1 Outline

In this lecture, we study

- matrix completion via the singular value decomposition,
- matrix completion via the power method.

2 Low Rank Matrix Completion

Let us consider

$$\min_{X \in \mathbb{R}^{n \times p}} \quad \|D - X\|_F \quad \text{subject to} \quad \text{rank}(X) \leq k$$

where

- D is an $n \times p$ matrix,
- $\|A\|_F$ denotes the Frobenius norm, i.e., $\|A\|_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{p} A_{ij}^2}$.

By the definition of the Frobenius norm, the problem is equivalent to

$$\min_{X \in \mathbb{R}^{n \times p}} \quad \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{p} (D_{ij} - X_{ij})^2 \quad \text{subject to} \quad \text{rank}(X) \leq k.$$

2.1 Applications

The most common application is matrix compression, where the goal is to provide a lossy compressed version of a given matrix. There are other practical applications, which we elaborate on below.

Movie Recommendation A common application is movie recommendation where matrix D collects user preference data of n users rating p movies. In movie recommendation, the user rating matrix D is typically very sparse, as it is rare that an user rates all movies. Hence, the goal is to find a matrix X that completes the missing entries of D. The general hypothesis is that the true rating matrix $X \in \mathbb{R}^{n \times p}$ is generated by the product of an user-feature matrix $U \in \mathbb{R}^{n \times k}$ and a movie-feature matrix $V \in \mathbb{R}^{p \times k}$ over k features as follows.

$$\begin{bmatrix} \\ \\ \\ \\ \end{bmatrix} \begin{matrix} \\ \\ X \\ \\ \end{matrix} \begin{bmatrix} \\ \\ \\ \\ \end{bmatrix} = \begin{bmatrix} \\ \\ \\ \\ \end{bmatrix} \begin{matrix} \\ \\ U \\ \\ \end{matrix} \begin{bmatrix} \\ \\ \\ \\ \end{bmatrix} \begin{matrix} \\ \\ V^\top \\ \\ \end{matrix} \begin{bmatrix} \\ \\ \\ \\ \end{bmatrix}.$$

Then matrix X has rank at most k.

Localization in Sensor Networks Suppose that there are n sensors whose physical locations are given by coordinate vectors $x_1, \ldots, x_n \in \mathbb{R}^3$. Here, these vectors are not given while we have access to the distance information about some pairs of distinct sensors. Note that the distance between x_i and x_j is given by

$$D_{ij} = \|x_i - x_j\|_2^2 = \|x_i\|_2^2 + \|x_j\|_2^2 - 2x_i^\top x_j.$$

Let X be the $n \times 3$ matrix whose rows are $x_1^\top, \ldots, x_n^\top$, $y \in \mathbb{R}^n$ be the vector whose components are $\|x_1\|_2^2, \ldots, \|x_n\|_2^2$, and let $\mathbf{1} \in \mathbb{R}^n$ be the vector of all ones. Then it follows that

$$D = y\mathbf{1}^\top + \mathbf{1}y^\top - 2XX^\top.$$

As the rank of X is at most 3, the rank of matrix D is at most 5. Basically, we are given only a subset of the entries of D, from which we want to predict matrix the sensor location matrix X.

2.2 Low-Rank Approximation by the Singular Value Decomposition

The singular value decomposition (SVD) of a matrix $D \in \mathbb{R}^{n\times p}$ is given by

$$D = U\Sigma V^\top$$

where

- $r = \min\{n, p\}$,
- $U \in \mathbb{R}^{n\times r}$ is an $n \times r$ matrix with orthonormal columns,
- $\Sigma \in \mathbb{R}^{r\times r}$ is a $r \times r$ diagonal matrix whose entries are the singular values of D: $\sigma_1 \geq \cdots \geq \sigma_r \geq 0$,
- $V \in \mathbb{R}^{p\times r}$ is a $p \times r$ matrix with orthonormal columns.

Here, for any $k \leq r$, let U_k denote the column submatrix of U that consists of the first k columns of U, and let V_k denote the column submatrix of V that consists of the first k columns of V. Moreover, let Σ_k be the $k \times k$ diagonal matrix whose entries are the top k singular values $\sigma_1, \ldots, \sigma_k$ of D. Then we may consider

$$D_k = U_k\Sigma_k V_k^\top. \tag{7.1}$$

Here, the matrix D_k is of rank k. We can argue is that D_k is the best rank-k approximation of D. To be more precise, we have the following result.

Theorem 7.1. *For any $k \leq \min\{n, p\}$, we have*

$$D_k \in \operatorname{argmin}_{X\in\mathbb{R}^{n\times p}} \{\|D - X\|_F : \ \operatorname{rank}(X) \leq k\}$$

where D_k is defined as in (7.1).

Proof. The Frobenius norm is unitarily invariant, so for any X,

$$\|D - X\|_F = \|U^\top(D - X)V\|_F = \|\Sigma - Y\|_F, \quad \text{where } Y := U^\top XV.$$

Moreover, $\operatorname{rank}(Y) = \operatorname{rank}(X)$. Hence we have

$$\min_{\operatorname{rank}(X)\leq k} \|D - X\|_F = \min_{\operatorname{rank}(Y)\leq k} \|\Sigma - Y\|_F.$$

Next, note that for any Y with $\text{rank}(Y) \leq k$,

$$\|\Sigma - Y\|_F^2 = \|\Sigma\|_F^2 + \|Y\|_F^2 - 2 \cdot \text{tr}\left(\Sigma^\top Y\right).$$

Von Neumann's trace inequality gives

$$\text{tr}\left(\Sigma^\top Y\right) \leq \sum_{i=1}^{r} \sigma_i \cdot \sigma_i(Y)$$

where $\sigma_1(Y) \geq \sigma_2(Y) \geq \cdots \geq \sigma_r(Y) \geq 0$ are the singular values of Y. Since $\text{rank}(Y) \leq k$, $\sigma_i(Y) = 0$ for $i > k$. Therefore, it follows that

$$\text{tr}\left(\Sigma^\top Y\right) \leq \sum_{i=1}^{k} \sigma_i \cdot \sigma_i(Y) \leq \left(\sum_{i=1}^{k} \sigma_i^2\right)^{1/2} \left(\sum_{i=1}^{k} \sigma_i(Y)^2\right)^{1/2} = \|\Sigma_k\|_F \, \|Y\|_F$$

where the second inequality holds due to the Cauchy-Schwarz inequality. Therefore,

$$\|\Sigma - Y\|_F^2 \;\geq\; \|\Sigma\|_F^2 + \|Y\|_F^2 - 2\|\Sigma_k\|_F \, \|Y\|_F = \left(\|Y\|_F - \|\Sigma_k\|_F\right)^2 + \left(\|\Sigma\|_F^2 - \|\Sigma_k\|_F^2\right).$$

The first term is nonnegative, hence

$$\|\Sigma - Y\|_F^2 \;\geq\; \|\Sigma\|_F^2 - \|\Sigma_k\|_F^2 = \sum_{i=k+1}^{r} \sigma_i^2. \qquad (*)$$

Taking $Y^\star := \begin{bmatrix} \Sigma_k & 0 \\ 0 & 0 \end{bmatrix}$ of rank k, we have

$$\|\Sigma - Y^\star\|_F^2 = \sum_{i=k+1}^{r} \sigma_i^2,$$

achieving the lower bound $(*)$. Here, $Y^\star$ corresponds to $X^\star = UY^\star V^\top = U_k \Sigma_k V_k^\top = D_k$. Thus D_k is optimal and the minimum error is $\sum_{i=k+1}^{r} \sigma_i^2$. □

Hence, one can compute an optimal low-rank approximation by computing the singular value decomposition. It is known that the complexity of deriving the singular value decomposition is

$$O(npr).$$

When D is an $n = p = r$, the complexity is of order $O(n^3)$, which is typically inefficient in practice.

3 Computing the Top Eigenvector

For low-rank matrix completion, what we really need is to compute the top k singular values and the corresponding singular vectors, while SVD computes all singular values and the corresponding singular vectors. In this section, we present an efficient method for computing the top singular value and the corresponding singular vectors.

Given a positive semidefinite matrix $A \in \mathbb{R}^{d \times d}$, we want to compute the top eigenvector of A which corresponds to the largest eigenvalue of A. Let $\lambda_{\max}$ denote the largest eigenvalue of A, and let $v_{\max}$ denote the top eigenvector. The closely related notion is the **Rayleigh quotinet**, defined as

$$R(A, x) = \frac{x^\top A x}{x^\top x} \quad \text{for any nonzero } x.$$

It is known that for any symmetric matrix A,

$$\max_{x\in\mathbb{R}^d:x\neq 0} R(A,x) = \lambda_{\max},$$
$$\operatorname*{argmax}_{x\in\mathbb{R}^d:x\neq 0} R(A,x) = v_{\max}.$$

This implies that we may compute the top eigenvector of a symmetric matrix A by solving

$$\max_{x\in\mathbb{R}^d} \quad x^\top Ax \quad \text{subject to} \quad \|x\|_2 = 1.$$

If A is positive semidefinite, then $x^\top Ax \geq 0$ for any $x \in \mathbb{R}^d$, which implies that one can replace the constraint $\|x\|_2 = 1$ by $\|x\|_2 \leq 1$ to deduce the equivalent formulation

$$\max_{x\in\mathbb{R}^d} \quad x^\top Ax \quad \text{subject to} \quad \|x\|_2 \leq 1.$$

Here, the feasible region $\{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$ is a convex set. However, the objective is to "maximize" the convex function $x^\top Ax$. Hence, the optimization problem is a nonconvex problem. Nevertheless, there exists an efficient algorithm for computing the top eigenvector.

3.1 Power Method

We present the **power method** which is known to compute the top eigenvector of a positive semidefinite matrix. The power method works with as in Algorithm 1.

Algorithm 1 Power Method

Initialize $\bar{x}_0 \in \mathbb{R}^d \setminus \{0\}$ and $x_0 = \bar{x}_0/\|\bar{x}_0\|_2$
for $t = 1, \ldots, T$ **do**
  Update $\bar{x}_t = A^t x_0$
  Obtain $x_t = \bar{x}_t/\|\bar{x}_t\|_2$
end for
Return x_T.

To analyze the power method, we explain some necessary background. Let A be a positive semidefinite matrix with eigenvalues

$$\lambda_1 \geq \cdots \geq \lambda_d \geq 0.$$

Let v_i denote the eigenvector associated with λ_i for $i \in [d]$. By the well-known **spectral theorem**, we may assume that $v_1, \ldots, v_d$ are mutually orthogonal and $\|v_i\| = 1$ for $i \in [d]$. Then $v_1, \ldots, v_d$ form an orthonormal basis of $\mathbb{R}^d$.

Theorem 7.2. *Suppose that $\lambda_1 > \lambda_2$. For any $\epsilon > 0$, if*

$$t \geq \frac{\lambda_1}{2(\lambda_1 - \lambda_2)} \log\left(\frac{1}{\epsilon(v_1^\top x_0)^2}\right),$$

then

$$(v_1^\top x_t)^2 \geq 1 - \epsilon$$

Proof. Since $v_1, \ldots, v_d$ form an orthonormal basis of $\mathbb{R}^d$, we have

$$x_0 = \sum_{i=1}^{d} \alpha_i v_i$$

where $\alpha_i = (v_i^\top x_0)$. Moreover, as $\bar{x}_t = A^t x_0$, it follows that

$$\bar{x}_t = \sum_{i=1}^{d} \alpha_i A^t v_i = \sum_{i=1}^{d} \alpha_i \lambda_i^t v_i.$$

Note that

$$x_t = \sum_{i=1}^{d} (v_i^\top x_t) v_i \quad \text{and} \quad \|x_t\|_2^2 = \sum_{i=1}^{d} (v_i^\top x_t)^2.$$

Since $\|x_t\|_2 = 1$, we have

$$1 - (v_1^\top x_t)^2 = \sum_{i=2}^{d} (v_i^\top x_t)^2 = \frac{1}{\|\bar{x}_t\|_2^2} \sum_{i=2}^{d} (v_i^\top \bar{x}_t)^2 = \frac{\sum_{i=2}^{d} \alpha_i^2 \lambda_i^{2t}}{\sum_{i=1}^{d} \alpha_i^2 \lambda_i^{2t}} \leq \frac{\lambda_2^{2t} \sum_{i=2}^{d} \alpha_i^2}{\alpha_1^2 \lambda_1^{2t}} \leq \frac{\lambda_2^{2t}}{\alpha_1^2 \lambda_1^{2t}}.$$

This implies that

$$1 - (v_1^\top x_t)^2 \leq \left(\frac{\lambda_2}{\lambda_1}\right)^{2t} \frac{1}{(v_1^\top x_0)^2}.$$

Hence, if

$$t \geq \frac{1}{2} \cdot \frac{1}{\log(\lambda_1/\lambda_2)} \log\left(\frac{1}{\epsilon (v_1^\top x_0)^2}\right),$$

then $(v_1^\top x_t)^2 \geq 1 - \epsilon$. Here,

$$\log\left(\frac{\lambda_1}{\lambda_2}\right) = \log\left(\frac{1}{1 - \frac{\lambda_1 - \lambda_2}{\lambda_1}}\right) = -\log\left(1 - \frac{\lambda_1 - \lambda_2}{\lambda_1}\right) \geq \frac{\lambda_1 - \lambda_2}{\lambda_1}.$$

Therefore, if

$$t \geq \frac{\lambda_1}{2(\lambda_1 - \lambda_2)} \log\left(\frac{1}{\epsilon (v_1^\top x_0)^2}\right),$$

then $(v_1^\top x_t)^2 \geq 1 - \epsilon$, as required. □

What does the theorem imply? Note that

$$\sum_{i=2}^{d} (v_i^\top x_t)^2 = \|x_t\|_2^2 - (v_1^\top x_t)^2 = 1 - (v_1^\top x_t)^2 \leq \epsilon.$$

Therefore, x_t is aligned with v_1, which is the top eigenvector. Moreover, the bound has the term $(v_1^\top x_0)^2$. If we sample x_0 from the multivariate normal distribution $\mathcal{N}(0, I_d)$ where I_d is the $d \times d$ identity matrix, then we can argue that $(v_1^\top x_0) = 1/\text{poly}(d)$ with high probability. In that case, the number of required iterations is

$$O\left(\frac{\lambda_1}{\lambda_1 - \lambda_2} \log\left(\frac{d}{\epsilon}\right)\right).$$

Here, the bound depends on the values of λ_1 and λ_2. The following result provides a convergence guarantee of the power method that does not depend on λ_1 and λ_2.

Theorem 7.3. *For any $\epsilon > 0$, if*

$$t \geq \frac{1}{\epsilon} \log\left(\frac{2}{\epsilon (v_1^\top x_0)^2}\right),$$

then

$$x_t^\top A x_t \geq (1-\epsilon)\lambda_1.$$

Proof. Let k denote the largest index such that

$$\lambda_k \geq \left(1 - \frac{\epsilon}{2}\right)\lambda_1.$$

Let $\alpha_i = (v_i^\top x_0)$ for $i \in [d]$. Note that

$$1 - \sum_{i=1}^{k} (v_i^\top x_t)^2 = \sum_{i=k+1}^{d} (v_i^\top x_t)^2 = \frac{1}{\|\bar{x}_t\|_2^2} \sum_{i=k+1}^{d} (v_i^\top \bar{x}_t)^2 = \frac{\sum_{i=k+1}^{d} \alpha_i^2 \lambda_i^{2t}}{\sum_{i=1}^{d} \alpha_i^2 \lambda_i^{2t}} \leq \left(1 - \frac{\epsilon}{2}\right)^{2t} \frac{1}{(v_1^\top x_0)^2}.$$

Moreover,

$$x_t^\top A x_t = \sum_{i=1}^{d} \lambda_i (v_i^\top x_t)^2 \geq \left(1 - \frac{\epsilon}{2}\right)\lambda_1 \sum_{i=1}^{k} (v_i^\top x_t)^2.$$

Combining the two inequalities, it follows that

$$x_t^\top A x_t \geq \left(1 - \frac{\epsilon}{2}\right)\lambda_1 \left(1 - \left(1 - \frac{\epsilon}{2}\right)^{2t} \frac{1}{(v_1^\top x_0)^2}\right).$$

Here, if

$$t \geq \frac{1}{\epsilon} \log\left(\frac{2}{\epsilon (v_1^\top x_0)^2}\right),$$

then

$$x_t^\top A x_t \geq \left(1 - \frac{\epsilon}{2}\right)^2 \lambda_1 \geq (1-\epsilon)\lambda_1,$$

as required. □

Again, If we sample x_0 from the multivariate normal distribution $\mathcal{N}(0, I_d)$, then the number of required iterations is

$$O\left(\frac{1}{\epsilon} \log\left(\frac{d}{\epsilon}\right)\right)$$

with high probability.

3.2 Low-Rank Approximation by the Power Method

Instead of computing the full singular value decomposition of a given matrix D, we may use the power method to compute the optimal low-rank approximation D_k of rank k.

Let u_1 denote the **top left singular vector** that corresponds to σ_1, and let v_1 denote the **top right singular vector** corresponding to σ_1. Note that

$$D^\top D = V \Sigma^2 V^\top \quad \text{and} \quad D D^\top = U \Sigma^2 U^\top$$

which implies that

$$\sigma_1 = \sqrt{\lambda_{\max}(D^\top D)} = \sqrt{\lambda_{\max}(D D^\top)}.$$

Algorithm 2 Power Method for the Top Left Singular Vector

Initialize $\bar{u}_0 \in \mathbb{R}^n \setminus \{0\}$ and $u_0 = \bar{u}_0 / \|\bar{u}_0\|_2$
for $t = 1, \ldots, T$ **do**
  Update $\bar{u}_t = (DD^\top)^t x_0$
  Obtain $u_t = \bar{u}_t / \|\bar{u}_t\|_2$
end for
Return u_T.

Algorithm 3 Power Method for the Top Right Singular Vector

Initialize $\bar{v}_0 \in \mathbb{R}^p \setminus \{0\}$ and $v_0 = \bar{v}_0 / \|\bar{v}_0\|_2$
for $t = 1, \ldots, T$ **do**
  Update $\bar{v}_t = (D^\top D)^t v_0$
  Obtain $v_t = \bar{v}_t / \|\bar{v}_t\|_2$
end for
Return v_T.

Therefore, by applying the power method to $D^\top D$, one can obtain the top right singular vector v_1. Similarly, by applying the power method to $DD^\top$, we can deduce the top left singular vector u_1. A pseudo-code of the method is given in Algorithms 2 and 3.

Then it follows that

$$D - u_1\sigma_1 v_1^\top = U\Sigma V^\top - u_1\sigma_1 v_1^\top = \begin{bmatrix} & & \\ u_2 & \cdots & u_r \\ & & \end{bmatrix} \begin{bmatrix} \sigma_2 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix} \begin{bmatrix} v_2^\top \\ \vdots \\ v_r^\top \end{bmatrix}.$$

Then we may apply the power method to the matrix $D - u_1\sigma_1 v_1^\top$ to compute u_2, v_2, and σ_2, which are the top left singular vector, the top right singular vector, and the largest singular value of $D - u_1\sigma_1 v_1^\top$.