Homework 1

Deep Learning

Duarte Calado de Almeida
95565

André Lopes Rodrigues
96576

Question 1

Question 2

Question 3

1. Let x_i denote the i-th component of $\boldsymbol{x}$ and let w_{ij} denote the entry in the i-th row and j-th column in matrix $\boldsymbol{W}$. We then have that:

$$
\begin{aligned}
h_i(\boldsymbol{x}) &= g\left(\sum_{j=1}^{D} w_{ij}x_j\right) = \left(\sum_{j=1}^{D} w_{ij}x_j\right)^2 = \left(\sum_{j=1}^{D} w_{ij}x_j\right)\left(\sum_{k=1}^{D} w_{ik}x_k\right) \\
&= \sum_{j=1}^{D}\sum_{k=1}^{D} w_{ij}x_j w_{ik}x_k = \sum_{j=1}^{D} w_{ij}^2 x_j^2 + \sum_{j=1}^{D}\sum_{k\neq j} w_{ij}w_{ik}x_jx_k \\
&= \sum_{j=1}^{D} w_{ij}^2 x_j^2 + \sum_{j=1}^{D}\sum_{k=1}^{j-1} w_{ij}w_{ik}x_jx_k + \sum_{j=1}^{D}\sum_{k=j+1}^{D} w_{ij}w_{ik}x_jx_k \\
&= \sum_{j=1}^{D} w_{ij}^2 x_j^2 + \sum_{k=1}^{D}\sum_{j=k+1}^{D} w_{ij}w_{ik}x_jx_k + \sum_{j=1}^{D}\sum_{k=j+1}^{D} w_{ij}w_{ik}x_jx_k \\
&= \sum_{j=1}^{D} w_{ij}^2 x_j^2 + \sum_{k=1}^{D}\sum_{j=k+1}^{D} w_{ij}w_{ik}(2x_jx_k) \\
&= \begin{bmatrix} w_{i1}^2 & w_{i1}w_{i2} & \dots & w_{i1}w_{iD} & w_{i2}^2 & w_{i2}w_{i3} & \dots & w_{i(D-1)}^2 & w_{i(D-1)}w_{iD} & w_{iD}^2 \end{bmatrix}
\begin{bmatrix} x_1^2 \\ 2x_1x_2 \\ \dots \\ 2x_1x_D \\ x_2^2 \\ 2x_2x_3 \\ \dots \\ x_{D-1}^2 \\ 2x_{D-1}x_D \\ x_D^2 \end{bmatrix}
\end{aligned}
$$

As such, $\boldsymbol{h}$ is linear in some feature transformation $\boldsymbol{\phi}$, that is, $\boldsymbol{h}$ can be written as $\boldsymbol{A}_\Theta\boldsymbol{\phi}(\boldsymbol{x})$. In particular, we have that such matrix $\boldsymbol{A}_\Theta$ can be defined as:

$$\boldsymbol{A}_\Theta = \begin{bmatrix} — \ \boldsymbol{a}_1^T \ — \\ — \ \boldsymbol{a}_2^T \ — \\ \vdots \\ — \ \boldsymbol{a}_K^T \ — \end{bmatrix}$$

where

$$\boldsymbol{a}_i = \begin{bmatrix} w_{i1}^2 & w_{i1}W_{i2} & \dots & w_{i1}w_{iD} & w_{i2}^2 & w_{i2}w_{i3} & \dots & w_{i(D-1)}^2 & w_{i(D-1)}w_{iD} & w_{iD}^2 \end{bmatrix}^T$$

and so $\boldsymbol{A}_\Theta \in \mathbb{R}^{K\times\frac{D(D+1)}{2}}$ (since $\sum_{k=1}^D k = \frac{D(D+1)}{2}$). Furthermore, we can define the feature transformation $\boldsymbol{\phi} : \mathbb{R}^D \mapsto \mathbb{R}^{\frac{D(D+1)}{2}}$ as:

$$\boldsymbol{\phi}(\boldsymbol{x}) = (x_1^2, 2x_1x_2, \dots, 2x_1x_D, x_2^2, 2x_2x_3, \dots, x_{D-1}^2, 2x_{D-1}x_D, x_D^2)$$

2. Given that the predicted output $\hat{y}$ is defined as:

$$\hat{y} = \boldsymbol{v}^T\boldsymbol{h}$$

the linearity of $\boldsymbol{h}$ in the feature transformation $\boldsymbol{\phi}(\boldsymbol{x})$ proven above leads to following equality:

$$\hat{y} = \boldsymbol{v}^T\boldsymbol{A}_\Theta\boldsymbol{\phi}(\boldsymbol{x}) = (\boldsymbol{A}_\Theta^T\boldsymbol{v})^T\boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{c}_\Theta^T\boldsymbol{\phi}(\boldsymbol{x})$$

where we take $\boldsymbol{c}_\Theta$ to be equal to $\boldsymbol{A}_\Theta^T\boldsymbol{v}$, thereby proving that $\hat{y}$ is also a linear transformation of $\boldsymbol{\phi}(\boldsymbol{x})$. However, $\hat{y}$ is **not** linear in terms of the original parameters Θ. To see this, note that the model is now a linear combination of **products** of entries of $\boldsymbol{W}$ and $\boldsymbol{v}$ rather than being linear in **each** individual entry:

$$\hat{y} = \boldsymbol{v}^T\boldsymbol{A}_\Theta\boldsymbol{\phi}(\boldsymbol{x}) = \sum_{i=1}^{D} v_i(\boldsymbol{w}_i^T\boldsymbol{x})^2 = \sum_{i=1}^{K}\sum_{j=1}^{D}\sum_{k=1}^{D} v_i w_{ij} w_{ik} x_i x_k$$

where we define $\boldsymbol{w}_i$ to be the vector in the i-th row of matrix $\boldsymbol{W}$.

3. To prove the desired result, for $\boldsymbol{c}_\Theta$ defined in the previous subquestion and for any $\boldsymbol{c} \in \mathbb{R}^{\frac{D(D+1)}{2}}$, we make the observation that the inner products $\boldsymbol{c}_\Theta^T\boldsymbol{\phi}(\boldsymbol{x})$ and $\boldsymbol{c}^T\boldsymbol{\phi}(\boldsymbol{x})$ actually correspond to quadratic forms in $\boldsymbol{x}$:

$$\begin{aligned} \boldsymbol{c}_\Theta^T\boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{v}^T\boldsymbol{h} &= \sum_{i=1}^{K} v_i(\boldsymbol{A}_\Theta\boldsymbol{\phi}(\boldsymbol{x}))_i^2 = \sum_{i=1}^{K} v_i(\boldsymbol{w}_i^T\boldsymbol{x})^2 \\ &= \begin{bmatrix} \boldsymbol{w}_1^T\boldsymbol{x} & \boldsymbol{w}_2^T\boldsymbol{x} & \dots & \boldsymbol{w}_K^T\boldsymbol{x} \end{bmatrix} \operatorname{diag}(\boldsymbol{v}) \begin{bmatrix} \boldsymbol{w}_1^T\boldsymbol{x} \\ \boldsymbol{w}_2^T\boldsymbol{x} \\ \dots \\ \boldsymbol{w}_K^T\boldsymbol{x} \end{bmatrix} \\ &= (\boldsymbol{W}\boldsymbol{x})^T\operatorname{diag}(\boldsymbol{v})\boldsymbol{W}\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{W}^T\operatorname{diag}(\boldsymbol{v})\boldsymbol{W}\boldsymbol{x} \end{aligned}$$

and

$$\begin{aligned} \boldsymbol{c}^T\boldsymbol{\phi}(\boldsymbol{x}) &= \sum_{i=1}^{\frac{D(D+1)}{2}} c_i\phi_i(x) = \sum_{i=1}^{D} c_{(i-1)D+i-\frac{(i-1)i}{2}}\ x_i^2 + \sum_{i=1}^{D}\sum_{j=i+1}^{D} c_{(i-1)D+j-\frac{(j-1)j}{2}}\ (2x_ix_j) \\ &= \boldsymbol{x}^T\mathcal{M}(\boldsymbol{c})\boldsymbol{x} \end{aligned}$$

where $\mathcal{M}(\boldsymbol{c}) \in \mathbb{R}^{D\times D}$ is a symmetric matrix obatined from $\boldsymbol{c}$ such that:

– the diagonal and the part above the diagonal of the matrix $\mathcal{M}(\boldsymbol{c})$ is filled row-wise with the elements of vector $\boldsymbol{c}$, i.e.:

$$\mathcal{M}(\boldsymbol{c}) = \begin{bmatrix} c_1 & c_2 & c_3 & \cdots & c_D \\ c_2 & c_{D+1} & c_{D+2} & \cdots & c_{2D-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_D & c_{D-1} & c_{D-2} & \cdots & c_{\frac{D(D+1)}{2}} \end{bmatrix}$$

– for $1 \le i \le D$ and $j < i$, $(\mathcal{M}(\boldsymbol{c}))_{ij} = (\mathcal{M}(\boldsymbol{c}))_{ji}$

Furthermore, we also recur to the following lemma:

Lemma 1. *Two vectors* $\boldsymbol{a}$, $\boldsymbol{b} \in \mathbb{R}^{\frac{D(D+1)}{2}}$ *are equal if and only if* $\boldsymbol{a}^T\boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{b}^T\boldsymbol{\phi}(\boldsymbol{x})$, *for all* $\boldsymbol{x} \in \mathbb{R}^D$

Proof. If $\boldsymbol{a} = \boldsymbol{b}$, then $\boldsymbol{a}^T\boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{b}^T\boldsymbol{\phi}(\boldsymbol{x})$ is trivially verified. For the reverse implication, note that, according to the previously made observation, for any $\boldsymbol{x} \in \mathbb{R}^D$:

$$\boldsymbol{a}^T\boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{b}^T\boldsymbol{\phi}(\boldsymbol{x}) \Rightarrow \boldsymbol{x}^T\mathcal{M}(\boldsymbol{a})\boldsymbol{x} = \boldsymbol{x}^T\mathcal{M}(\boldsymbol{b})\boldsymbol{x}$$

Since both $\mathcal{M}(\boldsymbol{a})$ and $\mathcal{M}(\boldsymbol{a})$ are symmetric and the associated quadratic forms are twice continuously differentiable, taking the hessian on both sides of the equation yields:

$$\mathcal{M}(\boldsymbol{a}) = \mathcal{M}(\boldsymbol{b})$$

that is, $\mathcal{M}(\boldsymbol{a})$ and $\mathcal{M}(\boldsymbol{b})$ are equal entry-wise. In particular, we have that $a_i = b_i$, for $i = 1, \ldots, \frac{D(D+1)}{2}$. $\square$

We are now equipped with the tools needed for the proof. Since $\mathcal{M}(\boldsymbol{c})$ is symmetric, the **Spectral Decomposition Theorem** tells us that there is an orthonormal matrix $\boldsymbol{Q}$ and a diagonal matrix $\boldsymbol{\Lambda}$ such that $\mathcal{M}(\boldsymbol{c}) = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T$. Let $\boldsymbol{q_i}$ denote the eigenvector of $\mathcal{M}(\boldsymbol{c})$ that is present in i-th column of $\boldsymbol{Q}$ and let λ_i be the corresponding eigenvector (note that $\{\boldsymbol{q}_i\}_{i=1}^D$ forms an orthonormal basis of $\mathbb{R}^D$). Then, we can write $\boldsymbol{c}^T\boldsymbol{\phi}(\boldsymbol{x})$ as:

$$\boldsymbol{c}^T\boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{x}^T\mathcal{M}(\boldsymbol{c})\boldsymbol{x} = (\boldsymbol{Q}^T\boldsymbol{x})^T\boldsymbol{\Lambda}(\boldsymbol{Q}^T\boldsymbol{x}) = \sum_{i=1}^{D} \lambda_i(\boldsymbol{q}_i^T\boldsymbol{x})^2$$

Now, if we assume that $K \ge D$, we can find a matrix $\boldsymbol{W}$ and a vector $\boldsymbol{v}$ that make $\boldsymbol{c}_\Theta^T\boldsymbol{\phi}(\boldsymbol{x})$ equal to $\boldsymbol{c}^T\boldsymbol{\phi}(\boldsymbol{x})$ in the following way:

– we make $\boldsymbol{v}$ to be equal to $(\lambda_1, \lambda_2, \ldots, \lambda_D, \underbrace{0, \ldots, 0}_{K-D \text{ times}})$;

– we make $\boldsymbol{W}$ to be equal to the vertical concatenation of $\boldsymbol{Q}^T$ with a $(K-D) \times D$ matrix of zeros, i.e.:

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{Q}^T \\ \boldsymbol{0}_{(K-D)\times D} \end{bmatrix}$$

We then have that:

$$\boldsymbol{c}_\Theta^T\boldsymbol{\phi}(\boldsymbol{x}) = \sum_{i=1}^{K} \boldsymbol{v}_i(\boldsymbol{w}_i^T\boldsymbol{x})^2 = \sum_{i=1}^{D} \lambda_i(\boldsymbol{q}_i^T\boldsymbol{x})^2 = (\boldsymbol{Q}^T\boldsymbol{x})^T\boldsymbol{\Lambda}(\boldsymbol{Q}^T\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T\boldsymbol{x} = \boldsymbol{c}^T\boldsymbol{\phi}(\boldsymbol{x})$$

and, by Lemma 1, we prove that the previous choice of $\boldsymbol{W}$ and $\boldsymbol{v}$ originate a vector $\boldsymbol{c}_\Theta$ such that $\boldsymbol{c}_\Theta = \boldsymbol{c}$. Furthermore, we have proven that the sets of classifiers $\mathcal{C}_1 = \{\boldsymbol{c}_\Theta^T\phi(x) : \Theta = (\boldsymbol{W}, \boldsymbol{v}) \in \mathbb{R}^{K\times D\times K}\}$ and $\mathcal{C}_2 = \{\boldsymbol{c}^T\phi(x) : \boldsymbol{c} \in \mathbb{R}^{\frac{D(D+1)}{2}}\}$ are exactly the same, and so the original neural network reduces to a **linear model** in terms of $\boldsymbol{c}_\Theta$. $\square$

In fact, we can relax the requirement $K \geq D$ to be $K \geq D - N$, where N is the dimension of the nullspace of $\mathcal{M}(\boldsymbol{c})$. Since $\mathcal{M}(\boldsymbol{c})$ is symmetric and thus diagonalizable, the set of indices $\mathcal{I}$ such that the eigenvectors $\{\boldsymbol{q}_i : i \in \mathcal{I}\}$ are not associated with eigenvalue zero has $D - N$ elements. As such:

$$\boldsymbol{c}^T \boldsymbol{\phi}(\boldsymbol{x}) = \sum_{i=1}^{D} \lambda_i (\boldsymbol{q}_i^T \boldsymbol{x})^2 = \sum_{i \in \mathcal{I}} \lambda_i (\boldsymbol{q}_i^T \boldsymbol{x})^2$$

and thus we can take $\boldsymbol{W}$ to be equal to the concatenation of the matrix $\tilde{\boldsymbol{Q}}^T$ with a matrix of $(K-(D-N)) \times D$ zeros (where $\tilde{\boldsymbol{Q}} \in \mathbb{R}^{D \times (D-N)}$ and each column vector $\tilde{\boldsymbol{q}}_k$ is equal to $\boldsymbol{q}_{i_k}$, where $1 \leq k \leq D - N$ and i_k is the k-th index present in $\mathcal{I}$). Following the previously made argument, we would obtain again a vector $\boldsymbol{c}_\Theta$ equal to $\boldsymbol{c}$.

Now, if $K < D - N$, then the nullspace of $\boldsymbol{W}$ has at least dimension $D - (D - N - 1) = N + 1$. Since the rank of $\mathcal{M}(\boldsymbol{c})$ is $D - N$, the dimensions of the rowspace of $\mathcal{M}(\boldsymbol{c})$ and the nullspace of $\boldsymbol{W}$ sum up to at least $D + 1$ and thus there is some non-null vector $\boldsymbol{x}^*$ that is in the rowspace of $\mathcal{M}(\boldsymbol{c})$ and in the nullspace of $\boldsymbol{W}$. Choosing $\boldsymbol{c}$ to be a vector such that $\mathcal{M}(\boldsymbol{c})$ is positive semidefinite, for example:

$$\mathcal{M}(\boldsymbol{c}) = \operatorname{diag}(\underbrace{1, \ldots, 1}_{D-N\text{times}}, \underbrace{0, \ldots, 0}_{N\text{times}})$$

we have:

$$\boldsymbol{c}_\Theta^T \boldsymbol{\phi}(\boldsymbol{x}^*) = \boldsymbol{x}^{*T} \boldsymbol{W}^T \operatorname{diag}(\boldsymbol{v}) \boldsymbol{W} \boldsymbol{x}^* = \boldsymbol{0}$$

and

$$\boldsymbol{c}^T \boldsymbol{\phi}(\boldsymbol{x}^*) = \sum_{i \in \mathcal{I}} \lambda_i (\boldsymbol{q}_i^T \boldsymbol{x}^*)^2 > 0$$

since $\lambda_i > 0$ (for $i \in \mathcal{I}$), and $(\boldsymbol{q}_i^T \boldsymbol{x}^*)^2 > 0$ for at least one $i \in \mathcal{I}$, as $\boldsymbol{x}^*$ belongs to the rowspace of $\mathcal{M}(\boldsymbol{c})$ (which is spanned by $\{\boldsymbol{q}_i : i \in \mathcal{I}\}$).

We have thus constructed an instance where $K < D$ and there is no choice of parameters $\boldsymbol{W}$ and $\boldsymbol{v}$ that make $\boldsymbol{c}_\Theta^T$ equal to $\boldsymbol{c}$ and so the model cannot be parametrized by $\boldsymbol{c}_\Theta^T$ in this case.

4. Given that $\hat{y} = \boldsymbol{c}_\Theta^T \boldsymbol{\phi}(\boldsymbol{x})$, we can write the squared loss as:

$$L(\boldsymbol{c}_\Theta^T; \mathcal{D}) = \sum_{n=1}^{N} (\hat{y}_n(\boldsymbol{x}_n; \boldsymbol{c}_\Theta^T) - y_n)^2 = \sum_{n=1}^{N} (\boldsymbol{c}_\Theta^T \boldsymbol{\phi}(\boldsymbol{x}) - y_n)^2 = \|\boldsymbol{X}\boldsymbol{c}_\Theta - \boldsymbol{y}\|_2^2$$

where $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$. As such, the minimzation of the squared lost corresponds to a linear least squares problem, which in this case has a unique solution $\boldsymbol{c}_\Theta^*$ that is found simply by setting the gradient to zero (since $\boldsymbol{X}$ has rank N, $\boldsymbol{X}^T\boldsymbol{X}$ has also rank N and so it is invertible):

$$\begin{aligned} \nabla_{\boldsymbol{c}_\Theta} L(\boldsymbol{c}_\Theta^T; \mathcal{D}) = \boldsymbol{0} &\Leftrightarrow \nabla_{\boldsymbol{c}_\Theta}(\boldsymbol{X}\boldsymbol{c}_\Theta) \nabla_{\boldsymbol{z}}(\|\boldsymbol{z}\|_2^2)|_{\boldsymbol{z}=\boldsymbol{X}\boldsymbol{c}_\Theta - \boldsymbol{y}} = \boldsymbol{0} \Leftrightarrow 2\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{c}_\Theta - \boldsymbol{y}) = \boldsymbol{0} \\ &\Rightarrow \boldsymbol{c}_\Theta^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \end{aligned}$$

Usually, loss functions of feedforward neural networks are non-convex in their parameters due to multiple compositions of non-linear activation functions, making the global minimization of said functions especially hard. In spite of that, since the presented architecture only has a single hidden layer with at least as many units as the input layer and uses only quadratic activations, we managed to reduce the underlying model to a **linear regression** by defining the underlying **feature transformation** $\boldsymbol{\phi}$ and **weight vector** $\boldsymbol{c}_\Theta$. Hence, the minimization of the loss function admits a closed form solution.