Christopher Zanoli
Student Number: 23-942-394

Exercise 1

Definition 1 (Non-negative matrix factorization (NMF)). *For an arbitrary real-valued, non-negative $n \times m$ matrix $\mathbf{A}$, NMF finds two non-negative matrices $\mathbf{U} \in \mathbb{R}_{\geq 0}^{n \times k}$, $\mathbf{V} \in \mathbb{R}_{\geq 0}^{m \times k}$, such that:*

$$\mathbf{A} \approx \mathbf{U}\mathbf{V}^T$$

NMF typically optimizes the standard matrix factorization objective previously seen in the lecture $f(\mathbf{U}, \mathbf{V}) = \|\mathbf{A} - \mathbf{U}\mathbf{V}^T\|_F^2$:

$$\min_{\mathbf{U},\mathbf{V} \geq 0} f(\mathbf{U}, \mathbf{V})$$

NMF is commonly employed for imputation (i.e., completing missing values as in our movie rating example) and dimensionality reduction.

Assumption 1 (Completeness). *Throughout the following NMF exercises, we assume that $\mathbf{A}$ is fully observed. This assumption is motivated by applications for which NMF is commonly used, such as gene-expression data. We note that this assumption can be relatively easily relaxed.*

Definition 2 (Frobenius inner product). *For arbitrary real-valued $n \times m$ matrices,*

$$\langle \mathbf{A}, \mathbf{B} \rangle_F = \mathbf{Tr}(\mathbf{B}^T \mathbf{A})$$

Definition 3 (Proximal operator). *Let $g : \mathbb{R}^{n \times k} \to \mathbb{R} \cup \{+\infty\}$ be a convex function. The proximal operator $\boldsymbol{prox}_g : \mathbb{R}^{n \times k} \to \mathbb{R}^{n \times k}$ on g is defined by*

$$\boldsymbol{prox}_g(\mathbf{V}) = \arg\min_{\mathbf{X}} g(\mathbf{X}) + \frac{1}{2}\|\mathbf{X} - \mathbf{V}\|_F^2$$

Note: *We disregard other requirements on g for the existence of **prox** here for the sake of exposition since all functions you will encounter fulfill them.*

Definition 4 (Proximal Gradient Descent (PGD))**.** *Consider the following optimization problem*

$$\min\left(f(\mathbf{V}) + g(\mathbf{V})\right)$$

where $f : \mathbb{R}^{n\times k} \to \mathbb{R}$ *and* $g : \mathbb{R}^{n\times k} \to \mathbb{R} \cup \{+\infty\}$ *are convex and only* f *is required to be differentiable.*

This problem can be optimized using PGD, which splits the objective into two parts, first applying a gradient step along f*, followed by applying the proximal operator* $\boldsymbol{prox}_g$ *of* g *(we ignore step size considerations here for simplicity).*

$$\mathbf{X}^{k+1} = \boldsymbol{prox}_g(\mathbf{X}^k - \nabla f(\mathbf{X}^k))$$

In practice, PGD is used prominently to optimize non-differentiable regularizers such as the ℓ_1 *norm in the Lasso. For our purposes, we will apply PGD to maintain the non-negativity constraints of* $\mathbf{U}, \mathbf{V}$ *while optimizing the NMF objective.*

1. Consider the following three datasets and briefly comment: What are the observed values in each dataset and do the observed values fulfill the assumptions of NMF regarding the domain of $\mathbf{A}$?

   Example for our running example movie rating data: The observed data are numeric ratings for each movie m by each user u. Since ratings are typically 1-10 or similar, they are non-negative and thus fulfill the assumptions of NMF on $\mathbf{A}$.

   (a) Item sales data: For the year 2023, we observed how many times each item i was sold in each Coop store c in Zürich.

   (b) Image data: We observe i **grayscale** images with their corresponding image representation which we assume to have 784 pixels.

   (c) Relative student performance in five exams at ETH: We assume that we followed s students at ETH who all took the same e exams (that are scored between 0 and 100) during one exam session. We calculate their relative scores for each exam as their score from 0-100 minus the mean score of that exam.

1 pts ☐

- (a): The observed data are numeric, i.e. the number of times each item i has been sold by coop c. Since an item cannot be sold for a negative amount, the minimum is 0. Hence they are non-negative and thus fulfill the assumptions of NMF on A.
- (b) The observed data are i matrices presumably $28x28$ since we assume each image has 784 pixels. Since grey-scale images' pixels are typically in the $0-255$ range (where 0 is associated with `black` and 255 with `white`), then they are non-negative and thus fulfill the assumptions of NMF on A.

- (c): For each exam e_i and student s_i the observed data are numeric. However the kind of computation performed on those observed data, makes the outcome unfulfilling the NMF assumption on A. Suppose for a given exam e_i our s students ($s = 8$) received the following grades= $[1, 40, 80, 99, 99, 99, 99, 99]$. The average grade for this exam is 77. If we compute the relative scores as defined in the request we obtain: $[-76, -37, 3, 22, 22, 22, 22, 22]$. Hence, our data can be negative and thus don't fulfill the assumption of NMF on A. In general, if the achieved grade by the student s_i on a given exam e_i is less than the average grade across all students for that given exam, the computed relative score will be negative.

2. Show that for an arbitrary real-valued $n \times m$ matrix $\mathbf{A}$, the Frobenius inner product induces the Frobenius norm, that is $\langle \mathbf{A}, \mathbf{A} \rangle_F = \|\mathbf{A}\|_F^2$.

1 pts ☐

To show that $\langle \mathbf{A}, \mathbf{A} \rangle_F = \|\mathbf{A}\|_F^2$ we can first employ the definition (2):

$$\langle \mathbf{A}, \mathbf{B} \rangle_F = \mathrm{Tr}(\mathbf{B^T A}) \Rightarrow \langle \mathbf{A}, \mathbf{A} \rangle_F = \mathrm{Tr}(\mathbf{A^T A})$$

Now, from the definition of Frobenius Norm: $\|\mathbf{A}\|_F = \sqrt{\mathrm{Tr}(\mathbf{A^T A})}$
If we square both sides we obtain:

$$\|\mathbf{A}\|_F^2 = \mathrm{Tr}(\mathbf{A^T A})$$

Now, since $\|\mathbf{A}\|_F^2 = \mathrm{Tr}(\mathbf{A^T A})$ and $\|\mathbf{A}\|_F = \sqrt{\mathrm{Tr}(\mathbf{A^T A})}$

$$\Rightarrow \|\mathbf{A}\|_F^2 = \mathrm{Tr}(\mathbf{A^T A})$$

□

3. Using only matrix notation, derive the partial gradients $\nabla_U f(\mathbf{U}, \mathbf{V})$ and $\nabla_V f(\mathbf{U}, \mathbf{V})$.

1.5 pts ☐

$$\begin{aligned}
\nabla_{\mathbf{U}} f(\mathbf{U}, \mathbf{V}) &= \nabla_{\mathbf{U}} \|\mathbf{A} - \mathbf{U}\mathbf{V}^T\|_F^2 \\
&= \nabla_{\mathbf{U}} \mathrm{Tr}((\mathbf{A} - \mathbf{U}\mathbf{V}^T)^T (\mathbf{A} - \mathbf{U}\mathbf{V}^T)) && (1) \\
&= \nabla_{\mathbf{U}} \mathrm{Tr}((\mathbf{A^T} - \mathbf{VU^T})(\mathbf{A} - \mathbf{U}\mathbf{V}^T)) && (2) \\
&= \nabla_{\mathbf{U}} \mathrm{Tr}((\mathbf{A^T A} - \mathbf{A^T U V^T} - \mathbf{V U^T A} + \mathbf{V U^T U V^T}))) && (3) \\
&= \nabla_{\mathbf{U}} (\mathrm{Tr}(\mathbf{A^T A}) - \mathrm{Tr}(\mathbf{A^T U V^T}) - \mathrm{Tr}(\mathbf{V U^T A}) + \mathrm{Tr}(\mathbf{V U^T U V^T})) && (4) \\
&= \nabla_{\mathbf{U}} \mathrm{Tr}(\mathbf{A^T A}) - \nabla_{\mathbf{U}} \mathrm{Tr}(\mathbf{A^T U V^T}) - \nabla_{\mathbf{U}} \mathrm{Tr}(\mathbf{V U^T A}) + \nabla_{\mathbf{U}} \mathrm{Tr}(\mathbf{V U^T U V^T}) && (5)
\end{aligned}$$

(1) From Exercise 1
(2) From properties of transposed matrices
(3) From the development of matrix products

(4) From linearity of the trace operator
(5) From vector calculus property: $\partial(\mathbf{X}+\mathbf{Y}) = \partial\mathbf{X} + \partial\mathbf{Y}$

Now we can compute the single terms:

$\nabla_{\mathbf{U}}\mathrm{Tr}(\mathbf{A^T A}) = 0$ since $\mathbf{A^T A}$ is a constant with respect to $\mathbf{U}$

$\nabla_{\mathbf{U}}\mathrm{Tr}(\mathbf{A^T U V^T}) = \mathbf{AV}$ since $\frac{\partial}{\partial \mathbf{X}}\mathrm{Tr}(\mathbf{AXB}) = \mathbf{A^T B^T}$

$\nabla_{\mathbf{U}}\mathrm{Tr}(\mathbf{V U^T A}) = \mathbf{AV}$ since $\frac{\partial}{\partial \mathbf{X}}\mathrm{Tr}(\mathbf{AX^T B}) = \mathbf{BA}$

$\nabla_{\mathbf{U}}\mathrm{Tr}(\mathbf{V U^T U V^T}) = \mathbf{2UV^T V}$ for the following reasons:

First, we employ the associative property of matrix multiplication and the cyclic property of the trace operator:

$$\mathrm{Tr}(\mathbf{VU^TUV^T}) = \mathrm{Tr}(\mathbf{VU^T}(\mathbf{UV^T})) = \mathrm{Tr}(\mathbf{UV^TVU^T}) = \mathrm{Tr}(\mathbf{U}(\mathbf{V^TV})\mathbf{U^T})$$

Next, we employ:

$$\frac{\partial}{\partial \mathbf{X}}\mathrm{Tr}(\mathbf{XBX^T}) = \mathbf{XB^T} + \mathbf{XB}$$

Putting all together:

$$\nabla_{\mathbf{U}} f(\mathbf{U},\mathbf{V}) = -\mathbf{AV} - \mathbf{AV} + 2\mathbf{UV^TV} = -2\mathbf{AV} + 2\mathbf{UV^TV}$$

$$\begin{aligned}
\nabla_{\mathbf{V}} f(\mathbf{U},\mathbf{V}) &= \nabla_{\mathbf{U}}\|\mathbf{A} - \mathbf{UV}^T\|_F^2 \\
&= \nabla_{\mathbf{V}}\mathrm{Tr}((\mathbf{A} - \mathbf{UV}^T)^T(\mathbf{A} - \mathbf{UV}^T)) && (1)\\
&= \nabla_{\mathbf{V}}\mathrm{Tr}((\mathbf{A^T} - \mathbf{VU^T})(\mathbf{A} - \mathbf{UV}^T)) && (2)\\
&= \nabla_{\mathbf{V}}\mathrm{Tr}((\mathbf{A^TA} - \mathbf{A^TUV^T} - \mathbf{VU^TA} + \mathbf{VU^TUV^T}))) && (3)\\
&= \nabla_{\mathbf{V}}(\mathrm{Tr}(\mathbf{A^TA}) - \mathrm{Tr}(\mathbf{A^TUV^T}) - \mathrm{Tr}(\mathbf{VU^TA}) + \mathrm{Tr}(\mathbf{VU^TUV^T})) && (4)\\
&= \nabla_{\mathbf{V}}\mathrm{Tr}(\mathbf{A^TA}) - \nabla_{\mathbf{V}}\mathrm{Tr}(\mathbf{A^TUV^T}) - \nabla_{\mathbf{V}}\mathrm{Tr}(\mathbf{VU^TA}) + \nabla_{\mathbf{V}}\mathrm{Tr}(\mathbf{VU^TUV^T}) && (5)
\end{aligned}$$

(1) From Exercise 1
(2) From properties of transposed matrices
(3) From the development of matrix products
(4) From linearity of the trace operator
(5) From vector calculus property: $\partial(\mathbf{X}+\mathbf{Y}) = \partial\mathbf{X} + \partial\mathbf{Y}$

Now we can compute the single terms:

$\nabla_{\mathbf{V}}\mathrm{Tr}(\mathbf{A^T A}) = 0$ since $\mathbf{A^T A}$ is a constant with respect to $\mathbf{V}$

$\nabla_{\mathbf{V}} \mathrm{Tr}(\mathbf{A^T U V^T}) = \mathbf{A^T U}$ for the following reasons:

First we employ the cyclic property of the trace operator:

$$\mathrm{Tr}(\mathbf{A^T U V^T}) = \mathrm{Tr}(\mathbf{U V^T A^T})$$

Next, we employ:

$$\frac{\partial}{\partial \mathbf{X}} \mathrm{Tr}(\mathbf{A X^T B}) = \mathbf{BA}$$

$\nabla_{\mathbf{V}} \mathrm{Tr}(\mathbf{V U^T A}) = A^T U$ for the following reasons:

First, we employ the cyclic property of the trace operator:

$$\frac{\partial}{\partial \mathbf{X}} \mathrm{Tr}(\mathbf{V U^T A}) = \mathrm{Tr}(\mathbf{A V U^T})$$

Next, we employ:

$$\frac{\partial}{\partial \mathbf{X}} \mathrm{Tr}(\mathbf{A X B}) = \mathbf{A^T B^T}$$

$\nabla_{\mathbf{V}} \mathrm{Tr}(\mathbf{V U^T U V^T}) = 2\mathbf{V U^T U}$ for the following reasons:

First, we employ the associativity property of matrix multiplication:

$$\mathrm{Tr}(\mathbf{V U^T U V^T}) = \mathrm{Tr}(\mathbf{V}(\mathbf{U^T U})\mathbf{V^T})$$

Next, we employ:

$$\frac{\partial}{\partial \mathbf{X}} \mathrm{Tr}(\mathbf{X B X^T}) = \mathbf{X B^T} + \mathbf{X B}$$

Putting all together:

$$\nabla_{\mathbf{V}} f(\mathbf{U}, \mathbf{V}) = -\mathbf{A^T U} - \mathbf{A^T U} + 2\mathbf{V U^T U} = -2\mathbf{A^T U} + 2\mathbf{V U^T U}$$

$\square$

4. Rewrite the NMF objective to include the constraints $\mathbf{U}, \mathbf{V} \geq 0$ to derive a functional form that is optimizable using PGD: $f^*(\mathbf{U}, \mathbf{V}) = f(\mathbf{U}, \mathbf{V}) + g(\mathbf{U}) + g(\mathbf{V})$. Derive the proximal operator $\mathbf{prox}_g(\mathbf{U})$ of g.

   Hint: Use the indicator function. Note that $\mathbf{U}, \mathbf{V} \geq 0$ is exactly enforcing each matrix element u_{ij}, v_{kj} to be non-negative.

1.5 pts $\square$

Recall that the NMF objective is the following:

$$\min_{\mathbf{U},\mathbf{V}\geq 0} f(\mathbf{U},\mathbf{V}) = \min_{\mathbf{U},\mathbf{V}\geq 0} \|\mathbf{A} - \mathbf{U}\mathbf{V}^T\|_F^2$$

First we can define the $g(\mathbf{X})$ function as:

$$g(\mathbf{X}) = \begin{cases} 0 & \text{if } \mathbf{X} \geq 0 \\ \infty & \text{otherwise} \end{cases}$$

This means that if the entry of $\mathbf{X}$ is negative then $g(\mathbf{X})$ assigns ∞, 0 otherwise.

Now, the proximal operator is defined as:

$$\mathbf{prox}_g(\mathbf{V}) = \arg\min_{\mathbf{X}} g(\mathbf{X}) + \frac{1}{2}\|\mathbf{X} - \mathbf{V}\|_F^2$$

Given the choice of $g(\mathbf{X})$, if $\mathbf{X}$ is negative then:

$$\mathbf{prox}_g(\mathbf{V}) = \arg\min_{\mathbf{X}} \infty + \frac{1}{2}\|\mathbf{X} - \mathbf{V}\|_F^2$$

This results in minimizing the following objective function:

$$\mathbf{prox}_g(\mathbf{V}) = \arg\min_{\mathbf{X}\geq 0} \frac{1}{2}\|\mathbf{X} - \mathbf{V}\|_F^2$$

Since the minimizer $\mathbf{U}^*$ has to comply with the non-negativity constraint of all elements of $\mathbf{U}$, then $\mathbf{U}^*$ can be expressed as follows: $\mathbf{U}^*_{i,j} = \mathbf{U} \cdot \mathbb{1}\{U_{i,j} \geq 0\}$

This means that, if the (i,j) entry of $\mathbf{U}$ is ≥ 0 then the (i,j) entry of $\mathbf{U}$ is kept, otherwise, if the (i,j) entry of $\mathbf{U}$ is < 0 then the (i,j) entry of $\mathbf{U}$ is set to 0, perfectly complying with the aforementioned constraint.

Exercise 2

You are given a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ and its singular value decomposition $\mathbf{A} = \mathbf{U\Sigma V}^{\mathrm{T}}$.

Hint: $\mathbf{A}$ does not have to be diagonalizable.

1. Show that $\sigma_1 \geq |\lambda|_{\max}$, i.e. show that the largest singular value dominates all eigenvalues of the matrix $\mathbf{A}$.

2 pts ☐

We can start the proof by stating the following property of orthogonal matrices:

$$\|\mathbf{Qx}\| = \|\mathbf{x}\|$$

I.e., multiplying a vector by an orthogonal matrix doesn't change the norm of the vector itself.

Hence, for $\mathbf{A} = \mathbf{U\Sigma V}^{\mathrm{T}}$:

$$\|\mathbf{Ax}\| = \|\mathbf{U\Sigma V}^{\mathrm{T}}\mathbf{X}\| = \|\mathbf{\Sigma V}^{\mathrm{T}}\mathbf{X}\|$$

Where the second equality is true since $\mathbf{U}$ is orthogonal and thus preserves lengths.

Now, let $\mathbf{y} = \mathbf{V^T x}$, then:

$$\|\mathbf{Ax}\| = \|\mathbf{\Sigma y}\|$$

We need to show now that $\|\mathbf{\Sigma y}\|$ is upper bounded by $\sigma_1\|\mathbf{x}\|$

We can express $\|\mathbf{\Sigma y}\|$ as:

$$\sum_{i=1}^{n}(\sigma_i y_i)^2$$

Now, since σ_1 is the largest singular value of $\mathbf{A}$ we have that $\sigma_1 \geq \sigma_i \ \forall i$

$$\Rightarrow \|\mathbf{\Sigma V}^{\mathrm{T}}\mathbf{X}\| \leq \sum_{i=1}^{n}(\sigma_i y_i)^2$$

Now we can recall that $\mathbf{y} = \mathbf{V^T x} \Rightarrow \|\mathbf{y}\| = \|\mathbf{V^T x}\| = \|\mathbf{A}\|$ since $\mathbf{V}$ is orthogonal.

$$\Rightarrow \|\mathbf{\Sigma y}\| \leq \sum_{i=1}^{n}(\sigma_i x_i)^2 = \sigma_1 \sum_{i=1}^{n} x_i^2 = \sigma_1\|\mathbf{x}\|$$

Since the inequality: $\|\mathbf{\Sigma y}\| \leq \sigma_1\|\mathbf{x}\|$ holds for $\forall$ eigenvector $\mathbf{x}$ with eigenvalue λ we can write:

$$\|\mathbf{Ax}\| = |\lambda|\|\mathbf{x}\|$$

This because for eigenvector$\mathbf{x}$ with eigenvalue λ we have:

$\mathbf{AX} = \lambda\mathbf{x}$ and $\|\lambda\mathbf{x}\| = |\lambda|\|\mathbf{x}\|$

Now, since $\|\mathbf{Ax}\| \leq \|\mathbf{\Sigma y}\|$, $\|\mathbf{\Sigma y}\| \leq \sigma_1\|\mathbf{x}\|$ and $|\lambda|\|\mathbf{x}\| \leq \sigma_1\|\mathbf{x}\| \Rightarrow$

$$|\lambda| \leq \sigma_1$$

This implies that for any eigenvalue λ its absolute value is less than or equal to $\sigma_1 \Rightarrow \sigma_1 \geq |\lambda|_{\max}$

□

2. Show that every entry of $\mathbf{A}$ must satisfy $|a_{ij}| \leq \sigma_1$

1 pts □

We know that multiplying a vector by an orthonormal matrix doesn't change the norm of the vector itself.

This means that: $\|\mathbf{U\Sigma V}^T\| = \|\mathbf{\Sigma V}^T\|$ since $\mathbf{U}$ is an orthogonal matrix.

We know that σ_1 is the largest singular value of $\mathbf{A}$, and in point 2.1 of Exercise 2 we derived the following inequality:

$$\|\mathbf{Ax}\| \leq \sigma_1\|\mathbf{x}\|$$

Hence we can write:

$$\|\mathbf{Ax}\| = \|\mathbf{\Sigma V}^T\mathbf{x}\| \leq \sigma_1\|\mathbf{x}\|$$

Now, let $\mathbf{x}$ be the unit vector $\mathbf{x} = (1, 0, 0, ..., 0)$ then $\mathbf{Ax}$ is the first column of $\mathbf{A}$.

Consequently, $\|\mathbf{Ax}\| \leq \sigma_1\|\mathbf{x}\|$ means that the length of the first column of $\mathbf{A}$ is bounded by $\sigma_1 \Rightarrow$ every entry in the first column of $\mathbf{A}$ must also be bounded by σ_1. This because the length of a vector (column in this case) is determined by the magnitudes of its individual entries: $\|\mathbf{c}\| = \sqrt{\sum_i c_i^2}$.

By generalizing this to all columns of $\mathbf{A}$, we obtain that $\|a_{i,j}\| \leq \sigma_1$

□

Exercise 3

You are given the following matrix $\mathbf{A} \in \mathbb{R}^{3\times3}$ for which you know that $\text{rank}(\mathbf{A}) = 2$. Can you find the values of $x_1, x_2 \in \mathbb{R}$ so that we can reconstruct it exactly? Explain your answer and provide values for x_1 and x_2 if reconstruction is possible.

$$\mathbf{A} = \begin{bmatrix} 4 & 6 & 1 \\ 2 & x_2 & 1 \\ x_1 & 3 & 2 \end{bmatrix}$$

2 pts □

A rank k square matrix $\mathbf{A}$ of dimension $n \times n$ is not reconstructable if the number of observed matrix entries $S < 2nk - k^2$.

In our case, $n = 3$ and $k = 2 \Rightarrow$ the inequality is not satisfied:

$$7 \overset{?}{<} 2 \cdot 3 \cdot 2 - 2^2$$
$$7 < 8$$

Hence, we can say for sure that the given matrix cannot be reconstructed.