbf{w}_2^\top \mathbf{x}) & & \\ & & \cdots & \\ & & & \sigma'(\mathbf{w}_m^\top \mathbf{x}) \end{pmatrix}.$$

b) Show that the linearization of f around initialization $\mathbf{W}_0 = \begin{bmatrix} \leftarrow \mathbf{w}_{0,1}^\top \rightarrow \\ \vdots \\ \leftarrow \mathbf{w}_{0,m}^\top \rightarrow \end{bmatrix}$ is equal to

$$f_0(\mathbf{W}) = \sum_{j=1}^{m} a_j \left(\left[\sigma(\mathbf{w}_{0,j}^\top \mathbf{x}) - \sigma'(\mathbf{w}_{0,j}) \mathbf{w}_{0,j}^\top \mathbf{x} \right] + \sigma'(\mathbf{w}_{0,j}) \mathbf{w}_j^\top \mathbf{x} \right).$$

c) Show that for any $\mathbf{W}, \mathbf{V} \in \mathbb{R}^{m \times d}$,

$$|f(\mathbf{W}) - f_0(\mathbf{V})| \leq \frac{\beta}{2\sqrt{m}} \|\mathbf{W} - \mathbf{V}\|_F^2 \|\mathbf{x}\|^2.$$

d) What do you conclude about the role of over-parametrization?

[1] Equivalently, one can define $f(\mathbf{W}) := \frac{1}{\sqrt{m}} \sum_{j=1}^{m} a_j \sigma(\mathbf{w}_j^\top \mathbf{x})$ with $a_j \overset{\text{i.i.d.}}{\sim} \text{Unif}(\pm 1)$.