Long Term Memory in AI - Vector Search and Databases **COS 597A Fall 2023**

Class 5 - Dimensionality Reduction

Lectures: Lectures: Edo Liberty and Matthijs Douze

Warning: *Please do not cite this note as a peer reviewed source. Please submit requests and corrections as issues or pull requests at github.com/edoliberty/vector-search-class-notes*

1 Singular Value Decomposition (SVD)

We will see that any matrix $A \in \mathbb{R}^{m \times n}$ (w.l.o.g. $m \leq n$) can be written as

$$A = \sum_{\ell=1}^{m} \sigma_\ell u_\ell v_\ell^T \quad (1)$$

$$\forall\ \ell \quad \sigma_\ell \in \mathbb{R},\ \ \sigma_\ell \geq 0 \quad (2)$$

$$\forall\ \ell, \ell' \quad \langle u_\ell, u_{\ell'} \rangle = \langle v_\ell, v_{\ell'} \rangle = \delta(\ell, \ell') \quad (3)$$

To prove this consider the matrix $AA^T \in \mathbb{R}^{m \times m}$. Set u_ℓ to be the ℓ'th eigenvector of AA^T. By definition we have that $AA^T u_\ell = \lambda_\ell u_\ell$. Since AA^T is positive semidefinite we have $\lambda_\ell \geq 0$. Since AA^T is symmetric we have that $\forall\ \ell, \ell'$ $\langle u_\ell, u_{\ell'} \rangle = \delta(\ell, \ell')$. Set $\sigma_\ell = \sqrt{\lambda_\ell}$ and $v_\ell = \frac{1}{\sigma_\ell} A^T u_\ell$. Now we can compute the following:

$$\langle v_\ell, v_{\ell'} \rangle = \frac{1}{\sigma_\ell^2} u_\ell^T A A^T u_{\ell'} = \frac{1}{\sigma_\ell^2} \lambda_\ell \langle u_\ell, u_{\ell'} \rangle = \delta(\ell, \ell')$$

We are only left to show that $A = \sum_{\ell=1}^{m} \sigma_\ell u_\ell v_\ell^T$. To do that consider the test vector $w = \sum_{i=1}^{m} \alpha_i u_i$.

$$\begin{aligned} ||w^T(A - \sum_{\ell=1}^{m} \sigma_\ell u_\ell v_\ell^T)|| &= ||(\sum_{i=1}^{m} \alpha_i u_i^T)(A - \sum_{\ell=1}^{m} \sigma_\ell u_\ell v_\ell^T)|| \\ &= ||(\sum_{i=1}^{m} \alpha_i u_i^T A - \sum_{i=1}^{m} \sum_{\ell=1}^{m} \delta(i,\ell) \alpha_i \sigma_\ell v_\ell^T|| \\ &= ||(\sum_{i=1}^{m} \alpha_i \sigma_i v_i^T - \sum_{i=1}^{m} \alpha_i \sigma_i v_i^T|| = 0 \end{aligned}$$

The vectors u_ℓ and v_ℓ are called the left and right singular vectors of A and σ_ℓ are the singular vectors of A. It is customary to order the singular values in descending order $\sigma_1 \geq \sigma_2, \ldots, \sigma_m \geq 0$. Also, we will denote by r the rank of A. Here is another very convenient way to write the fact that $A = \sum_{\ell=1}^{m} \sigma_\ell u_\ell v_\ell^T$

- Let $\Sigma \in \mathbb{R}^{r \times r}$ be a diagonal matrix whose entries are $\Sigma(i,i) = \sigma_i$ and $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r$.
- Let $U \in \mathbb{R}^{m \times r}$ be the matrix whose i'th column is the left singular vectors of A corresponding to singular value σ_i.
- Let $V \in \mathbb{R}^{n \times r}$ be the matrix whose i'th column is the right singular vectors of A corresponding to singular value σ_i.

We have that $A = USV^T$ and that $U^T U = V^T V = I_r$. Note that the sum goes only up to r which is the rank of A. Clearly, not summing up zero valued singular values does not change the sum.

Applications of the SVD

1. Determining range, null space and rank (also numerical rank).

2. Matrix approximation.

3. Inverse and Pseudo-inverse: If $A = U\Sigma V^T$ and Σ is full rank, then $A^{-1} = V\Sigma^{-1}U^T$. If Σ is singular, then its pseudo-inverse is given by $A^\dagger = V\Sigma^\dagger U^T$, where $\Sigma^\dagger$ is formed by replacing every nonzero entry by its reciprocal.

4. Least squares: If we need to solve $Ax = b$ in the least-squares sense, then $x_{LS} = V\Sigma^\dagger U^T b$.

5. De-noising – Small singular values typically correspond to noise. Take the matrix whose columns are the signals, compute SVD, zero small singular values, and reconstruct.

6. Compression – We have signals as the columns of the matrix S, that is, the i signal is given by

$$S_i = \sum_{i=1}^{r} (\sigma_j v_{ij})\, u_j.$$

   If some of the σ_i are small, we can discard them with small error, thus obtaining a compressed representation of each signal. We have to keep the coefficients $\sigma_j v_{ij}$ for each signal and the dictionary, that is, the vectors u_i that correspond to the retained coefficients.

SVD and eigen-decomposition are related but there are quite a few differences between them.

1. Not every matrix has an eigen-decomposition (not even any square matrix). Any matrix (even rectangular) has an SVD.

2. In eigen-decomposition $A = X\Lambda X^{-1}$, that is, the eigen-basis is not always orthogonal. The basis of singular vectors is always orthogonal.

3. In SVD we have two singular-spaces (right and left).

4. Computing the SVD of a matrix is more numerically stable.

Rank-k approximation in the spectral norm

The following will claim that the best approximation to A by a rank deficient matrix is obtained by the top singular values and vectors of A. More accurately:

Fact 1.1. *Set*

$$A_k = \sum_{j=1}^{k} \sigma_j u_j v_j^T.$$

Then,

$$\min_{\substack{B\in\mathbb{R}^{m\times n}\\ \mathrm{rank}(B)\le k}} \|A - B\|_2 = \|A - A_k\|_2 = \sigma_{k+1}.$$

Proof.

$$\|A - A_k\| = \|\sum_{j=1}^{r} \sigma_j u_j v_j^T - \sum_{j=1}^{k} \sigma_j u_j v_j^T\| = \|\sum_{j=k+1}^{r} \sigma_j u_j v_j^T\| = \sigma_{k+1}$$

and thus σ_{k+1} is the largest singular value of $A - A_k$. Alternatively, look at $U^T A_k V = \mathrm{diag}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0)$, which means that $\mathrm{rank}(A_k) = k$, and that

$$\|A - A_k\|_2 = \|U^T(A - A_k)V\|_2 = \|\,\mathrm{diag}(0, \ldots, 0, \sigma_{k+1}, \ldots, \sigma_r)\|_2 = \sigma_{k+1}.$$

Let B be an arbitrary matrix with $\operatorname{rank}(B_k) = k$. Then, it has a null space of dimension $n - k$, that is,

$$\operatorname{null}(B) = \operatorname{span}(w_1, \ldots, w_{n-k}).$$

A dimension argument shows that

$$\operatorname{span}(w_1, \ldots, w_{n-k}) \cap \operatorname{span}(v_1, \ldots, v_{k+1}) \neq \{0\}.$$

Let w be a unit vector from the intersection. Since

$$Aw = \sum_{j=1}^{k+1} \sigma_j (v_j^T w) u_j,$$

we have

$$\|A - B\|_2^2 \geq \|(A - B)w\|_2^2 = \|Aw\|_2^2 = \sum_{j=1}^{k+1} \sigma_j^2 \left|v_j^T w\right|^2 \geq \sigma_{k+1}^2 \sum_{j=1}^{k+1} \left|v_j^T w\right|^2 = \sigma_{k+1}^2,$$

since $w \in \operatorname{span}\{v_1, \ldots, v_{n+1}\}$, and the v_j are orthogonal. □

Rank-k approximation in the Frobenius norm

The same theorem holds with the Frobenius norm.

Theorem 1.1. *Set*

$$A_k = \sum_{j=1}^{k} \sigma_j u_j v_j^T.$$

Then,

$$\min_{\substack{B \in \mathbb{R}^{m \times n} \\ \operatorname{rank}(B) \leq k}} \|A - B\|_F = \|A - A_k\|_F = \sqrt{\sum_{i=k+1}^{m} \sigma_i^2}.$$

Proof. Suppose $A = U\Sigma V^T$. Then

$$\min_{\operatorname{rank}(B) \leq k} \|A - B\|_F^2 = \min_{\operatorname{rank}(B) \leq k} \|U\Sigma V^T - UU^T BVV^T\|_F^2 = \min_{\operatorname{rank}(B) \leq k} \|\Sigma - U^T BV\|_F^2.$$

Now,

$$\|\Sigma - U^T BV\|_F^2 = \sum_{i=1}^{n} \left(\Sigma_{ii} - \left(U^T BV\right)_{ii}\right)^2 + \text{off-diagonal terms.}$$

If B is the best approximation matrix and $U^T BV$ is not diagonal, then write $U^T BV = D + O$, where D is diagonal and O contains the off-diagonal elements. Then the matrix $B = UDV^T$ is a better approximation, which is a contradiction.

Thus, $U^T BV$ must be diagonal. Hence,

$$\|\Sigma - D\|_F^2 = \sum_{i=1}^{n} (\sigma_i - d_i)^2 = \sum_{i=1}^{k} (\sigma_i - d_i)^2 + \sum_{i=k+1}^{n} \sigma_i^2,$$

and this is minimal when $d_i = \sigma_i$, $i = 1, \ldots, k$. The best approximating matrix is $A_k = UDV^T$, and the approximation error is $\sqrt{\sum_{i=k+1}^{n} \sigma_i^2}$. □

2 Linear regression in the least-squared loss

In Linear regression we aim to find the best linear approximation to a set of observed data. For the m data points $\{x_1, \ldots, x_m\}$, $x_i \in \mathbb{R}^n$, each receiving the value y_i, we look for the weight vector w that minimizes:

$$\sum_{i=1}^{n} (x_i^T w - y_i)^2 = \|Aw - y\|_2^2$$

Where A is a matrix that holds the data points as rows $A_i = x_i^T$.

Proposition 2.1. *The vector w that minimizes $\|Aw - y\|_2^2$ is $w = A^\dagger y = V\Sigma^\dagger U^T y$ for $A = U\Sigma V^T$ and $\Sigma^\dagger_{ii} = 1/\Sigma_{ii}$ if $\Sigma_{ii} > 0$ and 0 else.*

Let us define $U_\|$ and $U_\perp$ as the parts of U corresponding to positive and zero singular values of A respectively. Also let $y_\| = 0$ and $y_\perp$ be two vectors such that $y = y_\| + y_\perp$ and $U_\| y_\perp = 0$ and $U_\perp y_\| = 0$.

Since $y_\|$ and $y_\perp$ are orthogonal we have that $\|Aw - y\|_2^2 = \|Aw - y_\| - y_\perp\|_2^2 = \|Aw - y_\|\|_2^2 + \|y_\perp\|_2^2$. Now, since $y_\|$ is in the range of A there is a solution w for which $\|Aw - y_\|\|_2^2 = 0$. Namely, $w = A^\dagger y = V\Sigma^\dagger U^T y$ for $A = U\Sigma V^T$. This is because $U\Sigma V^T V\Sigma^\dagger U^T y = y_\|$. Moreover, we get that the minimal cost is exactly $\|y_\perp\|_2^2$ which is independent of w.

3 PCA, Optimal squared loss dimension reduction

Given a set of n vectors $x_1, \ldots, x_n$ in $\mathbb{R}^m$. We look for a rank k projection matrix $P \in \mathbb{R}^{m\times m}$ that minimizes:

$$\sum_{i=1} ||Px_i - x_i||_2^2$$

If we denote by A the matrix whose i'th column is x_i then this is equivalent to minimizing $||PA - A||_F^2$ Since the best possible rank k approximation to the matrix A is $A_k = \sum_{i=1}^k \sigma_i u_i v_i^T$ the best possible solution would be a projection P for which $PA = A_k$. This is achieved by $P = U_k U_k^T$ where U_k is the matrix corresponding to the first k left singular vectors of A.

If we define $y_i = U_k^T x_i$ we see that the values of $y_i \in \mathbb{R}^k$ are optimally fitted to the set of points x_i in the sense that they minimize:

$$\min_{y_1,\ldots,y_n} \min_{\Psi \in \mathbb{R}^{k\times m}} \sum_{i=1} ||\Psi y_i - x_i||_2^2$$

The mapping of $x_i \to U_k^T x_i = y_i$ thus reduces the dimension of any set of points $x_1, \ldots, x_n$ in $\mathbb{R}^m$ to a set of points $y_1, \ldots, y_n$ in $\mathbb{R}^k$ optimally in the squared loss sense. This is commonly referred to as Principal Component Analysis (PCA).

4 Closest orthogonal matrix

The SVD also allows to find the orthogonal matrix that is closest to a given matrix. Again, suppose that $A = U\Sigma V^T$ and W is an orthogonal matrix that minimizes $\|A - W\|_F^2$ among all orthogonal matrices. Now,

$$\|U\Sigma V^T - W\|_F^2 = \|U\Sigma V^T - UU^T WVV^T\| = \|\Sigma - \tilde{W}\|,$$

where $\tilde{W} = U^T WV$ is another orthogonal matrix. We need to find the orthogonal matrix $\tilde{W}$ that is closest to Σ. Alternatively, we need to minimize $\|\tilde{W}^T\Sigma - I\|_F^2$.

If U is orthogonal and D is diagonal and positive, then

$$\begin{aligned}\operatorname{trace}(UD) &= \sum_{i,k} u_{ik} d_{ki} \le \sum_i \left(\left(\sum_k u_{ik}^2 \right)^{1/2} \left(\sum_k d_{ik}^2 \right)^{1/2} \right) \\ &= \sum_i \left(\sum_k d_{ki}^2 \right)^{1/2} = \sum_i \left(d_{ii}^2 \right)^{1/2} = \sum_i d_{ii} = \operatorname{trace}(D).\end{aligned} \tag{4}$$

Now

$$\begin{aligned}\|\tilde{W}^T \Sigma - I\|_F^2 &= \operatorname{trace}\left(\left(\tilde{W}^T \Sigma - I \right) \left(\tilde{W}^T \Sigma - I \right)^T \right) \\ &= \operatorname{trace}\left(\left(\tilde{W}^T \Sigma - I \right) \left(\Sigma \tilde{W} - I \right) \right) \\ &= \operatorname{trace}\left(\tilde{W}^T \Sigma^2 \tilde{W} \right) - \operatorname{trace}\left(\tilde{W}^T \Sigma \right) - \operatorname{trace}\left(\Sigma \tilde{W} \right) + n \\ &= \operatorname{trace}\left(\left(\Sigma \tilde{W} \right)^T \left(\Sigma \tilde{W} \right) \right) - 2 \operatorname{trace}\left(\Sigma \tilde{W} \right) + n \\ &= \|\Sigma \tilde{W}\|_F^2 - 2 \operatorname{trace}\left(\Sigma \tilde{W} \right) + n \\ &= \|\Sigma\|_F^2 - 2 \operatorname{trace}\left(\Sigma \tilde{W} \right) + n.\end{aligned}$$

Thus, we need to maximize trace $\left(\Sigma \tilde{W} \right)$. But this is maximized by $\tilde{W} = I$ by (4). Thus, the best approximating matrix is $W = UV^T$.

5 Computing the SVD: The power method

We give a simple algorithm for computing the Singular Value Decomposition of a matrix $A \in \mathbb{R}^{m \times n}$. We start by computing the first singular value σ_1 and left and right singular vectors u_1 and v_1 of A, for which $min_{i<j} \log(\sigma_i/\sigma_j) \ge \lambda$:

```
Generate $x_0$ such that $x_0(i) \sim \mathcal{N}(0,1)$.
$s \leftarrow \log(4\log(2n/\delta)/\varepsilon\delta)/2\lambda$
for $i$ in $[1, \ldots, s]$:
    $x_i \leftarrow A^T A x_{i-1}$
$v_1 \leftarrow x_i/\|x_i\|$
$\sigma_1 \leftarrow \|Av_1\|$
$u_1 \leftarrow Av_1/\sigma_1$
return $(\sigma_1, u_1, v_1)$
```

Let us prove the correctness of this algorithm. First, write each vector x_i as a linear combination of the right singular values of A i.e. $x_i = \sum_j \alpha_j^i v_j$. From the fact that v_j are the eigenvectors of $A^T A$ corresponding to eigenvalues σ_j^2 we get that $\alpha_j^i = \alpha_j^{i-1} \sigma_j^2$. Thus, $\alpha_j^s = \alpha_j^0 \sigma_j^{2s}$. Looking at the ratio between the coefficients of v_1 and v_i for x_s we get that:

$$\frac{| < x_s, v_1 > |}{| < x_s, v_i > |} = \frac{|\alpha_1^0|}{|\alpha_i^0|} \left(\frac{\sigma_1}{\sigma_i} \right)^{2s}$$

Demanding that the error in the estimation of σ_1 is less than ε gives the requirement on s.

$$\begin{aligned} \frac{|\alpha_1^0|}{|\alpha_i^0|}\left(\frac{\sigma_1}{\sigma_i}\right)^{2s} &\geq \frac{n}{\varepsilon} \qquad (5)\\ s &\geq \frac{\log(n|\alpha_i^0|/\varepsilon|\alpha^0|_1)}{2\log(\sigma_1/\sigma_i)} \qquad (6)\end{aligned}$$

From the two-stability of the Gaussian distribution we have that $\alpha_i^0 \sim \mathcal{N}(0,1)$. Therefore, $\Pr[\alpha_i^0 > t] \leq e^{-t^2}$ which gives that with probability at least $1-\delta/2$ we have for all i, $|\alpha_i^0| \leq \sqrt{\log(2n/\delta)}$. Also, $\Pr[|\alpha_1^0| \leq \delta/4] \leq \delta/2$ (this is because $\Pr[|z| < t] \leq max_r \Psi_z(r) \cdot 2t$ for any distribution and the normal distribution function at zero takes it maximal value which is less than 2) Thus, with probability at least $1-\delta$ we have that for all i, $\frac{|\alpha_1^0|}{|\alpha_i^0|} \leq \frac{\sqrt{\log(2n/\delta)}}{\delta/4}$. Combining all of the above we get that it is sufficient to set $s = \log(4n\log(2n/\delta)/\varepsilon\delta)/2\lambda = O(\log(n/\varepsilon\delta)/\lambda)$ in order to get ε precision with probability at least $1-\delta$.

We now describe how to extend this to a full SVD of A. Since we have computed (σ_1, u_1, v_1), we can repeat this procedure for $A - \sigma_1 u_1 v_1^T = \sum_{i=2}^{n} \sigma_i u_i v_i^T$. The top singular value and vectors of which are (σ_2, u_2, v_2). Thus, computing the rank-k approximation of A requires $O(mnks) = O(mnk\log(n/\varepsilon\delta))/\lambda)$ operations. This is because computing $A^T Ax$ requires $O(mn)$ operations and for each of the first k singular values and vectors this is performed s times.

The main problem with this algorithm is that its running time is heavily influenced by the value of λ. Other variants of this algorithm are much less sensitive to the value of this parameter, but are out of the scope of this class.

6 Random-projection

We will give a simple proof of the following, rather amazing, fact. Every set of n points in a Euclidian space (say in dimension d) can be embedded into the Euclidean space of dimension $k = O(\log(n)/\varepsilon^2)$ such that all pairwise distances are preserved up distortion $1 \pm \varepsilon$. We will prove the construction of [3] which is simpler than the one in [4].

We will argue that a certain distribution over the choice of a matrix $\mathbb{R} \in \mathbb{R}^{k\times d}$ gives that:

$$\forall x \in \mathbb{S}^{d-1} \;\; \Pr\left[\left| ||\frac{1}{\sqrt{k}}Rx|| - 1\right| > \varepsilon\right] \leq \frac{1}{n^2} \tag{7}$$

Before we pick this distribution and show that Equation 7 holds for it, let us first see that this gives the opening statement.

Consider a set of n points $x_1, \ldots, x_n$ in Euclidean space $\mathbb{R}^d$. Embedding these points into a lower dimension while preserving all distances between them up to distortion $1 \pm \varepsilon$ means approximately preserving the norms of all $\binom{n}{2}$ vectors $x_i - x_j$. Assuming Equation 7 holds and using the union bound, this property will fail to hold for at least one $x_i - x_j$ pair with probability at most $\binom{n}{2}\frac{1}{n^2} \leq 1/2$. Which means that all $\binom{n}{2}$ point distances are preserved up to distortion ε with probability at least $1/2$.

7 Matrices with normally distributed independent entries

We consider the distribution of matrices R such that each $R(i,j)$ is drawn independently from a normal distribution with mean zero and variance 1, $R(i,j) \sim \mathcal{N}(0,1)$. We show that for this distribution Equation 7 holds for some $k \in O(\log(n)/\varepsilon^2)$.

First consider the random variable $z = \sum_{j=1}^{d} r(j)x(j)$ where $r(j) \sim \mathcal{N}(0,1)$. To understand how the variable z distributes we recall the two-stability of the normal distribution. Namely, if $z_3 = z_2 + z_1$ and $z_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ and $z_2 \sim \mathcal{N}(\mu_2, \sigma_2)$ then,

$$z_3 \sim \mathcal{N}(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}).$$

In our case, $r(i)x(i) \sim \mathcal{N}(0, x_i)$ and therefore, $z = \sum_{i=1}^{d} r(i)x(i) \sim \mathcal{N}(0, \sqrt{\sum_{i=1}^{d} x_i^2}) \sim \mathcal{N}(0,1)$. Now, note that each element in the vector Rx distributes exactly like z. Defining k identical copies of z, $z_1, \ldots, z_k$, We get that $||\frac{1}{\sqrt{k}}Rx||$ distributes exactly like $\sqrt{\frac{1}{k}\sum_{i=1}^{k} z_i^2}$. Thus, proving Equation 7 reduces to showing that:

$$\Pr\left[\left|\sqrt{\frac{1}{k}\sum_{i=1}^{k} z_i^2} - 1\right| > \varepsilon\right] \leq \frac{1}{n^2} \tag{8}$$

for a set of independent normal random variables $z_1, \ldots, z_k \sim \mathcal{N}(0,1)$. It is sufficient to demanding that $\Pr[\sum_{i=1}^{k} z_i^2 \geq k(1+\varepsilon)^2]$ and $\Pr[\sum_{i=1}^{k} z_i^2 \leq k(1-\varepsilon)^2]$ are both smaller than $1/2n^2$. We start with bounding the probability that $\sum_{i=1}^{k} z_i^2 \geq k(1+\varepsilon)$ (this is okay because $k(1+\varepsilon) < k(1+\varepsilon)^2$).

$$\Pr[\sum z_i^2 \geq k(1+\varepsilon)] = \Pr[e^{\lambda \sum z_i^2} \leq e^{\lambda k(1+\varepsilon)}] \leq (\mathbb{E}[e^{\lambda z^2}])^k / e^{\lambda k(1+\varepsilon)}$$

Since $z \sim \mathcal{N}(0,1)$ we can compute $\mathbb{E}[e^{\lambda z^2}]$ exactly:

$$\mathbb{E}[e^{\lambda z^2}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\lambda t^2} e^{-\frac{t^2}{2}} dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(t\sqrt{1-2\lambda})^2}{2}} dt = e^{\frac{1}{2}\log(1-2\lambda)}$$

The final step is by substituting $t' = t\sqrt{1-2\lambda}$ and recalling that $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{t'^2}{2}} dt' = 1$. Finally, using the fact that $log(\frac{1}{1-2\lambda}) \leq 2\lambda + 4\lambda^2$ for $\lambda \in [0, 1/4]$ we have:

$$\mathbb{E}[e^{\lambda z^2}] = \frac{1}{\sqrt{1-2\lambda}} = e^{\frac{1}{2}\log(\frac{1}{1-2\lambda})} \leq e^{\lambda + 2\lambda^2}$$

Substituting this into the equation above we have that:

$$\Pr \leq e^{k(\lambda+2\lambda^2) - k\lambda(1+\varepsilon)} = e^{2k\lambda^2 - k\lambda\varepsilon} = e^{-k\varepsilon^2/8}$$

for $\lambda \leftarrow \varepsilon/4$. Finally, our condition that

$$\Pr[\sum_{i=1}^{k} z_i^2 \geq k(1+\varepsilon)] \leq e^{-k\varepsilon^2/8} \leq 1/2n^2$$

is achieved by $k = c\log(n)/\varepsilon^2$. Calculating for $\Pr[\sum_{i=1}^{k} z_i^2 \leq k(1-\varepsilon)]$ in the same manner shows that $k = c\log(n)/\varepsilon^2$ is also sufficient for this case. This completes the proof.

8 Fast Random Projections

We discussed in class the fact that random projection matrices cannot be made sparse in general. That is because projecting sparse vectors and preserving their norm requires the projecting matrix is almost fully dense see also [6] and [5].

But, the question is, can we actively make sure that x is not sparse? If so, can we achieve a sparse random projection for non sparse vectors? These two questions received a positive answer in the seminal work by Ailon and Chazelle [1]. The results of [1] were improved and simplified over the years. See [2] for the latest result and an overview.

In this lesson we will produce a very simple algorithm based on the ideas in [1]. This algorithm will require a target dimension of $O(\log^2(n)/\varepsilon^2)$ instead of $O(\log(n)/\varepsilon^2)$ but will be much simpler to prove.

8.1 Fast vector ℓ_4 norm reduction

The goal of this subsection is to devise a linear mapping which preserves vector's ℓ_2 norms but reduces their ℓ_4 norms with high probability. This will work to our advantage because, intuitively, vectors whose ℓ_4 norm is small cannot be too sparse. For this we will need to learn what Hadamard matrices are.

Hadamard matrices are commonly used in coding theory and are conceptually close for Fourier matrices. We assume for convenience that d is a power of 2 (otherwise we can pad out vectors with zeros). The Walsh Hadamard transform of a vector $x \in \mathbb{R}^d$ is the result of the matrix-vector multiplication Hx where H is a $d \times d$ matrix whose entries are $H(i,j) = \frac{1}{\sqrt{d}}(-1)^{\langle i,j \rangle}$. Here $\langle i,j \rangle$ means the dot product over F_2 of the bit representation of i and j as binary vectors of length $\log(d)$. Another way to view this is to define Hadamard Matrices recursively.

$$H_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad H_d = \frac{1}{\sqrt{2}} \left(\begin{array}{c:c} H_{d/2} & H_{d/2} \\ \hdashline H_{d/2} & -H_{d/2} \end{array} \right)$$

Here are a few interesting (and easy to show) facts about Hadamard matrices.

1. H_d is a unitary matrix $\|Hx\| = \|x\|$ for any vector $x \in \mathbb{R}^d$.

2. Computing $x \mapsto Hx$ requires $O(d\log(d))$ operations.

We also define a diagonal matrix D to be such that $D(i,i) \in \{1,-1\}$ uniformly. Clearly, we have that $\|HDx\|_2 = \|x\|_2$ since both H and D are isotropies. Let us now bound $\|HDx\|_\infty$. $(HDx)(1) = \sum_{i=1}^{d} H(1,i)D(i,i)x_i = \sum_{i=1}^{d} \frac{x_i}{\sqrt{d}} s_i$ where $s_i \in \{-1,1\}$ uniformly. To bound this we recap Hoeffding's inequality.

Fact 8.1 (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables s.t. $X_i \in [a_i, b_i]$. Let $X = \sum_{i=1}^{n} X_i$.*

$$\Pr[|X - \mathbb{E}[X]| \geq t] \leq 2e^{-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}} \tag{9}$$

Invoking Hoeffding's inequality and then the union bound we get that if $\|HDx\|_\infty \leq \sqrt{\frac{c\log(n)}{d}}$ for all points x. Remark, for this we assumed $\log(d) = O(\log(n))$ otherwise we should have had $\log(nd)$ in the bound. The situation, however, that the dimension is super polynomial in the number of points is unlikely. Usually it is common to have $n > d$.

Lemma 8.1. *Let $x \in \mathbb{R}^d$ by such that $\|x\| = 1$. Then:*

$$\|HDx\|_4^4 = O(log(n)/d)$$

with probability at least $1 - 1/\text{poly}(n)$

Proof. Let us define $y = HDx$ and $z_i = y_i^2$. From the above we have that $z_i \leq \frac{c\log(n)}{d} = \eta$ with probability at least $1 - 1/\text{poly}(n)$. The quantity $\|HDx\|_4^4 = \|y\|_4^4 = \sum_i z_i^2$ is a convex function of the z variables which is defined over a polytop $z_i \in [0,1]$ and $\sum_i z_i = 1$ (this is because $\|y\|_2^2 = 1$). This means that its maximal value is obtained on an extreme point of this polytope. In other words, the point $z_1, \ldots, z_{1/\eta} = \eta$ and $z_{1/\eta+1}, \ldots, z_d = 0$ or $z = [\eta, \eta, \ldots, \eta, \eta, 0, 0, 0, \ldots, 0, 0, 0]$. Computing the value of the function in this point gives $\sum_i z_i^2 \leq (1/\eta) \cdot (\eta^2) = \eta$. Recalling the $\eta = \frac{c\log(n)}{d}$ completes the proof. □

8.2 Sampling from vectors with low ℓ_4 norms

Here we prove a very simple fact. For vectors whose ℓ_4 is low, dimensionality reduction can be obtained by sampling.

Let y be a vector such that $\|y\|_2 = 1$. Let z be a sampled version of y such that $z_i = y_i/\sqrt{p}$ with probability p and 0 else. This is akin to sampling, in expectation, $d \cdot p$ coordinates from y (and scaling them up by $1/\sqrt{p}$). Note the $\mathbb{E}[\|z\|^2] = \mathbb{E}[\|y\|^2] = 1$ moreover:

$$\Pr[|\|z\|^2 - 1| > \varepsilon] = \Pr[|\sum z_i^2 - 1| > \varepsilon] = \Pr[|\sum b_i y_i^2/p - 1| > \varepsilon]$$

Where b_i are independent random indicator variables taking the $b_i = 1$ with probability p and $b_i = 0$ else. To apply Chernoff's bound we must assert that $y_i^2/p \le 1$. Let's assume this for now and return to it later. Applying Chernoff's bound we get

$$\Pr[|\sum b_i y_i^2/p - 1| > \varepsilon] \le e^{-\frac{c\varepsilon^2}{\sigma^2}}$$

where $\sigma^2 = \sum_i \mathbb{E}[(b_i y_i^2/p)^2] = \|y\|_4^4/p$. Concluding that

$$\Pr[|\|z\|^2 - 1| > \varepsilon] \le e^{-\frac{cp\varepsilon^2}{\|y\|_4^4}}$$

This shows that the concentration of the sampling procedure really depends directly on the ℓ_4 norm of the sampled vector. If we plug in the bound on $\|y\|_4^4 = \|HDx\|_4^4$ from the previous section we get

$$\Pr[|\|z\|^2 - 1| > \varepsilon] \le e^{-\frac{cp\varepsilon d}{\log(n)}} \le \frac{1}{\text{poly}(n)}$$

For some $p \in O(\log^2(n)/d\varepsilon^2)$.

8.3 Random Projection by Sampling

Putting it all together we obtain the following.

Lemma 8.2. *Define the following matrices*

- *D: A diagonal matrix such that $D_{i,i} \in \{+1, -1\}$ uniformly.*
- *H: The $d \times d$ Walsh Hadamard Transform matrix.*
- *P: A 'sampling matrix' which contains each row of matrix $I_d \cdot \sqrt{p}$ with probability $p = c\log^2(n)/d\varepsilon^2$.*

Then, with at least constant probability the following holds.

1. *The target dimension of the mapping is $k = c\log^2(n)/\varepsilon^2$ (a factor log(n) worse than optimal).*
2. *The mapping $x \mapsto PHDx$ is a $(1 \pm \varepsilon)$-distortion mapping for any set of n points. That is, for any set $x_1, \ldots, x_n \in \mathbb{R}^d$ we have*

$$\|x_i - x_j\|(1-\varepsilon) \le \|PHDx_i - PHDx_j\| \le \|x_i - x_j\|(1+\varepsilon)$$

3. *Storing PHD requires at most $O(d + k\log(d))$ space.*
4. *Applying the mapping $x \mapsto PHDx$ requires at most $d\log(d)$ floating point operations.*

References

[1] Nir Ailon and Bernard Chazelle. Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In *Proceedings of the 38st Annual Symposium on the Theory of Compututing (STOC)*, pages 557–563, Seattle, WA, 2006.

[2] Nir Ailon and Edo Liberty. An almost optimal unrestricted fast johnson-lindenstrauss transform. In *SODA*, pages 185–191, 2011.

[3] S. DasGupta and A. Gupta. An elementary proof of the Johnson-Lindenstrauss lemma. *Technical Report, UC Berkeley*, 99-006, 1999.

[4] W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

[5] Daniel M. Kane and Jelani Nelson. Sparser johnson-lindenstrauss transforms. In *SODA*, pages 1195–1206, 2012.

[6] Jelani Nelson and Huy L. Nguyen. Sparsity lower bounds for dimensionality reducing maps. In *arXiv:1211.0995v1*, 2012.