Exercises
Computational Intelligence Lab
SS 2018

Machine Learning Institute
Dept. of Computer Science, ETH Zürich
Prof. Dr. Thomas Hofmann
Web http://cil.inf.ethz.ch/

Series 4 Solutions
(Matrix Approximation & Reconstruction)

Solution 1 (Alternating Least Squares for Collaborative Filtering):

1. Is the objective function $L(\mathbf{U}, \mathbf{V})$ convex? If not, prove it.

   The objective is not convex. To prove that, it is sufficient to provide a counter example for $m = n = 1$. This counter example can be generalized to other dimensions by setting all the entries in $\mathbf{U}$ and $\mathbf{V}$ to zero except for those with indexes $(1, 1)$:

$$\mathbf{U} = \begin{bmatrix} u & 0 & \dots \\ 0 & 0 & \dots \\ \vdots & \vdots & \ddots \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} v & 0 & \dots \\ 0 & 0 & \dots \\ \vdots & \vdots & \ddots \end{bmatrix}.$$

   If $n = m = 1$, the objective can be written as follows:

$$L(u, v) = (a - uv)^2 + \lambda u^2 + \lambda v^2.$$

   We are going to use the following theorem: a twice differentiable function is convex on a convex set if and only if its Hessian is positive semidefinite on the interior of that convex set.

   One can easily verify that the objective $L(u, v)$ is twice differentiable and its Hessian is

$$\nabla^2 L(u, v) = 2 \begin{bmatrix} v^2 + \lambda & 2uv - a \\ 2uv - a & u^2 + \lambda \end{bmatrix}. \tag{1}$$

   By setting $u = v = \sqrt{2\lambda + 2|a|}$, we can find that

$$\begin{aligned} \det(\nabla^2 L(u, v)) &= 4(v^2 + \lambda)(u^2 + \lambda) - 4(2uv - a)^2 \\ &= 4\Big[(3\lambda + 2|a|)^2 - (4\lambda + \underbrace{4|a| - a}_{>2|a|})^2\Big] < 0. \end{aligned}$$

   Thus, the Hessian (1) is not positive semidefinite everywhere in $\mathbb{R}^2$ by Sylvester's criterion, and hence $L(u, v)$ is not convex in $\mathbb{R}^2$.

2. Is the objective $L(\mathbf{U}, \mathbf{V})$ convex with respect to $\mathbf{U}$?

   Yes.

3. Derive the update rule for $\mathbf{u}_i$. Note that the update rule for $\mathbf{v}_j$ is symmetric to that for $\mathbf{u}_i$.

   Hint: differentiate the objective with respect to $\mathbf{u}_i$ holding $\mathbf{V}$ constant and set the gradient to zero.

$$\frac{\partial L(\mathbf{U}, \mathbf{V})}{\partial \mathbf{u}_i} = -2 \sum_{j:(i,j) \in \mathcal{I}} (a_{ij} - \mathbf{u}_i^\top \mathbf{v}_j)\mathbf{v}_j + 2\lambda \mathbf{u}_i \tag{2}$$

Setting it to zero, one can get

$$\sum_{j:(i,j)\in\mathcal{I}} a_{ij}\mathbf{v}_j = \sum_{j:(i,j)\in\mathcal{I}} (\mathbf{u}_i^\top \mathbf{v}_j)\mathbf{v}_j + \lambda \mathbf{u}_i \tag{3}$$

$$= \sum_{j:(i,j)\in\mathcal{I}} \mathbf{v}_j(\mathbf{u}_i^\top \mathbf{v}_j) + \lambda \mathbf{u}_i \tag{4}$$

$$= \sum_{j:(i,j)\in\mathcal{I}} \mathbf{v}_j(\mathbf{v}_j^\top \mathbf{u}_i) + \lambda \mathbf{u}_i \tag{5}$$

$$= (\sum_{j:(i,j)\in\mathcal{I}} \mathbf{v}_j\mathbf{v}_j^\top)\mathbf{u}_i + \lambda \mathbf{u}_i \tag{6}$$

$$= (\sum_{j:(i,j)\in\mathcal{I}} \mathbf{v}_j\mathbf{v}_j^\top + \lambda \mathbf{I}_k)\mathbf{u}_i. \tag{7}$$

Then it arrives that

$$\mathbf{u}_i = (\sum_{j:(i,j)\in\mathcal{I}} \mathbf{v}_j\mathbf{v}_j^\top + \lambda \mathbf{I}_k)^{-1} \sum_{j:(i,j)\in\mathcal{I}} a_{ij}\mathbf{v}_j. \tag{8}$$

4. Suppose the computational complexity of inverting a $k \times k$ matrix is $O(k^3)$, let n_i be the number of items rated by user i. Find the computational complexity of the step

   $$\mathbf{u}_i = (\sum_{j:(i,j)\in\mathcal{I}} \mathbf{v}_j\mathbf{v}_j^\top + \lambda \mathbf{I}_k)^{-1} \sum_{j:(i,j)\in\mathcal{I}} a_{ij}\mathbf{v}_j$$

   in the ALS algorithm above. Use big O notation.

   $O(n_ik^2 + k^3)$.

5. For a recommender system, $\mathbf{u}_i$ and $\mathbf{v}_j$ can be interpreted as the low-dimensional representations of the user i and the item j correspondingly. Interpret the update steps of the ALS algorithm in terms of obtaining low-dimensional representations for a recommender system.

   The updates of $\mathbf{u}_i$ can be interpreted as follows: given low-dimensional representations of the items, compute independently the best representation of each user, whereas for the updates of $\mathbf{v}_j$, we have: given low-dimensional representations of the users, compute independently the best representation of each item.

Solution 2 (Convex Relaxation for Exact Matrix Recovery):

Let us consider the singular vector decomposition of matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$,

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^\top, \tag{9}$$

where matrices $\mathbf{U} \in \mathbb{R}^{m\times m}$ and $\mathbf{V} \in \mathbb{R}^{n\times n}$ are orthogonal, and $\mathbf{D} \in \mathbb{R}^{m\times n}$ is a diagonal rectangular matrix with non-negative real numbers on its diagonal, which, for instance, for the case $m < n$ can be represented as follows:

$$\mathbf{D} = \begin{bmatrix}\mathbf{\Sigma} & \mathbf{0}\end{bmatrix}, \quad \mathbf{\Sigma} = \text{diag}(\sigma_1, \dots, \sigma_m) = \begin{bmatrix}\sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_m\end{bmatrix}, \quad \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_m \geq 0. \tag{10}$$

1. Since matrices $\mathbf{U}$ and $\mathbf{V}$ are orthogonal, and hence full rank matrices, the rank of the matrix $\mathbf{A}$ is equal to the number of its positive singular values

   $$\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{D}) = \#\{\sigma_i > 0\}. \tag{11}$$

   On the other hand, the Euclidean operator norm of $\mathbf{A}$ is equal to its largest singular value σ_1,

   $$\|\mathbf{A}\|_2 = \sigma_{\max}(\mathbf{A}) = \sigma_1. \tag{12}$$

   Therefore, if $\|\mathbf{A}\|_2 \leq 1$ and hence $\forall i\ \sigma_i \leq 1$, one can derive the following inequality,

   $$\text{rank}(\mathbf{A}) = \#\{\sigma_i > 0\} = \sum_{i:\ \sigma_i>0} 1 \geq \sum_{i:\ \sigma_i>0} \sigma_i = \sum_i \sigma_i = \|\mathbf{A}\|_*. \tag{13}$$

2. A function $f:\ X \to \mathbb{R}$ is called convex if $\forall x, y \in X$

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \quad \forall \lambda \in [0,1]. \tag{14}$$

Let $\mathbf{U}_\lambda \mathbf{D}_\lambda \mathbf{V}_\lambda^\top$ be the SVD decomposition of $\lambda\mathbf{A} + (1-\lambda)\mathbf{B}$. Now we start the proof.

$$\begin{aligned}
\|\lambda\mathbf{A} + (1-\lambda)\mathbf{B}\|_* &= \text{trace}\,(\mathbf{D}_\lambda) && (15)\\
&= \text{trace}\left((\mathbf{U}_\lambda^\top \mathbf{U}_\lambda)\mathbf{D}_\lambda(\mathbf{V}_\lambda^\top \mathbf{V}_\lambda)\right) && (16)\\
&= \text{trace}\left(\mathbf{U}_\lambda^\top (\mathbf{U}_\lambda \mathbf{D}_\lambda \mathbf{V}_\lambda^\top)\mathbf{V}_\lambda\right) && (17)\\
&= \text{trace}\left(\mathbf{U}_\lambda^\top (\lambda\mathbf{A} + (1-\lambda)\mathbf{B})\mathbf{V}_\lambda\right) && (18)\\
&= \lambda\,\text{trace}\left(\mathbf{U}_\lambda^\top \mathbf{A}\mathbf{V}_\lambda\right) + (1-\lambda)\,\text{trace}\left(\mathbf{U}_\lambda^\top \mathbf{B}\mathbf{V}_\lambda\right) && (19)
\end{aligned}$$

Our proof is done once we bound both terms: $\text{trace}\left(\mathbf{U}_\lambda^\top \mathbf{A}\mathbf{V}_\lambda\right) \leq \|\mathbf{A}\|_*$ and $\text{trace}\left(\mathbf{U}_\lambda^\top \mathbf{B}\mathbf{V}_\lambda\right) \leq \|\mathbf{B}\|_*$.
Let $\mathbf{U}_A \mathbf{D}_A \mathbf{V}_A^\top$ be the SVD decomposition of $\mathbf{A}$. Then we get

$$\begin{aligned}
\text{trace}\left(\mathbf{U}_\lambda^\top \mathbf{A}\mathbf{V}_\lambda\right) &= \sum_{i=1}^{\min(m,n)} \left[\mathbf{U}_\lambda^\top \mathbf{A}\mathbf{V}_\lambda\right]_i^i && (20)\\
&= \sum_{i=1}^{\min(m,n)} \left[\mathbf{U}_\lambda^\top \mathbf{U}_A \mathbf{D}_A \mathbf{V}_A^\top \mathbf{V}_\lambda\right]_i^i && (21)\\
&= \sum_{i=1}^{\min(m,n)} \sum_{j=1}^{\min(m,n)} \left[\mathbf{U}_\lambda^\top \mathbf{U}_A\right]_j^i \sigma_j(\mathbf{A}) \left[\mathbf{V}_A^\top \mathbf{V}_\lambda\right]_i^j && (22)\\
&= \sum_{j=1}^{\min(m,n)} \sigma_j(\mathbf{A}) \sum_{i=1}^{\min(m,n)} \left[\mathbf{U}_\lambda^\top \mathbf{U}_A\right]_j^i \left[\mathbf{V}_A^\top \mathbf{V}_\lambda\right]_i^j && (23)\\
&\leq \sum_{j=1}^{\min(m,n)} \sigma_j(\mathbf{A}) \left\| \left[\mathbf{U}_\lambda^\top \mathbf{U}_A\right]_j \right\|_2 \left\| \left[\mathbf{V}_A^\top \mathbf{V}_\lambda\right]^j \right\|_2 && (24)\\
&= \sum_{j=1}^{\min(m,n)} \sigma_j(\mathbf{A}) = \|\mathbf{A}\|_*, && (25)
\end{aligned}$$

where the superscript i above a matrix denotes its i-th row and the subscript i below a matrix denotes its i-th column.
Similarly, one can bound $\text{trace}\left(\mathbf{U}_\lambda^\top \mathbf{B}\mathbf{V}_\lambda\right) \leq \|\mathbf{B}\|_*$, and therefore,

$$\|\lambda\mathbf{A} + (1-\lambda)\mathbf{B}\|_* \leq \lambda \|\mathbf{A}\|_* + (1-\lambda)\|\mathbf{B}\|_*, \tag{26}$$

which finishes the proof.

3. **Note:** This one is a bonus question, it requires lots of knowledge from mathematical optimization, similar questions will not be asked in the exam.
We are going to rewrite the problem

$$\min_{\mathbf{B}} \|\mathbf{B}\|_*, \quad \text{s.t. } \|\mathbf{A} - \mathbf{B}\|_{\mathbf{G}} = 0, \tag{27}$$

as a problem of semidefinite programming (SDP) in the following form,

$$\min_{\mathbf{B}, \mathbf{W}_1, \mathbf{W}_2} \ \tfrac{1}{2}\text{trace}(\mathbf{W}_1) + \tfrac{1}{2}\text{trace}(\mathbf{W}_2), \quad \text{s.t. } \begin{bmatrix} \mathbf{W}_1 & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{W}_2 \end{bmatrix} \succeq 0,\ \|\mathbf{A} - \mathbf{B}\|_{\mathbf{G}} = 0. \tag{28}$$

Note: we assume $m = n$ for simplicity.
We are going to prove the equivalence of (27) and (28), with the help of the Schur complement lemma.
Lemma [1]

$$\begin{bmatrix} \mathbf{W}_1 & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{W}_2 \end{bmatrix} \succeq 0 \qquad \iff \qquad \begin{aligned} &\mathbf{W}_1 \succeq 0, \\ &\mathbf{W}_2 - \mathbf{B}^\top \mathbf{W}_1^+ \mathbf{B} \succeq 0, \\ &(\mathbf{I} - \mathbf{W}_1 \mathbf{W}_1^+)\mathbf{B} = 0, \end{aligned} \tag{29}$$

where $\mathbf{A}^+$ denotes the pseudoinverse of a matrix $\mathbf{A}$, which is a generalization of the inverse matrix defined for any rectangular matrix. The pseudoinverse of a matrix is tightly connected to its SVD decomposition. If $\mathbf{U}\mathbf{D}\mathbf{V}^\top$ is the SVD decomposition of matrix $\mathbf{A}$, Then the pseudoinverse is equal to $\mathbf{A}^+ = \mathbf{V}\mathbf{D}^+\mathbf{U}^\top$.

Using the Schur complement lemma, the SDP problem (28) can be reformulated as follows:

$$\begin{aligned}\min_{\mathbf{B},\mathbf{W}_1,\mathbf{W}_2} \ \tfrac{1}{2}\text{trace}(\mathbf{W}_1) + \tfrac{1}{2}\text{trace}(\mathbf{W}_2), \quad \text{s.t. } & \|\mathbf{A}-\mathbf{B}\|_{\mathbf{G}} = 0,\\ & \mathbf{W}_1 \succeq 0,\\ & \mathbf{W}_2 - \mathbf{B}^\top\mathbf{W}_1^+\mathbf{B} \succeq 0,\\ & (\mathbf{I}-\mathbf{W}_1\mathbf{W}_1^+)\mathbf{B} = 0.\end{aligned} \tag{30}$$

Since matrix $\mathbf{W}_1$ is symmetric positive semidefinite ($\mathbf{W}_1 \succeq 0$), its SVD decomposition can be parametrized by an orthogonal matrix $\mathbf{U}$ and a diagonal positive semidefinite matrix $\mathbf{D} \succeq 0$:

$$\mathbf{W}_1 = \mathbf{U}\mathbf{D}\mathbf{U}^\top \succeq 0. \tag{31}$$

Therefore, the pseudoinverse of $\mathbf{W}_1$ is equal to

$$\mathbf{W}_1^+ = \mathbf{U}\mathbf{D}^+\mathbf{U}^\top \succeq 0. \tag{32}$$

Note that replacing the constraint $\mathbf{W}_2 - \mathbf{B}^\top\mathbf{W}_1^+\mathbf{B} \succeq 0$ with equation $\mathbf{W}_2 = \mathbf{B}^\top\mathbf{W}_1^+\mathbf{B}$ does not affect the solution.

Now, our proof is done once we prove that the problem (27) is equivalent to the following one,

$$\begin{aligned}\min_{\mathbf{B},\mathbf{U},\mathbf{D}} \ \tfrac{1}{2}\text{trace}(\mathbf{U}\mathbf{D}\mathbf{U}^\top) + \tfrac{1}{2}\text{trace}(\mathbf{B}^\top(\mathbf{U}\mathbf{D}^+\mathbf{U}^\top)\mathbf{B}), \quad \text{s.t. } & \|\mathbf{A}-\mathbf{B}\|_{\mathbf{G}} = 0,\\ & \mathbf{D} = \text{diag}(d_1,\dots,d_n) \succeq 0,\\ & \mathbf{U} \text{ orthogonal},\\ & (\mathbf{I}-\mathbf{U}\mathbf{D}\mathbf{D}^+\mathbf{U}^\top)\mathbf{B} = 0.\end{aligned} \tag{33}$$

Let us find minimum over $\mathbf{D}$ holding the other variables constant.

$$\min_{\mathbf{D}} \ \tfrac{1}{2}\text{trace}(\mathbf{U}\mathbf{D}\mathbf{U}^\top) + \tfrac{1}{2}\text{trace}(\mathbf{B}^\top(\mathbf{U}\mathbf{D}^+\mathbf{U}^\top)\mathbf{B}), \quad \text{s.t. } \mathbf{D} = \text{diag}(d_1,\dots,d_n) \succeq 0. \tag{34}$$

$$\begin{aligned}&\tfrac{1}{2}\text{trace}(\mathbf{U}\mathbf{D}\mathbf{U}^\top) + \tfrac{1}{2}\text{trace}(\mathbf{B}^\top(\mathbf{U}\mathbf{D}^+\mathbf{U}^\top)\mathbf{B})\\ &= \tfrac{1}{2}\text{trace}(\mathbf{U}^\top\mathbf{U}\mathbf{D}) + \tfrac{1}{2}\text{trace}((\mathbf{U}\mathbf{D}^+\mathbf{U}^\top)\mathbf{B}\mathbf{B}^\top)\\ &= \tfrac{1}{2}\text{trace}(\mathbf{D}) + \tfrac{1}{2}\text{trace}(\mathbf{D}^+(\mathbf{U}^\top\mathbf{B}\mathbf{B}^\top\mathbf{U}))\\ &= \tfrac{1}{2}\sum_{i:\ d_i>0} d_i + \tfrac{1}{2}\sum_{i:\ d_i>0}\frac{1}{d_i}[\mathbf{U}^\top\mathbf{B}\mathbf{B}^\top\mathbf{U}]_i^i \to \min_{d_1,\dots,d_n\geq 0}.\end{aligned} \tag{35}$$

Here we can find the optimal solution $\mathbf{D}$.

$$\mathbf{D} = \text{diag}(d_1,\dots,d_n), \quad d_i = \sqrt{[\mathbf{U}^\top\mathbf{B}\mathbf{B}^\top\mathbf{U}]_i^i}. \tag{36}$$

Hence for all orthogonal matrices $\mathbf{U}$, the optimal value is

$$\begin{aligned}\tfrac{1}{2}\sum_{i:\ d_i>0} d_i + \tfrac{1}{2}\sum_{i:\ d_i>0}\frac{1}{d_i}[\mathbf{U}^\top\mathbf{B}\mathbf{B}^\top\mathbf{U}]_i^i &= \sum_{i=1}^n \sqrt{[\mathbf{U}^\top\mathbf{B}]^i[\mathbf{B}^\top\mathbf{U}]_i}\\ &= \sum_{i=1}^n \sqrt{[\mathbf{B}^\top\mathbf{U}]_i^\top[\mathbf{B}^\top\mathbf{U}]_i}\\ &= \sum_{i=1}^n \left\|[\mathbf{B}^\top\mathbf{U}]_i\right\|_2\\ &= \sum_{i=1}^n \left\|\mathbf{B}^\top\mathbf{u}_i\right\|_2\\ &\geq \|\mathbf{B}\|_*.\end{aligned} \tag{37}$$

On the other hand, if $\mathbf{B} = \mathbf{U}\mathbf{D}_B\mathbf{V}$ is the SVD decomposition of $\mathbf{B}$, then

$$\tfrac{1}{2}\sum_{i:\, d_i>0} d_i + \tfrac{1}{2}\sum_{i:\, d_i>0} \frac{1}{d_i}[\mathbf{U}^\top\mathbf{B}\mathbf{B}^\top\mathbf{U}]_i^i = \|\mathbf{B}\|_*,$$

Also

$$d_i = \sqrt{[\mathbf{U}^\top\mathbf{B}\mathbf{B}^\top\mathbf{U}]_i^i} = \sqrt{[\mathbf{U}^\top\mathbf{U}\mathbf{D}_B\mathbf{V}^\top\mathbf{V}\mathbf{D}_B\mathbf{U}^\top\mathbf{U}]_i^i} = \sqrt{[\mathbf{D}_B^2]_i^i},$$

and therefore, $\mathbf{D} = \mathbf{D}_B$, which satisfies the constraint $(\mathbf{I} - \mathbf{U}\mathbf{D}\mathbf{D}^+\mathbf{U}^\top)\mathbf{B} = 0$ in the problem (33). Thus, the problem (33) is equivalent to the problem (27).

[1] Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press. `http://stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf`

Solution 3 (SGD for Collaborative Filtering):

Consider the given objective function as a sum

$$f(\mathbf{U},\mathbf{Z}) = \frac{1}{|\Omega|}\sum_{(d,n)\in\Omega} \underbrace{\frac{1}{2}\big[\mathbf{X}_{dn} - (\mathbf{U}\mathbf{Z}^T)_{dn}\big]^2}_{f_{d,n}}$$

where $\mathbf{U} \in \mathbb{R}^{D\times K}$, $\mathbf{Z} \in \mathbb{R}^{N\times K}$.

- **Stochastic Gradient:** For one fixed element (d,n) of the sum, we derive the gradient entry (d',k) of $\mathbf{U}$, that is $\frac{\partial}{\partial u_{d',k}} f_{d,n}(\mathbf{U},\mathbf{Z})$, and analogously for the $\mathbf{Z}$ part.

$$\frac{\partial}{\partial u_{d',k}} f_{d,n}(\mathbf{U},\mathbf{Z}) = \begin{cases} -\big[\mathbf{X}_{dn} - (\mathbf{U}\mathbf{Z}^T)_{dn}\big] z_{n,k} & \text{if } d' = d \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial}{\partial z_{n',k}} f_{d,n}(\mathbf{U},\mathbf{Z}) = \begin{cases} -\big[\mathbf{X}_{dn} - (\mathbf{U}\mathbf{Z}^T)_{dn}\big] u_{d,k} & \text{if } n' = n \\ 0 & \text{otherwise} \end{cases}$$

- **Full Gradient:** We have access to all elements $(d,n) \in \Omega$, so we can calculate the partial derivatives of the full gradient for all $(d,n) \in \Omega$. For one specific $(d,n) \in \Omega$, the partial derivatives are the same as that in the stochastic gradient above.