EPFL | MGT-418 : Convex Optimization | Project 5

Questions – Fall 2019

Recommender Systems (graded)

This project is due on **December 11, 2019, at 23:59**. You may form teams of up to three people. Each team should upload a single zip-file containing their report and Matlab code to Moodle. Make sure to clearly state the team members in your report.

Description

Recommender systems are typically designed to solve matrix completion problems: Given a subset of the elements of a (normally non-square) matrix, the goal is to reconstruct the remaining elements of the matrix. Such problems arise in a number of applications, for instance in the context of streaming services like Netflix. Given the ratings that customers gave to movies they already watched, one wishes to infer the ratings that these customers are likely to give to unwatched movies in order to make better recommendations. In matrix completion terms, the aim is to reconstruct a ratings matrix $R \in \mathbb{R}^{m\times n}$, where m is the number of customers and n is the number of movies, based on a few observations R_{ij} whose indices (i, j) are grouped in an index set $\Omega \subseteq \{1, \dots, m\} \times \{1, \dots, n\}$.

Not surprisingly, without further assumptions on the structure of the ratings matrix R, the outlined problem is ill-posed as the known ratings imply nothing about the remaining elements of the matrix. The key observation is that, in reality, customers have a limited number of movie tastes. Some might like action movies and science fiction, others might prefer dramas and comedies, again others might enjoy some other combination of movie genres. But the total number of combinations is in the order of ten or twenty, not in the order of hundreds. Therefore, one can expect many of the rows of the ratings matrix to be approximately linearly dependent, namely when they correspond to customers with similar movie tastes. Similarly, one can expect many of its columns to be approximately linearly dependent, namely when they correspond to movies of the same genre. In other words, we can expect the ratings matrix R to be close to a low-rank matrix in the sense that only few of its singular values are significant while the vast majority of them is essentially zero.

This motivates us to determine a hopefully good reconstruction $X \in \mathbb{R}^{m\times n}$ of the true ratings matrix R by searching for a matrix that is compatible with the observed ratings and has lowest rank among all compatible matrices. This gives rise to the optimization problem below.

$$\begin{array}{ll} \underset{X\in\mathbb{R}^{m\times n}}{\text{minimize}} & \operatorname{rank}(X) \\ \text{subject to} & X_{ij} = R_{ij} \quad \forall (i,j) \in \Omega \end{array}$$

Unfortunately, $\operatorname{rank}(X)$ is not a convex function, so the above is not a convex optimization problem. However, we can consider a convex relaxation that replaces the rank function with the nuclear norm $\|X\|_* = \sum_{i=1}^{r} \sigma_i(X)$, where $r = \min\{m, n\}$ and $\sigma_i(X)$ denotes the i-th singular value of X.

$$\begin{array}{ll} \underset{X\in\mathbb{R}^{m\times n}}{\text{minimize}} & \|X\|_* \\ \text{subject to} & X_{ij} = R_{ij} \quad \forall (i,j) \in \Omega \end{array} \tag{1}$$

While problem (1) is convex, it is not in a form that is compatible with the currently available solvers. In this project, we will first equivalently reformulate problem (1) as a semidefinite program (SDP) that can be solved with MOSEK, and then we will try to evaluate whether the low-rank assumption proves to be instrumental in solving the matrix completion problem.

Questions (*Hint: Review trace properties, singular value decomposition, and matrix norms.*)

1. **SDP Reformulation of the Nuclear Norm**: The troubling aspect of problem (1) is its objective function, namely the nuclear norm $\|X\|_*$. In this first part, we find a way to express $\|X\|_*$ as the optimal value of a semidefinite program. To this end, we first establish that the nuclear norm is the dual norm of the spectral norm $\|Y\| = \sigma_{\max}(Y)$, *i.e.*, we show that

$$\begin{aligned} \|X\|_* \;=\; & \max_{Y \in \mathbb{R}^{m\times n}} \quad \operatorname{tr}(X^\top Y) \\ & \text{s.t.} \quad \|Y\| \le 1. \end{aligned} \tag{2}$$

   1.1 Based on the singular value decomposition of X, find some $\hat{Y} \in \mathbb{R}^{m\times n}$ satisfying $\|\hat{Y}\| \le 1$ and achieving $\operatorname{tr}(X^\top \hat{Y}) = \|X\|_*$. This establishes the following inequality. **(10 points)**

$$\begin{aligned} \|X\|_* \;\le\; & \max_{Y \in \mathbb{R}^{m\times n}} \quad \operatorname{tr}(X^\top Y) \\ & \text{s.t.} \quad \|Y\| \le 1 \end{aligned}$$

   1.2 Substitute the singular value decomposition of X into the right-hand side of the inequality below and show that it holds. *Hint: Recall that* $\|Y\| = \sup_{\|v\|_2=1} \sup_{\|u\|_2=1} u^\top Y v$. **(15 points)**

$$\begin{aligned} \|X\|_* \;\ge\; & \max_{Y \in \mathbb{R}^{m\times n}} \quad \operatorname{tr}(X^\top Y) \\ & \text{s.t.} \quad \|Y\| \le 1 \end{aligned}$$

   Together the results obtained in (1.1) and (1.2) prove the equality (2). Observe, however, that the right-hand side of (2) is not yet a semidefinite program. This is our next task.

   1.3 Recalling Schur's lemma, the definition of the spectral norm as $\|Y\| = \sqrt{\lambda_{\max}(Y^\top Y)}$, and the fact that $\lambda_{\max}(A) \le t \iff A \preceq t\mathbb{I}$ for any symmetric matrix A and scalar t, show the equivalence below. **(10 points)**

$$\|Y\| \le 1 \quad \iff \quad \begin{bmatrix} \mathbb{I} & Y^\top \\ Y & \mathbb{I} \end{bmatrix} \succeq 0$$

   With (1.3), we can now express the nuclear norm as the optimal value of the semidefinite program

$$\begin{aligned} \|X\|_* \;=\; & \max_{Y \in \mathbb{R}^{m\times n}} \quad \operatorname{tr}(X^\top Y) \\ & \text{s.t.} \quad \begin{bmatrix} \mathbb{I} & Y^\top \\ Y & \mathbb{I} \end{bmatrix} \succeq 0. \end{aligned} \tag{3}$$

2. **SDP Reformulation of the Matrix Completion Problem**: Note that the matrix completion problem (1) is a minimization problem. In contrast, the SDP reformulation (3) is a maximization problem. In this second part, we will use Lagrangian duality to convert problem (3) into an equivalent minimization problem with the aim to obtain one overall minimization problem.

   2.1 Find a solution $\bar{Y} \in \mathbb{R}^{m\times n}$ of problem (3) that satisfies Slater's condition. **(5 points)**

   Through (2.1) we establish that strong duality holds between problem (3) and its dual problem, and thus that they share the same optimal value. Now, we need to determine the dual problem.

   2.2 Compute the Lagrangian dual problem of problem (3) and show that it can be expressed as problem (4) below. Pay special attention to the fact that we are dualizing a maximization problem (see remark on page 4). Note that the optimal value of problem (4) corresponds to $\|X\|_*$ because strong duality holds between problems (3) and (4). **(30 points)**

$$\begin{aligned} \|X\|_* \;=\; & \min_{\Lambda_1 \in \mathbb{S}^n, \Lambda_2 \in \mathbb{S}^m} \quad \operatorname{tr}(\Lambda_1) + \operatorname{tr}(\Lambda_2) \\ & \text{s.t.} \quad \begin{bmatrix} \Lambda_1 & -\frac{1}{2}X^\top \\ -\frac{1}{2}X & \Lambda_2 \end{bmatrix} \succeq 0 \end{aligned} \tag{4}$$

Based on the computations performed in (2.2), we are now able to express the nuclear norm $\|X\|_*$ as the optimal objective value of the semidefinite minimization problem (4). Since problem (1) and problem (4) are both minimization problems, we can merge them into one overall minimization problem. The matrix completion problem (1) can thus be reformulated equivalently as the semidefinite program (5) shown below.

$$\begin{array}{ll} \underset{X\in\mathbb{R}^{m\times n},\Lambda_1\in\mathbb{S}^n,\Lambda_2\in\mathbb{S}^m}{\text{minimize}} & \operatorname{tr}(\Lambda_1)+\operatorname{tr}(\Lambda_2) \\ \text{subject to} & X_{ij}=R_{ij} \quad \forall (i,j)\in\Omega \\ & \begin{bmatrix} \Lambda_1 & -\frac{1}{2}X^\top \\ -\frac{1}{2}X & \Lambda_2 \end{bmatrix} \succeq 0 \end{array} \tag{5}$$

3. **Numerical Evaluation**:

Consider the datasets `20c50m.mat` (containing 20 customers and 50 movies) and `50c200m.mat` (containing 50 customers and 200 movies). The ij-th entry of each dataset is either a natural number between one and ten (encoding the rating that customer i gave to movie j, 1 = worst, 10 = best) or zero (encoding an unknown rating). Our goal is to use the low-rank SDP method developed in the previous two parts to reconstruct the unknown ratings from the known ones.

3.1 Given the information that the ratings are natural numbers between one and ten, we know that a reasonable reconstruction X of the true ratings matrix R should satisfy the convex constraints

$$1 \leq X_{ij} \leq 10 \quad \forall i=1,\ldots,m,\ \forall j=1,\ldots,n.$$

Add these additional constraints to the reformulated matrix completion problem (5) and implement the enriched reformulation using YALMIP. Solve it with MOSEK first for the dataset `20c50m.mat` and then for the dataset `50c200m.mat` (allow one or two hours runtime for the latter). You can find a skeleton of the code in the Matlab file `p5q3.m`. **(15 points)**

The datasets `20c50m_truth.mat` and `50c200m_truth.mat` contain the true ratings belonging to the datasets `20c50m.mat` and `50c200m.mat`, respectively. Given the true ratings, we would now like to evaluate the performance of the low-rank SDP method implemented in (3.1).

3.2 Complete the following table by computing the quantities defined below for both datasets. Briefly comment on the performance achieved by the low-rank SDP method. **(15 points)**

Dataset	f_{obs}	$f_{\pm 0}$	$f_{\pm 1}$	$f_{\pm 2}$
20c50m				
50c200m				

Notation: Below, $\mathbf{1}_{\{E\}}$ is the indicator function of event E, $X^\star$ denotes the rounded solution obtained in (3.1), and Ω is the index set of observations introduced in the description.

- the fraction f_{obs} of true ratings observed, *i.e.*,

$$f_{\text{obs}} = \tfrac{1}{mn}|\Omega|$$

- the fraction $f_{\pm 0}$ of correctly predicted ratings, *i.e.*,

$$f_{\pm 0} = \tfrac{1}{mn}\sum_{i=1}^m \sum_{j=1}^n \mathbf{1}_{\{X^\star_{ij}=R_{ij}\}}$$

- the fraction $f_{\pm 1}$ of correctly predicted ratings up to a precision of ± 1, *i.e.*,

$$f_{\pm 1} = \tfrac{1}{mn}\sum_{i=1}^m \sum_{j=1}^n \mathbf{1}_{\{X^\star_{ij}\in\{R_{ij}-1,R_{ij},R_{ij}+1\}\}}$$

- the fraction $f_{\pm 2}$ of correctly predicted ratings up to a precision of ± 2, *i.e.*,

$$f_{\pm 2} = \tfrac{1}{mn}\sum_{i=1}^m \sum_{j=1}^n \mathbf{1}_{\{X^\star_{ij}\in\{R_{ij}-2,R_{ij}-1,R_{ij},R_{ij}+1,R_{ij}+2\}\}}$$

Remark:

Recall that a maximization problem with decisions x, feasible set $\mathcal{X}$, and objective function $f_0(x)$ can be equivalently transformed into a minimization problem through the identity

$$\sup_{x\in\mathcal{X}} f_0(x) \;=\; -\inf_{x\in\mathcal{X}} -f_0(x).$$

The given maximization problem (3) can be dualized in several ways – all yielding the same result:

- by transforming the maximization problem to an equivalent minimization problem, applying the standard procedure seen in class, and transforming the dual problem back; or
- by directly dualizing the maximization problem, whereby the standard procedure seen in class needs to be slightly adapted.

Here is the adapted procedure for the second approach. Consider a generic maximization problem

$$\begin{array}{ll} \underset{x\in\mathbb{R}^n}{\text{maximize}} & f_0(x) \\ \text{subject to} & f_i(x) \le 0 \quad \forall i = 1,\dots,m \\ & h_i(x) = 0 \quad \forall i = 1,\dots,p. \end{array}$$

Associating multipliers $\lambda_i \in \mathbb{R}_+$ and $\mu_i \in \mathbb{R}$ with the inequality and equality constraints, respectively, the Lagrangian is computed as (note that the "+" signs of the standard procedure are now "–" signs)

$$L(x,\lambda,\mu) = f_0(x) - \textstyle\sum_{i=1}^m \lambda_i f_i(x) - \sum_{i=1}^p \mu_i h_i(x).$$

The dual objective function is obtained by taking the supremum of the Lagrangian over the primal variables (as opposed to taking the infimum like in the standard procedure), that is,

$$g(\lambda,\mu) = \sup_{x\in\mathbb{R}^n} L(x,\lambda,\mu).$$

Finally, this time round the Lagrangian dual problem is a minimization problem and takes the form

$$\underset{\lambda\in\mathbb{R}^m_+,\mu\in\mathbb{R}^p}{\text{minimize}} \quad g(\lambda,\mu).$$