

## Recitation 9

Lecturer: Regev Schweiger

Scribe: Regev Schweiger

## 9.1 Singular Value Decomposition (SVD)

### 9.1.1 Definitions and Intuition

#### Symmetric Matrices

The SVD is related to the familiar theory of diagonalizing a symmetric matrix. Recall that if  $A$  is a symmetric real  $n \times n$  matrix, there is an orthonormal matrix  $V$  and a diagonal matrix  $D$  such that  $A = VDV^T$ . Here, the columns of  $V$  are eigenvectors for  $A$  and form an orthonormal basis for  $\mathbb{R}^n$ ; the diagonal entries of  $D$  are the eigenvalues of  $A$ .

A useful way to think about such matrices, is as a composition of three operations:

1. First,  $V^T = V^{-1}$  - a *change of basis transformation*, that receives as input a vector given in the coordinates of the standard basis and outputs the coordinates of the vector in the basis of the columns of  $V$ ;
2. Secondly,  $D$  - a *coordinate-wise transformation*, that independently multiplies each coordinate by a scalar; the transformation simply dilates some components and contracts others, according to the magnitudes of the eigenvalues (with a reflection through the origin also possible, for negative eigenvalues);
3. Finally,  $V$  - the *reverse change of basis transformation*, that receives as input a vector given in the basis of the columns of  $V$  and outputs the coordinates of the vector in the standard basis.

Often, when choosing the right basis according to which to examine the matrix, its properties become clearer.

**Example: Projection Matrix.** Let  $W$  be a  $m$  dimensional subspace. The *projection matrix*  $P$  with respect to  $W$  is the linear transformation that returns the projection of a vector  $\mathbf{v}$  on the subspace spanned by  $W$ . Namely, if  $\mathbf{v} = \mathbf{w}^\perp + \mathbf{w}^\parallel$ , where  $\mathbf{w}^\perp$  is orthogonal to  $W$  and  $\mathbf{w}^\parallel \in W$ , then  $P(\mathbf{v}) = \mathbf{w}^\parallel$ .

Let us choose an orthonormal basis  $\mathbf{v}_1, \dots, \mathbf{v}_m$  that spans  $W$ , and complete it to a full orthonormal basis  $\mathbf{v}_1, \dots, \mathbf{v}_m, \mathbf{v}_{m+1}, \dots, \mathbf{v}_n$ . Then, in this basis, the projection matrix  $P$  is easily seen to be:

$$D = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & & 0 & & \\ & 0 & & & \ddots & \\ & & & & & 0 \end{pmatrix}$$

or, in the standard basis,  $VDV^T$ . In this basis, we can easily see properties of projection matrices such as  $P^2 = P$  and that  $P\mathbf{w} = \mathbf{w}$  for all  $\mathbf{w} \in W$ .

## SVD

For the SVD we begin with an arbitrary real  $m \times n$  matrix  $A$ . There are orthonormal matrices  $U$  and  $V$  and a diagonal matrix, this time denoted  $\Sigma$ , such that  $A = U\Sigma V^T$ . In this case,  $U$  is  $m \times m$  and  $V$  is  $n \times n$ , so that  $\Sigma$  is rectangular with the same dimensions as  $A$ . The diagonal entries of  $\Sigma$ , that is  $\Sigma_{i,i} = \sigma_i$ , can be arranged to be nonnegative and in order of decreasing magnitude. These are called the singular values of  $A$ . The columns of  $U$  and  $V$  are called left and right singular vectors, for  $A$ .

Now we can look at  $A$  as a composition of three operations:

1. First,  $V^T = V^{-1}$  - a *change of basis transformation*, that receives as input a vector given in the coordinates of the standard basis and outputs the coordinates of the vector in the basis of the columns of  $V$ ;
2. Secondly,  $\Sigma$  - a *coordinate-wise transformation* that simply dilates some components and contracts others, according to the magnitudes of the singular values, and possibly discards components or appends zeros as needed to account for a change in dimension;
3. Finally,  $U$  - a **different** *change of basis transformation*, that receives as input a vector given in the basis of the columns of  $U$  and outputs the coordinates of the vector in the standard basis.

Note that  $U$  and  $V$  are different bases, and in fact may be of different dimensions.

## Null and Range Subspaces

Recall that for a matrix  $A$ ,  $\ker(A) = \{\mathbf{x} \in \mathbb{R}^n \mid A\mathbf{x} = \mathbf{0}\}$  and  $\text{im}(A) = \{\mathbf{y} \in \mathbb{R}^m \mid \exists \mathbf{x}, A\mathbf{x} = \mathbf{y}\}$  is the subspace spanned by the columns of  $A$ . The SVD gives a natural interpretation of these subspaces and their orthogonal complements: First, let's assume  $m < n$ , and for simplicity, assume all  $\sigma_i$  are positive. The kernel of  $A$  is exactly the space which  $\Sigma$  sends to  $\mathbf{0}$  - these are the coordinates that  $\Sigma$  discards. Therefore, the kernel is spanned by  $\mathbf{v}_{m+1}, \dots, \mathbf{v}_n$ , and

its orthogonal complement is spanned by  $\mathbf{v}_1, \dots, \mathbf{v}_m$ . The image subspace of  $A$  here will be simply  $\mathbb{R}^m$ .

In the case of  $m \geq n$  (again assuming all singular values are positive), the kernel is trivial - it's  $\{\mathbf{0}\}$ . The image subspace in this case would be  $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ .

### Alternative Definition

When viewed in a purely algebraic sense, any zero rows and columns of the matrix  $\Sigma$  are superfluous. We can therefore decompose into the type of decomposition we have seen in class:  $A = U\Sigma V^T$ , with  $A$  of size  $m \times n$ , and  $U, \Sigma, V$  of sizes  $m \times m, n \times n, n \times n$ .

### 9.1.2 Application: Pseudo-Inverse

Let  $A$  be a full-rank  $m \times n$  matrix, now with  $m > n$ ,  $\mathbf{x} \in \mathbb{R}^n$  and  $A\mathbf{x} = \mathbf{b} \in \mathbb{R}^m$ . Given  $\mathbf{b}$ , we want to recover  $\mathbf{x}$ . The pseudo-inverse,  $A^+$ , is the transformation matrix for which  $A^+\mathbf{b} = \mathbf{x}$ .

What would have been the solution, if we were looking "at the right bases" - if both  $V, U$  were the identity matrices? The transformation would be

$$\Sigma = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \\ & & & \mathbf{0} \end{pmatrix}$$

The pseudo-inverse matrix, in the "right" bases, would simply be:

$$\Sigma^+ = \begin{pmatrix} \sigma_1^{-1} & & & \\ & \ddots & & \\ & & & \\ & & & \sigma_n^{-1} \\ & & & & \mathbf{0} \end{pmatrix}$$

We want to achieve that doing only matrix multiplications, because that carries over when the input and output spaces are not the standard bases. We first look at  $A^T A = V\Sigma^T \Sigma V^T$ . Then:

$$\Sigma^T \Sigma = \begin{pmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_n^2 \end{pmatrix}$$

Inverting, we get  $(A^T A)^{-1} = V(\Sigma^T \Sigma)^{-1} V^T$ . Then:

$$(\Sigma^T \Sigma)^{-1} = \begin{pmatrix} \sigma_1^{-2} & & \\ & \ddots & \\ & & \sigma_n^{-2} \end{pmatrix}$$

We can easily see that  $(\Sigma^T \Sigma)^{-1} \Sigma^T = \Sigma^+$  as defined above. This gives us the formula and intuition for the *pseudo-inverse* formula:

$$A^+ = (A^T A)^{-1} A^T$$

Indeed,  $A^+ \mathbf{b} = (A^T A)^{-1} A^T A \mathbf{x} = \mathbf{x}$ .

### 9.1.3 Application: Projection Matrix

Let  $A$  be a full-rank  $m \times n$  matrix ( $m > n$ ). We wish to calculate the projection matrix that projects a vector to the column space of  $A$ . We may not assume the columns of  $A$  are orthonormal.

Again - what would have been the solution, if we were looking "at the right bases" - if both  $V, U$  were the identity matrices? Then the transformation would be

$$\Sigma = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \\ & & & \mathbf{0} \end{pmatrix}$$

The column space of  $\Sigma$  is the space spanned by the first  $m$  standard vectors. The projection matrix, in the "right" bases, would simply be:

$$\begin{pmatrix} 1 & & & \\ & \ddots & & \vdots \\ & & 1 & \\ & \dots & & \mathbf{0} \end{pmatrix}$$

We can see that this is achieved by  $\Sigma \Sigma^+$ . This gives us the formula:

$$P = A(A^T A)^{-1} A^T$$

These two last examples will play a prominent role in linear regression.

### 9.1.4 Application: Low Rank Approximations

In this application we begin with an  $m \times n$  matrix  $A$  of numerical data, and our goal is to describe a close approximation to  $A$  using many fewer numbers than the  $mn$  original entries. The matrix is not considered as a linear transformation, or indeed as an algebraic object at all. It is simply a table of  $mn$  numbers and we would like to find an approximation that captures the most significant features of the data. Another common example is when we

observe a real-life matrix which is supposed to be low rank, but in fact is full rank due to noise. We want to recover the real, low-rank signal.

We therefore wish to find the best low-rank approximation to our matrix  $A$ . Let us define this properly. The *Frobenius norm*  $\|A\|_F$  of a matrix  $A$ , is defined by:

$$\|A\|_F^2 = \sum_{i=1}^m \sum_{j=1}^n A_{ij}^2$$

We want to find the best rank  $r$  approximation  $A_r$  to  $A$ , such that the Frobenius norm of the difference  $\|A_r - A\|_F$  would be minimal.

It can be easily seen (exercise) that for orthonormal matrices  $U, V$ ,

$$\|UA\|_F^2 = \|AV^T\|_F^2 = \|A\|_F^2$$

Therefore,

$$\|A\|_F^2 = \|U\Sigma V^T\|_F^2 = \|\Sigma\|_F^2 = \sum_{i=1}^m \sigma_i^2$$

It is therefore quite intuitive (although we will not prove it here), that the best rank  $r$  approximation would be:  $A_r = U\Sigma_r V^T$ , where

$$\Sigma_r = \begin{pmatrix} \sigma_1 & & & & & \\ & \ddots & & & & \vdots \\ & & \sigma_r & & & \\ & & & 0 & & \mathbf{0} \\ & & & & \ddots & \vdots \\ & & & & & 0 \end{pmatrix}$$

Namely, replacing the last  $m - r$  singular values with 0-s.