PRINCETON UNIV. F'16 COS 521: ADVANCED ALGORITHM DESIGN

Lecture 12: SVD, Power method, and Planted Graph problems (+ eigenvalues of random matrices)

Lecturer: *Pravesh Kothari* Scribe:*Sanjeev Arora*

Today we continue the topic of low-dimensional approximation to datasets and matrices. Last time we saw the singular value decomposition of matrices.

0.1 SVD computation

Recall this theorem from last time. Note that in Linear Algebra classes this might have been stated as: "Every symmetric matrix can be written as $U^T DU$, where D is a diagonal matrix and U is a matrix with orthonormal columns."

THEOREM 1 (SINGULAR VALUE DECOMPOSITION AND BEST RANK-k-APPROXIMATION) *An $m \times n$ real matrix has $t \leq \min\{m,n\}$ nonnegative real numbers $\sigma_1, \sigma_2, \ldots, \sigma_t$ (called singular values) and two sets of unit vectors $U = \{u_1, u_2, \ldots, u_t\}$ which are in $\Re^m$ and $V = v_1, v_2, \ldots, v_t \in \Re^n$ (all vectors are column vectors) where U, V are orthogonormal sets and*

$$u_i^T M = \sigma_i v_i \qquad \text{and} \qquad M v_i = \sigma_i u_i^T. \tag{1}$$

(When M is symmetric, each $u_i = v_i$ and the σ_i's are eigenvalues and can be negative.) Furthermore, M can be represented as

$$M = \sum_i \sigma_i u_i v_i^T. \tag{2}$$

The best rank k approximation to M consists of taking the first k terms of (2) and discarding the rest (where $\sigma_1 \geq \sigma_2 \cdots \geq \sigma_r$).

Taking the best rank k approximation is also called *Principal Component Analysis* or PCA.

You probably have seen eigenvalue and eigenvector computations in your linear algebra course, so you know how to compute the PCA for symmetric matrices. The nonsymmetric case reduces to the symmetric one by using the following observation. If M is the matrix in (2) then

$$MM^T = (\sum_i \sigma_i u_i v_i^T)(\sum_i \sigma_i v_i u_i^T) \quad = \sum_i \sigma_i^2 u_i u_i^T \quad \text{since } v_i^T v_j = 1 \text{ iff } i = j \text{ and 0 else.}$$

Thus we can recover the u_i's and σ_i's by computing the eigenvalues and eigenvectors of MM^T, and then recover v_i by using (1).

Another application of singular vectors is the *Pagerank* algorithm for ranking webpages.

0.1.1 The power method

The eigenvalue computation you saw in your linear algebra course takes at least n^3 time. Often we are only interested in the top few eigenvectors, in which case there's a method that can work much faster (especially when the matrix is *sparse*, i.e., has few nonzero entries).

As usual, we first look at the subcase of symmetric matrices. To compute the largest eigenvector of matrix M we do the following. Pick a random unit vector x. Then repeat the following a few times: replace x by Mx. We show this works under the following *Gap assumption*: There is a *gap* of γ between the the top two eigenvalues: $|\lambda_1| - |\lambda_2| = \gamma$.

The analysis is the same calculation as the one we used to analyse Markov chains. We can write x as $\sum_i \alpha_i e_i$ where e_i's are the eigenvectors and λ_i's are numbered in decreasing order by absolute value. Then t iterations produces $M^t x = \sum_i \alpha_i \lambda_i^t e_i$. Since x is a unit vector, $\sum_i \alpha_i^2 = 1$.

Since $|\lambda_i| \leq |\lambda_1| - \gamma$ for $i \geq 2$, we have

$$\sum_{i \geq 2} |\alpha_i| \left|\lambda_i^t\right| \leq n\alpha_{\max}(|\lambda_1| - \gamma)^t \quad = n \left|\lambda_1\right|^t (1 - \gamma/\left|\lambda_1\right|)^t,$$

where $\alpha_{\max}$ is the largest coefficient in magnitude.

Furthermore, since x was a random unit vector (and recalling that its projection α_1 on the fixed vector e_1 is normally distributed), the probability is at least 0.99 that $\alpha_1 > 1/(10n)$. Thus setting $t = O(\log n \left|\lambda_1\right| / \gamma)$ the components for $i \geq 2$ become miniscule and $x \approx \alpha_1 \left|\lambda_1\right|^t e_1$. Thus rescaling to make it a unit vector, we get e_1 up to some error. Then we can project all vectors to the subspace perpendicular to e_1 and continue with the process to find the remaining eigenvectors and eigenvalues.

This process works under the above gap assumption. What if the gap assumption does not hold? Say, the first 3 eigenvalues are all close together, and separated by a gap from the fourth. Then the above process ends up with some random vector in the subspace spanned by the top three eigenvectors. For real-life matrices the gap assumption often holds.

Interpretation as message passing. If the matrix represents a weighted graph, then the power method can be interpreted as a *message passing* algorithm. Every node gets a random initial value with the ith node getting x_i. At each step each node takes the weighted average of its neighbors, where the weight is given by the weight on the edge connecting them. This is nothing but computing $M \cdot x$ in a distributed fashion.

This distributed view of the power method has been used in many settings, including in the original algorithms for ranking web-pages.

Also, one can use other update rules than $x \leftarrow Mx$ to obtain a host of other distributed algorithms. *Belief propagation* is a famous one.

0.2 Recovering planted bisections

Now we return to the planted bisection problem, also introduced last time. The method we show is called the *spectral method*, since it uses the *spectrum*, i.e., eigenvalue/singular value information. We also mentioned last time that the generalization of this model to graphs with k parts is called *Stochastic Block Model*.

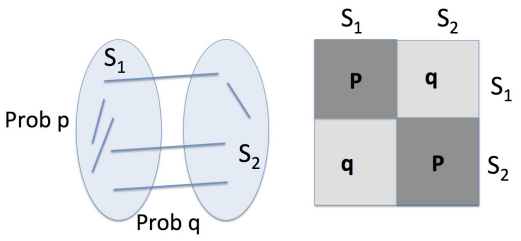

Figure 1: Planted Bisection problem: Edge probability is p within S_1, S_2 and q between S_1, S_2 where $q < p$. On the right hand side is the adjacency matrix. If we somehow knew S_1, S_2 and grouped the corresponding rows and columns together, and squint at the matrix from afar, we'd see more density of edges within S_1, S_2 and less density between S_1, S_2. Thus from a distance the adjacency matrix looks like a rank 2 matrix.

The observation in Figure 1 suggests that the adjacency matrix is close to a rank 2 matrix shown there: the block within S_1, S_2 have value p in each entry; the blocks between S_1, S_2 have q in each entry. This is rank 2 since it has only two distinct column vectors.

Now we sketch why the best rank-2 approximation to the adjacency matrix will more or less recover the planted bisection. Specifically, the idea is to find the rank 2 approximation; with very high probability its columns can be cleanly clustered into 2 clusters. This gives a grouping of the vertices into 2 groups as well, which turns out to be the planted bisection.

Why this works has to do with the properties of rank k approximations. First we define two norms of a matrix.

Definition 1 (Frobenius and spectral norm) *If M is an $n \times n$ matrix then its* Frobenius norm $|M|_F$ *is* $\sqrt{\sum_{ij} M_{ij}^2}$ *and its* spectral norm $|M|_2$ *is the maximum value of $|Mx|_2$ over all unit vectors $x \in \Re^n$. (By Courant-Fisher, the spectral norm is also the highest eigenvalue.) For matrices that are not symmetric the definition of Frobenius norm is analogous and the spectral norm is the highest singular value.*

Last time we defined the *best rank k approximation to M* as the matrix $\tilde{M}$ that is rank k and minimizes $\left|M - \tilde{M}\right|_F^2$. The following theorem shows that we could have defined it equivalently using spectral norm.

Lemma 2
Matrix $\tilde{M}$ as defined above also satisfies that $\left|M - \tilde{M}\right|_2 \leq |M - B|_2$ for all B that have rank k.

Theorem 3
If $\tilde{M}$ is the best rank-k approximation to M, then for every rank k matrix C:

$$\left|\tilde{M} - C\right|_F^2 \leq 5k\,|M - C|_2^2\,.$$

PROOF: Follows by Spectral decomposition and Courant-Fisher theorem, and the fact that the column vectors in $\tilde{M}$ and C together span a space of dimension at most $2k$. Thus $\left|\tilde{M}-C\right|_F^2$ involves a matrix of rank at most $2k$. Rest of the details are cut and pasted from Hopcroft-Kannan in Figure 2.

□

Returning to planted graph bisection, let M be the adjacency matrix of the graph with planted bisection. Let C be the rank-2 matrix that we *think* is a good approximation to M, namely, the one in Figure 1. Let $\tilde{M}$ be the true rank 2 approximation found via SVD. In general $\tilde{M}$ is not the same as C. But Theorem 3 implies that we can upper bound the average coordinate-wise squared difference of $\tilde{M}$ and C by the quantity on the right hand side, which is the spectral norm (i.e., largest eigenvalue) of $M-C$.

Notice, $M-C$ is a *random matrix* whose each coordinate is one of four values $1-p, -p, 1-q, -q$. More importantly, the expectation of each coordinate is 0 (since the entry of M is a coin toss whose expected value is the corresponding entry of C). The study of eigenvalues of such random matrices is a famous subfield of science with unexpected connections to number theory (including the famous Riemann hypothesis), quantum physics (quantum gravity, quantum chaos), etc. We show below that $|M-C|_2^2$ is at most $O(np)$. We conclude that the average column vector in $\tilde{M}$ and C (whose square norm is about np) are apart by $O(p)$. Thus intuitively, clustering the columns of C into two will find us the bipartition. Actually showing this requires more work which we will not do.

Here is a generic clustering algorithm into two clusters: Pick a random column of $\tilde{M}$ and put into one cluster all columns whose distance from it is at most $10p$. Put all other columns in the other cluster.

0.2.1 Eigenvalues of random matrices

We sketch a proof of the following classic theorem to give a taste of this beautiful area.

THEOREM 4
Let R be a random matrix such that R_{ij}'s are independent random variables in $[-1,1]$ of expectation 0 and variance at most σ^2. Then with probability $1-\exp(-n)$ the largest eigenvalue of R is at most $O(\sigma\sqrt{n})$.

For simplicity we prove this for $\sigma = 1$.
PROOF: Recalling that the largest eigenvalue is $\max_{x:|x|_2=1}\left|x^TRx\right|$, we hope to use something like Chernoff bound. The difficulty is that the set of unit vectors is uncountably infinite, so we cannot use the union bound. We break the proof as follows.

Idea 1) *For any fixed unit vector $x \in \Re^n$, $\left|x^TRx\right| \leq O(\sqrt{n})$ with probability $1-\exp(-Cn)$ where C is an arbitrarily large constant.* This follows from Chernoff-type bounds. Note that $x^TRx = \sum_{ij} R_{ij}x_ix_j$. By Chernoff bounds (specifically, Hoeffding's inequality) the probability that this exceeds t is at most

$$\exp(-\frac{t^2}{\sum_i x_i^2 x_j^2}) \quad \leq \exp(-\Omega(t^2)),$$

since $(\sum_{ij} x_i^2x_j^2)^{1/2} \leq \sum_i x_i^2 = 1$.

Idea 2) *There is a set of* $\exp(n)$ *special directions* $x_{(1)}, x_{(2)}, \ldots,$ *that approximately "cover"the set of unit vectors in* $\Re^n$. Namely, for every unit vector v, there is at least one $x_{(i)}$ such that $< v, x_{(i)} > > 0.9$. This is an example of the so-called ϵ-*net method* a standard way to deal with the difficulty that the set of unit vectors is uncountably infinite. Instead of trying to apply the union bound to all unit vectors, we apply it to just the set of special directions, which is finite and approximates the set of all unit vectors.

First, note that $< v, x^{(i)} > > 0.9$ iff

$$\left|v - x_{(i)}\right|^2 = |v|^2 + \left|x_{(i)}\right|^2 - 2 < v, x_{(i)} > \quad \leq 0.2.$$

In other words we are trying to cover the unit sphere with spheres of radius 0.2.

Try to pick this set greedily. Pick $x_{(1)}$ arbitrarily, and throw out the unit sphere of radius 0.2 around it. Then pick $x_{(2)}$ arbitrarily out of the remaining sphere, and throw out the unit sphere of radius 0.2 around it. And so on.

How many points did we end up with? By construction, each point that was picked has distance at least 0.2 from every other point that was picked, so the spheres of radius 0.1 around the picked points are mutually disjoint. Thus the maximum number of points we could have picked is the number of disjoint spheres of radius 0.1 in a ball of radius at most 1.1. Denoting by $B(r)$ denote the volume of spheres of volume r, this is at most $B(1.1)/B(0.1) = \exp(n)$.

Idea 3) Combining Ideas 1 and 2, and the union bound, we have with high probability, $\left|x_{(i)}^T R x_{(i)}\right| \leq O(\sqrt{n})$ for all the special directions.

Idea 4): *If* v *is the eigenvector corresponding to the largest eigenvalue satisfies then there is some special direction satisfying* $\left|x_{(i)}^T R x_{(i)}\right| > 0.4 v^T R v$.

This is saying that the set of special directions is some kind of *neighborhood watch* looking for crime (i.e., a *bad event*). If a bad event happens for any unit vector at all, one of the special directions in that neighborhood will notice something almost as bad.

By the covering property, there is some special direction $x_{(i)}$ that is close to v. Represent it as $\alpha v + \beta u$ where $u \perp v$ and u is a unit vector. So $\alpha \geq 0.9$ and $\beta \leq \sqrt{0.19} \leq 0.5$. Then $\left|x_{(i)}^T R x_{(i)}\right| = \alpha v^T R v + \beta u^T R u$. But v is the largest eigenvalue so $\left|u^T R u\right| \leq v^T R v$. We conclude $\left|x_{(i)}^T R x_{(i)}\right| \geq (0.9 - 0.5) v^T R v$, as claimed.

The theorem now follows from Idea 3 and 4. □

BIBLIOGRAPHY

1. F. McSherry. *Spectral partitioning of random graphs.* In IEEE FOCS 2001 Proceedings.

2. Relevant chapter of Hopcroft-Kannan.

3. T. Tao. *Topics in random matrix theory.*

Lemma 8.7 *Suppose A is an $n \times d$ matrix and suppose C is an $n \times d$ rank k matrix. Let $\bar{A}$ be the best rank k approximation to A found by SVD. Then, $||\bar{A} - C||_F^2 \leq 5k||A - C||_2^2$.*

Proof: Let $\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_k}$ be the top k singular vectors of A. Extend the set of the top k singular vectors to an orthonormal basis $\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_p}$ of the vector space spanned by the rows of $\bar{A}$ and C. Note that $p \leq 2k$ since $\bar{A}$ is spanned by $\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_k}$ and C is of rank at most k. Then,

$$||\bar{A} - C||_F^2 = \sum_{i=1}^{k} |(\bar{A} - C)\mathbf{u_i}|^2 + \sum_{i=k+1}^{p} |(\bar{A} - C)\mathbf{u_i}|^2.$$

Since $\{\mathbf{u_i}| 1 \leq i \leq k\}$ are the top k singular vectors of A and since $\bar{A}$ is the rank k approximation to A, for $1 \leq i \leq k$, $A\mathbf{u_i} = \bar{A}\mathbf{u_i}$ and thus $\left|\left(\bar{A} - C\right)\mathbf{u_i}\right|^2 = |(A - C)\mathbf{u_i}|^2$. For $i > k$, $\bar{A}\mathbf{u_i} = 0$, thus $\left|\left(\bar{A} - C\right)\mathbf{u_i}\right|^2 = |C\mathbf{u_i}|^2$. From this it follows that

$$\begin{aligned}\left\|\bar{A} - C\right\|_F^2 &= \sum_{i=1}^{k} |(A - C)\mathbf{u_i}|^2 + \sum_{i=k+1}^{p} |C\mathbf{u_i}|^2 \\ &\leq k||A - C||_2^2 + \sum_{i=k+1}^{p} |A\mathbf{u_i} + (C - A)\mathbf{u_i}|^2\end{aligned}$$

Using $|a + b|^2 \leq 2|a|^2 + 2|b|^2$

$$\begin{aligned}\left\|\bar{A} - C\right\|_F^2 &\leq k||A - C||_2^2 + 2\sum_{i=k+1}^{p} |A\mathbf{u_i}|^2 + 2\sum_{i=k+1}^{p} |(C - A)\mathbf{u_i}|^2 \\ &\leq k||A - C||_2^2 + 2\left(p - k - 1\right)\sigma_{k+1}^2(A) + 2\left(p - k - 1\right)||A - C||_2^2\end{aligned}$$

Using $p \leq 2k$ implies $k > p - k - 1$

$$\left\|\bar{A} - C\right\|_F^2 \leq k||A - C||_2^2 + 2k\sigma_{k+1}^2(A) + 2k||A - C||_2^2. \tag{8.1}$$

As we saw in Chapter 4, for any rank k matrix B, $||A - B||_2 \geq \sigma_{k+1}(A)$ and so $\sigma_{k+1}(A) \leq ||A - C||_2$ and plugging this in, we get the Lemma. ∎

Figure 2: Proof of Theorem 3 from Hopcroft-Kannan book