CSE 521: Design and Analysis of Algorithms I **Spring 2016**

Lecture 9: SVD, Low Rank Approximation

Lecturer: Shayan Oveis Gharan *April 25th* *Scribe: Koosha Khalvati*

9.1 Singular Value Decomposition (SVD)

Claim 9.1. *For matrix $M \in \mathbb{R}^{m\times n}(m \le n)$, there exist two sets of k orthonormal columns $v_1, \dots, v_k$ and $u_1, \dots, u_k$ such that:*

$$M = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$

where for each i, $\sigma > 0$ and $k \le \min\{m, n\}$.

In the above representation, σ_i is called a singular value, u_i is a left singular vector, and v_i is a right singular vector.

Proof. To prove the existence of these two sets of columns, we construct them: First, note that $M^T M \in \mathbb{R}^{n\times n}$ is a symmetric and positive semidefinite matrix. To see the latter observe that for any vector $x \in \mathbb{R}^n$,

$$x^T(M^T M)x = \langle x, M\rangle^2 \ge 0.$$

Secondly, by the spectral theorem, there exists eigenvalues $\lambda_1, \dots, \lambda_n$ with corresponding orthonormal eigenvectors $v_1, \dots, v_n$ such that:

$$M^T M = \sum_{i=1}^{n} \lambda_i v_i v_i^T$$

Use these v_i's (eigenvectors of $M^T M$), the right singular vectors in SVD of M. Note that in the above representation for all i, $\lambda_i \ge 0$. We also need to throw away all v_i, λ_i where $\lambda_i = 0$.

It remains to construct u_i. We let each u_i be proportional to Mv_i. Firstly, observe that for any i, j, Mv_i, Mv_j are orthogonal,

$$\langle Mv_i, Mv_j\rangle = \langle M^T Mv_i, v_j\rangle = \langle \lambda_i v_i, v_j\rangle = \begin{cases} \lambda_i & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases} \tag{9.1}$$

So, we just need to normalize u_i. Define,

$$u_i = \frac{Mv_i}{\|Mv_i\|} = \frac{Mv_i}{\sqrt{\lambda_i}} = \frac{Mv_i}{\sigma_i}$$

Note that singular values σ_i are equal to $\sqrt{\lambda_i}$; since $M^T M$ is PSD, $\lambda_i \ge 0$ and σ_i is well defined. In particular, observe that if M is a symmetric matrix, σ_i is the absolute value of the i-th eigenvalue of M.

Now, we want to show that these v_is and u_is meet SVD conditions. Recall that v_i's are orthonormal because they are eigenvectors of $M^T M$, and u_is are orthonormal by (9.1).

We claim that

$$M = \sum \sigma_i u_i v_i^T$$

. One way to prove that two operators are the same is to show that they act identically on a set of basis vectors $w_1, \ldots, w_n$. In particular, we show that for any j,

$$Mv_j = \left(\sum \sigma_i u_i v_i^T\right) v_j. \tag{9.2}$$

If the above holds for all i, since $v_1, \ldots, v_n$ form a basis, for any vector v one can write

$$Mv = M\left(\sum_{j=1}^{n} v_j \langle v, v_j\rangle\right) = \sum_{j=1}^{n} Mv_j \langle v, v_j\rangle = \sum_{j=1}^{n} \left(\sum \sigma_i u_i v_i^T\right) v_j \langle v, v_i\rangle = \left(\sum \sigma_i u_i v_i^T\right) v.$$

So, the two matrix are indeed equal.

It remains to prove (9.2) for all j.

$$\left(\sum \sigma_i u_i v_i^T\right) v_j = \sum_{i=1}^{k} \sigma_i u_i \langle v_i, v_j\rangle = \sigma_j u_j = \sigma_j \frac{Mv_j}{\sigma_j} = Mv_j.$$

This is because $\langle v_i, v_j\rangle$ is equal to 0 when i and j are different and it is equal to 1 when $i = j$. In the last equality we used the definition of u_i. This proves (9.2), as desired. □

9.2 Low Rank Approximation

In the rest of this lecture and part of the next one we study low rank approximation of matrices.

First, let's define the rank of the matrix: There are many ways one can define the rank of a matrix. Rank of Matrix M, $rank(M)$, is the number of linearly independent columns in M. It is also equal to the number of linearly independent rows in M. In addition, it is equal to the number of none-zero eigenvalues (or singular values) of M.

It is possible to prove that all of the definitions above are one concept.

9.2.1 Motivation

In the low rank approximation, the goal is to approximate a given matrix M, with a low rank Matrix, i.e. $\tilde{M}_k$ such that $\|M - \tilde{M}_k\|$ is approximately zero. We have not specified the norm; in general one should choose the norm based on the specific application.

This area is also known as principal component analysis. It has a similar flavor as the dimension reduction technique that we studied a few lectures ago. We have a high dimensional data represented as a matrix and we want to approximate it with a much lower dimensional object. Note that a rank k approximation of M can be expressed using only k singular values and the corresponding k singular vectors. So, in principal, a rank k approximation only needs $O(kn)$ bits of memory.

So, at a very high level one can use low rank approximation to approximately store an $m \times n$ matrix with only $O(kn)$ bits. In the next lecture we will see some of its applications in optimization. Besides these applications, one can use low rank approximation to reveal hidden structures in a given data set. Let us give several examples of this phenomenon.

9.2.2 Application1: Consumer-Product Matrix

A Consumer-product matrix is a matrix ($M \in \mathbb{R}^{n\times d}$) where each row corresponds to a consumer and each column corresponds to a product. Entry i, j of the matrix represents the probability that consumer i purchases product j. We can hypothesize that there are of k hidden features in consumers such as age, gender, and annual income, etc and the decision of each consumer is only a function of these hidden features.

With this hypothesis we can rewrite

$$M = AB$$

as a product of two matrices: a factor weigh matrix, $A \in \mathbb{R}^{n\times k}$ and a purchase probability matrix $B \in \mathbb{R}^{k\times d}$. In particular, each row of A represents a consumer as a weighted sum of the k underlying features and each column of B represents the purchase probability for a consumer with only one feature.

Ideally, all elements of consumer-product matrix are available and we can use a low rank approximation of M to find these k hidden features for a small value of k. In the real world however, we usually have only some of the elements of the matrix. So, our goal is to predict the unknown elements. There are several examples of such a question. In the Netflix challenge, we were given a partial ratings of the Netflix users and our goal was to predict the rating of each user for the rest of the movies. In online advertisement, the goal is to find the best ad to show to a given user, given that he has purchased one or two items in the past.

9.2.3 Application2: Hidden Partitioning

Given graph, assume that there are two hidden communities such that the probability p of existence of an edge between the nodes of one community is much more than the probability q of an edge connecting the nodes in different communities. Given a sample of such a graph our goal is to recover the hidden communities with high probability. If we reorder the vertices such that the first community consists of nodes $1, \ldots, n/2$ and the second one consists of $n/2 + 1, \ldots, n$, the expected matrix look like the following:

$$\tilde{M}_2 = \begin{bmatrix} p & q \\ q & p \end{bmatrix}$$

Observe that the above matrix is a rank 2 matrix. So, one may expect that by using a rank 2 approximation of the adjacency matrix of the given graph he can recover the hidden partition. This is indeed possible assuming p is sufficiently larger than q [McS01].

9.2.4 Application3: Image Compression

As each image is represented by a matrix, it is possible to compress an image by approximating it by a lower-rank matrix. To see an example of image compression by lower-rank matrix approximation in MATLAB, please check the course homepage.

9.2.5 Principal Component Analysis

In this section we study low rank approximation with respect to the operator norm.

Definition 9.2 (Norm of the Matrix)**.** *The operator norm of a matrix M is defined as follows:*

$$\|M\|_2 = \max_x \frac{\|Mx\|_2}{\|x\|_2}.$$

In words, the operator norm shows how large a vector x can get once it is multiplied with M. It follows by the Rayleigh quotient that the operator norm of M is equal to its largest singular value.

Claim 9.3. $\|M\|_2 = \sigma_{\max}(M)$

Proof. We can write,

$$\max_x \frac{\|Mx\|_2}{\|x\|_2} = \max_x \sqrt{\frac{\|Mx\|_2^2}{\|x\|_2^2}} = \sqrt{\frac{x^T M^T M x}{x^T x}} = \sqrt{\lambda_{\max}(M^T M)} = \sigma_{\max}(M),$$

where the second to last equality follows by the Rayleigh quotient. □

In the following theorem we show that the best rank k approximation with respect to operator norm satisfies $\|M - \tilde{M}_k\| \leq \sigma_{k+1}$. Note that this approximation does not directly imply that the entries of $M - \tilde{M}_k$ are closed to zero; it only upper bounds the action of $M - \tilde{M}_k$ on vectors.

Theorem 9.4. *Let $M \in \mathbb{R}^{m \times n}$ with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$. For any integer $k \geq 1$,*

$$\min_{\tilde{M}_k} \|M - \tilde{M}_k\|_2 = \sigma_{k+1} \tag{9.3}$$

Proof. We need to prove two statements: i) There exist a rank-k matrix $\tilde{M}_k$ such that $\|M - \tilde{M}_k\|_2 = \sigma_{k+1}$. (2) For any rank k matrix $\tilde{M}_k$, $\|M - \tilde{M}_k\|_2 \geq \sigma_{k+1}$.

We start with part (i). We let

$$\tilde{M}_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$

By definition of M,

$$M - \tilde{M}_k = \sum_{i=k+1}^{m} \sigma_i u_i v_i^T \Rightarrow \|M - \tilde{M}_k\|_2 = \sigma_{max}\left(\sum_{i=k+1}^{m} \sigma_i u_i v_i^T\right) = \sigma_{k+1}$$

where the last equality uses Claim 9.3.

Now, we prove part (ii). So, assume $\tilde{M}_k$ is an arbitrary rank k matrix. For a matrix M, the null space of M

$$\text{NULL}(M) = \{x : Mx = 0\}$$

is the set of vector that M map to zero. Let null(M) be the dimension of the null space of M, NULL(M). It is a well known fact that for any $M \in \mathbb{R}^{m \times n}$,

$$\text{rank}(M) + \text{null}(M) = n.$$

This is because any vector which is orthogonal to the right singular vectors of M is in NULL(M). So, the above equality follows from the fact that M has rank(M) right singular vectors (with positive singular values). As an application, since $\tilde{M}_k$ has rank k, we must have

$$\text{null}(\tilde{M}_k) = n - \text{rank}(\tilde{M}_k) = n - k. \tag{9.4}$$

By the Rayleigh quotient, we can write,

$$\begin{aligned}\sigma_{k+1}(M)^2 = \lambda_{k+1}(M^TM) &= \min_{S:(n-k)\text{dims}} \max_{x\in S} \frac{x^TM^TMx}{x^Tx} \\ &\leq \max_{x\in \text{NULL}(\tilde{M}_k)} \frac{x^TM^TMx}{x^Tx} \\ &= \max_{x\in \text{NULL}(\tilde{M}_k)} \frac{x^T(M-\tilde{M}_k)^T(M-\tilde{M}_k)x}{x^Tx} \\ &\leq \max_{x} \frac{x^T(M-\tilde{M}_k)^T(M-\tilde{M}_k)x}{x^Tx}.\end{aligned}$$

The first inequality uses the fact that NULL(M) is a $n-k$ dimensional linear space; so a special case of S being a $n-k$ dimensional linear space is $S = NULL(M)$. The second equality uses that $\tilde{M}_kx = 0$ for any $x \in \text{NULL}(\tilde{M}_k)$.

Now, we are done using another application of the Rayleigh quotient.

$$\max_x \frac{x^T(M-\tilde{M}_k)^T(M-\tilde{M}_k)x}{x^Tx} = \lambda_{\max}((M-\tilde{M}_k)^T(M-\tilde{M}_k)) = \sigma_{\max}(M-\tilde{M}_k)^2.$$

This completes the proof of Theorem 9.4. □

9.2.6 Approximation of Frobenius Norm

In this section we study low rank approximation with respect to different norm known as the Frobenius norm.

Definition 9.5 (Frobenius Norm)**.** *For any matrix M, the Frobenius norm of M is defined as follows:* $\|M\|_F = \sqrt{\sum i,jM_{i,j}^2}$

Again, we are looking for the best rank-k approximation with respect to the Frobenius Norm.

Theorem 9.6. *Given a matrix $M \in \mathbb{R}m \times n$ with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$, we have*

$$\min_{\tilde{M}_k} \|M-\tilde{M}_k\|_F^2 = \sum_{i=k+1}^{m} \sigma_i^2 \tag{9.5}$$

Before, proving the theorem, let us see how we can relate the Frobenius norm to the singular values of a matrix.

Claim 9.7. *For any matrix $M \in \mathbb{R}^{m\times n}$,*

$$\|M\|_F^2 = \sum \sigma_i^2$$

Proof. First, observe that i,i-th entry of the matrix M^TM, is the inner product of the i-th column of M with itself.

$$(M^TM)_{i,i} = \sum_j M_{j,i}^2.$$

Summing up over all i we have

$$\sum_{i=1}^{n}(M^TM)_{i,i} = \sum_{i,j} M_{i,j}^2 = \|M\|_F^2.$$

Now, observe the the LHS is indeed $\mathrm{Tr}(M^T M)$. In the last lecture we proved that the trace of a symmetric matrix is equal to sum of its eigenvalues. So, the LHS is equal to $\sum \sigma_i^2$. □

To prove Theorem 9.6 we need to prove two statements: i) There is a rank-k matrix $\tilde{M}_k$ such that $\|M - \tilde{M}_k\|_F^2 = \sum_{i=k+1}^m \sigma_i^2$ ii) For any rank k matrix $\tilde{M}_k$, $\|M - \tilde{M}_k\|_F^2 \geq \sum_{i=k+1}^m \sigma_i^2$. Here, we only prove (i). Similar to Theorem 9.4 let

$$\tilde{M}_k = \sum_{i=k+1}^{m} \sigma_i u_i v_i^T,$$

Then,

$$\|M - \tilde{M}_k\|_F^2 = \left\| \sum_{i=k+1}^{m} \sigma_i u_i v_i^T \right\| = \sum_{i=k+1}^{m} \sigma_i^2.$$

as desired.

References

[McS01] F. McSherry. “Spectral Partitioning of Random Graphs”. In: *FOCS*. 2001, pp. 529–537 (cit. on p. 9-3).