Mathematical Toolkit **Autumn 2021**

Lecture 8: October 21, 2021

Lecturer: Madhur Tulsiani

1 Applications of SVD: least squares approximation

We discuss another application of singular value decomposition (SVD) of matrices. Let $a_1, \ldots, a_n \in \mathbb{R}^d$ be points which we want to fit to a low-dimensional subspace. The goal is to find a subspace S of $\mathbb{R}^d$ of dimension at most k to minimize $\sum_{i=1}^n (\mathrm{dist}(a_i, S))^2$, where $\mathrm{dist}(a_i, S)$ denotes the distance of a_i from the closest point in S. We first prove the following.

Claim 1.1 *Let $u_1, \ldots, u_k$ be an orthonormal basis for S. Then*

$$(\mathrm{dist}(a_i, S))^2 \;=\; \|a_i\|_2^2 - \sum_{j=1}^{k} \langle a_i, u_j \rangle^2 .$$

Proof: Complete $u_1, \ldots, u_k$ to an orthonormal basis $u_{k+1}, \ldots, u_d$ for all of $\mathbb{R}^d$. For any point $v \in \mathbb{R}^d$, where exist $c_1, \ldots, c_d \in \mathbb{R}$ such that $v = \sum_{j=1}^d c_j \cdot u_j$. To find the distance $\mathrm{dist}(v, S) = \min_{u \in S} \|v - u\|$, we need to find the point $u \in S$, which is closest to v.

Let $u = \sum_{j=1}^k b_j \cdot u_k$ be an arbitrary point in S (any $u \in S$ can be written in this form, since $u_1, \ldots, u_k$ form a basis for S). We have that

$$\|v - u\|^2 \;=\; \left\| \sum_{j=1}^{k} (c_j - b_j) \cdot u_j + \sum_{j=k+1}^{d} c_j \cdot u_j \right\|^2 \;=\; \sum_{j=1}^{k} (c_j - b_j)^2 + \sum_{j=k+1}^{d} c_j^2 ,$$

which is minimized when $b_j = c_j$ for all $j \in [k]$. Thus, the cloest point $u \in S$ to $v = \sum_{j=1}^d c_j \cdot u_j$ is given by $u = \sum_{j=1}^k c_j \cdot u_j$, with $v - u = \sum_{j=k+1}^d c_j \cdot u_j$. Moreover, since $u_1, \ldots, u_d$ form an *orthonormal* basis, we have $c_j = \langle u_j, v \rangle$ for all $j \in [d]$, which gives

$$\|v - u\|^2 \;=\; \sum_{j=k+1}^{d} c_j^2 \;=\; \sum_{j=1}^{d} c_j^2 - \sum_{j=1}^{k} c_j^2 \;=\; \|v\|^2 - \sum_{j=1}^{k} \langle u_j, v \rangle^2 .$$

Using the above for each a_i (as the point v) completes the proof. ■

Thus, the goal is to find a set of k orthonormal vectors $u_1, \ldots, u_k$ to maximize the quantity $\sum_{i=1}^{n} \sum_{j=1}^{k} \left\langle a_i, u_j \right\rangle^2$. Let $A \in \mathbb{R}^{n \times d}$ be a matrix with the i^{th} row equal to a_i^T. Then, we need to find orthonormal vectors $u_1, \ldots, u_k$ to maximize $\|Au_1\|_2^2 + \cdots + \|Au_k\|_2^2$. We will prove the following.

Proposition 1.2 *Let $v_1, \ldots, v_r$ be the right singular vectors of A corresponding to singular values $\sigma_1 \geq \cdots \geq \sigma_r > 0$. Then, for all $k \leq r$ and all orthonormal sets of vectors $u_1, \ldots, u_k$*

$$\|Au_1\|_2^2 + \cdots + \|Au_k\|_2^2 \ \leq \ \|Av_1\|_2^2 + \cdots + \|Av_k\|_2^2$$

Thus, the optimal solution is to take $S = \text{Span}\,(v_1, \ldots, v_k)$. We prove the above by induction on k. For $k = 1$, we note that

$$\|Au_1\|_2^2 \ = \ \left\langle A^T A u_1, u_1 \right\rangle \ \leq \ \max_{v \in \mathbb{R}^d \setminus \{0\}} \mathcal{R}_{A^T A}(v) \ = \ \sigma_1^2 \ = \ \|Av_1\|_2^2 \ .$$

To prove the induction step for a given $k \leq r$, define

$$V_{k-1}^{\perp} = \left\{ v \in \mathbb{R}^d \mid \langle v, v_i \rangle = 0 \ \ \forall i \in [k-1] \right\} .$$

First prove the following claim.

Claim 1.3 *Given an orthonormal set $u_1, \ldots, u_k$, there exist orthonormal vectors $u_1', \ldots, u_k'$ such that*

- $u_k' \in V_{k-1}^{\perp}$.
- $\text{Span}\,(u_1, \ldots, u_k) \ = \ \text{Span}\,(u_1', \ldots, u_k')$.
- $\|Au_1\|_2^2 + \cdots + \|Au_k\|_2^2 \ = \ \|Au_1'\|_2^2 + \cdots + \|Au_k'\|_2^2$.

Proof: We only provide a sketch of the proof here. Let $S = \text{Span}\,(\{u_1, \ldots, u_k\})$. Note that $\dim\left(V_{k-1}^{\perp}\right) = d - k + 1$ (why?) and $\dim(S) = k$. Thus,

$$\dim\left(V_{k-1}^{\perp} \cap S\right) \ \geq \ k + (d - k + 1) - d \ = \ 1 \,.$$

Hence, there exists $u_k' \in V_{k-1}^{\perp} \cap S$ with $\|u_k'\| = 1$. Completing this to an orthonormal basis of S gives orthonormal $u_1', \ldots, u_k'$ with the first and second properties. We claim that this already implies the third property (why?). ■

Thus, we can assume without loss of generality that the given vectors $u_1, \ldots, u_k$ are such that $u_k \in V_{k-1}^{\perp}$. Hence,

$$\|Au_k\|_2^2 \;\leq\; \max_{\substack{v \in V_{k-1}^{\perp} \\ \|v\|=1}} \|Av\|_2^2 \;=\; \sigma_k^2 \;=\; \|Av_k\|_2^2 \,.$$

Also, by the inductive hypothesis, we have that

$$\|Au_1\|_2^2 + \cdots + \|Au_{k-1}\|_2^2 \;\leq\; \|Av_1\|_2^2 + \cdots + \|Av_{k-1}\|_2^2 \,,$$

which completes the proof. The above proof can also be used to prove that SVD gives the best rank k approximation to the matrix A in Frobenius norm. We will see this in the next homework.

2 Bounding the eigenvalues: Gershgorin Disc Theorem

We will now see a simple but extremely useful bound on the eigenvalues of a matrix, given by the Gershgorin disc theorem. Many useful variants of this bound can also be derived from the observation that for any invertible matrix S, the matrices $S^{-1}MS$ and M have the same eigenvalues (prove it!).

Theorem 2.1 (Gershgorin Disc Theorem) *Let $M \in \mathbb{C}^{n \times n}$. Let $R_i = \sum_{j \neq i} |M_{ij}|$. Define the set*

$$\mathrm{Disc}(M_{ii}, R_i) \;:=\; \{z \mid z \in C, |x - M_{ii}| \leq R_i\} \,.$$

If λ is an eigenvalue of M, then

$$\lambda \;\in\; \bigcup_{i=1}^{n} \mathrm{Disc}(M_{ii}, R_i) \,.$$

Proof: Let $x \in \mathbb{C}^n$ be an eigenvector corresponding to the eigenvalue λ. Let $i_0 = \mathrm{argmax}_{i \in [n]}\{|x_i|\}$. Since x is an eigenvector, we have

$$Mx = \lambda x \quad \Rightarrow \quad \forall i \in [n] \;\; \sum_{j=1}^{n} M_{ij} z_j \;=\; \lambda x_i \,.$$

In particular, we have that for $i = i_0$,

$$\sum_{j=1}^{n} M_{i_0 j} x_j \;=\; \lambda x_{i_0} \;\;\Rightarrow\;\; \sum_{j=1}^{n} M_{i_0 j} \frac{x_j}{x_{i_0}} \;=\; \lambda \;\;\Rightarrow\;\; \sum_{j \neq i_0} M_{i_0 j} \frac{x_j}{x_{i_0}} \;=\; \lambda - M_{i_0 i_0} \,.$$

Thus, we have

$$|\lambda - M_{i_0 i_0}| \;\leq\; \sum_{j \neq i_0} |M_{i_0 j}| \cdot \left|\frac{x_j}{x_{i_0}}\right| \;\leq\; \sum_{j \neq i_0} |M_{i_0 j}| \;=\; R_{i_0} \,.$$

■

2.1 An application to compressed sensing

The Gershgorin disc theorem is quite useful in compressed sensing, to ensure what is known as the "Restricted Isometry Property" for the measurement matrices.

Definition 2.2 *A matrix $A \in \mathbb{R}^{k \times n}$ is said to have the* restricted isometry property *with parameters (s, δ_s) if*

$$(1 - \delta_s) \cdot \|x\|^2 \;\leq\; \|Ax\|^2 \;\leq\; (1 + \delta_s) \cdot \|x\|^2$$

for all $x \in \mathbb{R}^n$ which satisfy $|\{i \mid x_i \neq 0\}| \leq s$.

Thus, we want the transformation A to be approximately norm preserving for all *sparse* vectors x. This can of course be ensured for all x by taking $A = \mathsf{id}$, but we require $k \ll n$ for the applications in compressed sensing. In general, the restricted isometry property is NP-hard to verify and can thus also be hard to reason about for a given matrix. The Gershgorin Disc Theorem lets us derive a much easier condition which is sufficient to ensure the restricted isometry property.

Definition 2.3 *Let $A \in \mathbb{R}_{k \times n}$ be such that $\left\|A^{(i)}\right\| = 1$ for each column $A^{(i)}$ of A. Define the* coherence *of A as*

$$\mu(A) \;=\; \max_{i \neq j} \left| \left\langle A^{(i)}, A^{(j)} \right\rangle \right| .$$

We will prove the following

Proposition 2.4 *Let $A \in \mathbb{R}^{k \times n}$ be such that $\left\|A^{(i)}\right\| = 1$ for each column $A^{(i)}$ of A. Then, for any s, the matrix A has the restricted isometry property with parameters $(s, (s-1)\mu(A))$.*

Note that the bound becomes meaningless if $s \geq 1 + \frac{1}{\mu(A)}$. However, the above proposition shows that it may be sufficient to bound $\mu(A)$ (which is also easier to check in practice).

Proof: Consider any x such that $|\{i \mid x_i \neq 0\}| \leq s$. Let S denote the support of x i.e., $S = \{i \mid x_i \neq 0\}$. Let A_S denote the $k \times |S|$ submatrix where we only keep the columns corresponding to indices in S. Let x_S denote x restricted to the non-zero entries. Then

$$\|Ax\|^2 \;=\; \|A_S x_S\|^2 \;=\; \left\langle A_S^T A_S x_S, x_S \right\rangle .$$

Thus, it suffices to bound the eigenvalues of the matrix $A_S^T A_S$. Note that $(A_S)_{ij} \;=\; \left\langle A^{(i)}, A^{(j)} \right\rangle$. Thus the diagonal entries are 1 and the off-diagonal entries are bounded by $\mu(A)$ in absolute value. By the Gershgorin Disc Theorem, for any eigenvalue λ of A, we have

$$|\lambda - 1| \;\leq\; (s-1) \cdot \mu(A) .$$

Thus, we have

$$(1-(s-1)\cdot\mu(A))\cdot\|x\|^2 \;\leq\; \|Ax\|^2 \;\leq\; (1+(s-1)\cdot\mu(A))\cdot\|x\|^2\,,$$

as desired. ■

The theorem is also very useful for bounding how much the eigenvalues of matrix change due to a perturbation. We will see an example of this in the homework.