Randomized Linear Algebra and Dimension Reduction

Tengyuan Liang[1]

DLA Lecture 1: Randomize

Randomized linear algebra, dimension reduction, and data visualization. Readings: Blum, Hopcroft and Kannan [2], Chapter 2.7, 3, and 6.3.

Contents

1 *Random Projection and Dimension Reduction*

Nearest neighbor search is a frequent subroutine in dealing with high-dimensional data. Consider a data set of n points in d dimensions with large n and d. Ideally, we wish to collect the nearest data points to a given query point more efficiently, say a small function of $\log(n)$ and $\log(d)$, whereas the preprocessing time could be a polynomial of n and d. Such preprocessing, or dimension reduction, can be achieved using randomized linear projections. To this end, we will show that one can project the data points to a k-dimensional subspace with $k \asymp \log(n) \ll d$ that approximately preserves the pairwise distances.

```
import numpy as np

# Algorithm: random projection
# Input: a data matrix X of size n by d
# Output: a dimension reduction of X, f(X) which is n by k
# Random projection function f: R^d -> R^k

def random_projection(X, k):
    n, d = np.shape(X)
    U = np.random.normal(loc=0, scale=1/np.sqrt(k), size=(d,k))
    f_X = np.matmul(X, U)
    return f_X
```

[1] The University of Chicago
Booth School of Business

[2] Avrim Blum, John E. Hopcroft, and Ravindran Kannan. *Foundations of Data Science*. Cambridge University Press, 2020

Theorem 1 (Johnson-Lindenstrauss Lemma). *Let $\epsilon \in (0,1)$ and n be any integer. For any data set of n points $v_1, \ldots, v_n$ in $\mathbb{R}^d$, and integer*

$$k \gtrsim \frac{\log(n)}{\epsilon^2} , \tag{1.1}$$

the random projection $f : \mathbb{R}^d \to \mathbb{R}^k$ defined above preserves the pairwise distances in the following sense:

$$\min_{i \neq j \in [n]} \frac{\|f(v_i) - f(v_j)\|}{\|v_i - v_j\|} \geq 1 - \epsilon ,$$
$$\max_{i \neq j \in [n]} \frac{\|f(v_i) - f(v_j)\|}{\|v_i - v_j\|} \leq 1 + \epsilon ,$$

with probability at least $1 - O(\frac{1}{n})$.

Proof of Theorem 1. The proof uses standard Gaussian concentration and a union bound. By the definition of random projection, for any point $v \in \mathbb{R}^d$, $f(v) := [\langle u_1, v\rangle, \ldots, \langle u_k, v\rangle]$ with $u_j \sim \mathcal{N}(0, \frac{1}{k} I_d)$, for $j = 1, \ldots, k$. By linearity, we know

$$f(v_i) - f(v_j) = f(v_i - v_j) .$$

Apply Lemma 1, for any fixed i, j

$$\mathbb{P}\left(\left|\frac{\|f(v_i - v_j)\|}{\|v_i - v_j\|} - 1\right| \geq \epsilon\right) \leq 2e^{-c \cdot k\epsilon^2} .$$

Use union bound, we know

$$\mathbb{P}\left(\left|\frac{\|f(v_i - v_j)\|}{\|v_i - v_j\|} - 1\right| \geq \epsilon, \ \exists i \neq j \in [n]\right) \leq \binom{n}{2} 2e^{-c \cdot k\epsilon^2} . \tag{1.2}$$

By setting $k \gtrsim \frac{\log(n)}{\epsilon^2}$, one can make the above probability as small as $\frac{1}{n}$.

□

Lemma 1 (Concentration: Gaussian Annulus). *For any point $v \in \mathbb{R}^d$, $f(v) := [\langle u_1, v\rangle, \ldots, \langle u_k, v\rangle]$ with $u_j \sim \mathcal{N}(0, \frac{1}{k} I_d)$, for $j = 1, \ldots, k$, the following concentration holds with some universal constant $c > 0$*

$$\mathbb{P}\left(\big|\|f(v)\| - \|v\|\big| \geq \epsilon\|v\|\right) \leq 2e^{-c \cdot k\epsilon^2} .$$

Here the probability is taken over the randomized vectors $u_j, j = 1, \ldots, k$ that construct the random projection.

Proof. It is immediate that $\langle u_j, v\rangle \sim \mathcal{N}(0, \frac{1}{k}\|v\|^2)$, therefore $g_j := \sqrt{k}\frac{\langle u_j, v\rangle}{\|v\|} \sim \mathcal{N}(0,1)$ are i.i.d. normals

$$\frac{\|f(v)\|^2}{\|v\|^2} = \sum_{j=1}^{k} \frac{\langle u_1, v\rangle^2}{\|v\|^2} = \frac{1}{k} \sum_{j=1}^{k} g_j^2 . \tag{1.3}$$

We first show that by standard sub-Gamma/Exponential concentration inequalities

$$\mathbb{P}\left(\left|\frac{1}{k}\sum_{j=1}^{k} g_j^2 - 1\right| \geq \epsilon\right) \leq 2e^{-c\cdot k\epsilon^2} \tag{1.4}$$

TL: HW, show (1.4) in two steps: first bound each moment, then bound the Laplace transform. See Boucheron, Lugosi, and Massart.

for $\epsilon \in (0,1)$, with c at least $1/8$. Note that

$$\left|\sqrt{\frac{1}{k}\sum_{j=1}^{k} g_j^2} - 1\right| \geq \epsilon \implies \left|\frac{1}{k}\sum_{j=1}^{k} g_j^2 - 1\right| \geq \epsilon \tag{1.5}$$

we complete the proof. □

2 *Best-Fit Subspaces and Dimension Reduction*

The decompositional approach to matrix computation (1951) is listed as one of the "Top 10 Algorithms" that have influenced the practice of science and engineering in the 20th century. As dimension reduction subroutines, matrix factorization or decomposition such as singular value decomposition, principal component analysis find appearances in ranking documents and web pages (HITS algorithm, page-rank), data visualization (multi-dimensional scaling), and even in social sciences (finance, time-series analysis).

The goal is simple: given a data matrix $A \in \mathbb{R}^{n\times d}$ (of rank r), can we construct a low-rank matrix A_k of rank $k \ll \min\{n,d\}$ such that $A \approx A_k$? Can we compute A_k using a fast algorithm?

The answer is given by the singular value decomposition (SVD). Define the right singular vectors in a greedy way

$$\begin{aligned}
v_1 &:= \operatorname*{arg\,max}_{\|v\|=1} \|Av\|, && \sigma_1 := \|Av_1\| \\
v_2 &:= \operatorname*{arg\,max}_{\substack{v\perp v_1\\ \|v\|=1}} \|Av\|, && \sigma_2 := \|Av_2\| \\
&\vdots \\
v_r &:= \operatorname*{arg\,max}_{\substack{v\perp v_1,\dots,v_{r-1}\\ \|v\|=1}} \|Av\|, && \sigma_r := \|Av_r\|
\end{aligned}$$

and the left singular vectors $u_i := \frac{1}{\sigma_i}Av_i$, $i = 1,\dots r$. Algorithmically, one can solve the singular vectors using power iteration with random initialization; the convergence depends on the spectral gap.

Theorem 2 (Singular Value Decomposition)**.** *Let A be an $n \times d$ matrix of rank r with right-singular vectors $v_1, v_2 \dots, v_r$, left singular vectors $u_1, u_2, \dots u_r$, and corresponding singular values $\sigma_1, \sigma_2, \dots, \sigma_r$. Then*

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top \tag{2.1}$$

The SVD constructs the best-fit subspaces in the following sense.

Theorem 3 (Optimal Low Rank Approximation)**.** *For $k < r$. Define the rank-k approximation $A_k := \sum_{i=1}^{k} \sigma_i u_i v_i^\top$. Then*

$$\min_{\operatorname{rank}(X)\le k} \|A - X\|_{\mathrm{F}} = \|A - A_k\|_{\mathrm{F}} = \left(\sum_{i=k+1}^{r} \sigma_i^2\right)^{1/2} \tag{2.2}$$

$$\min_{\operatorname{rank}(X)\le k} \|A - X\| = \|A - A_k\| = \sigma_{k+1} \tag{2.3}$$

where $\|\cdot\|_{\mathrm{F}}, \|\cdot\|$ denote the Frobenius norm and the spectral norm. Moreover, the best-fit subspaces are given by the projection matrix $\underbrace{Q}_{n\times k} Q^\top := \sum_{i=1}^{k} u_i u_i^\top$ and $A_k = QQ^\top A$.

When n and d are large, the above computation is heavy, which requires $O(knd)$ floating-point operations. This is standard linear algebra. Can we do it fast? We can use randomized linear algebra to do this . Consider the Randomized SVD algorithm, analyzed in Halko, Martinsson, and Tropp (2010), Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. [3]

TL: The trick is to consider a slightly larger subspace, say of dimension $2k$.

```
# Algorithm: randomized SVD
# Input: a data matrix A of size n by d,
#        a rank parameter k,
#        and an exponent q = 0, 1, 2
# Output: an approximate SVD of rank k, A \approx U S V^\top
# See Halko Martinsson and Tropp, Algorithm 4.4

def randomized_SVD(A, k, q = 0):
    n, d = np.shape(A)
    m = min(2*k, d) # m ~ k << d
    # Stage A: generate d by 2k Gaussian test matrix G,
    # and reduce the problem to size n by 2k
    G = np.random.standard_normal(size=(d, m))
    Y = np.matmul(A, G) # n by m
    Q, R = np.linalg.qr(Y) # QR step to enforce orthogonality
    ## Iterative matrix exponent step to increase the spectral gap
    for i in range(q):
        Z = np.matmul(np.transpose(A), Q) # d by m
        Q, R = np.linalg.qr(Z) # QR step to enforce orthogonality
        Y = np.matmul(A, Q) # n by m
        Q, R = np.linalg.qr(Y)
    # Stage B
    B = np.matmul(np.transpose(Q), A) # m by d
```

[3] Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions, December 2010

```
    U_tilde, S, V = np.linalg.svd(B) # m by m, m by m, m by d
   U = np.matmul(Q, U_tilde) # n by m
    return U, S, V
```

The algorithm implements the following steps

1. Generate an $d \times 2k$ random Gaussian matrix G
2. Form $Y = (AA^\top)^q AG$ by multiplying alternatively with A and $A^\top$
3. Construct a matrix Q whose columns form an orthonormal basis for the range of Y
4. Form $B = Q^\top A$
5. Compute an SVD of the small matrix of size $2k \times d$: $B = \tilde{U}\Sigma_{2k}V^\top$
6. Set $U = Q\tilde{U}$, and return U, Σ_{2k}, V.

Note that the above randomized SVD is equivalent to constructing an approximate best-fit subspace $Q \in \mathbb{R}^{n\times 2k}$ using randomized linear algebra, and then approximating with

$$A \approx QQ^\top A \tag{2.4}$$

Theorem 4 (Randomized SVD)**.** *Let $A \in \mathbb{R}^{n\times d}$. Select an exponent $q \in \mathbb{N}_{\geq 0}$ and a target number of singular vectors, where $2 \leq k \leq 0.5\min\{n,d\}$. The randomized SVD algorithm to obtain a rank-2k factorization $U\Sigma_{2k}V^\top$. Then*

$$\mathbb{E}\,\|A - U\Sigma_{2k}V^\top\| \leq \Big[1 + 4\sqrt{\frac{2\min\{n,d\}}{k-1}}\Big]^{1/(2q+1)}\sigma_{k+1} + \sigma_{k+1}\,. \tag{2.5}$$

Proof of Theorem 4. Let $A = U\Sigma V^\top$ and the test matrix $G \in \mathbb{R}^{d\times \ell}$ matrix where $\ell \geq k$ (later, we take $\ell = 2k$ as specified in the algorithm). We will first derive the results for $q = 0$ case, then lift the result for general q using a power scheme. Recall that $Y = AG$ and that Q is the orthonormal basis for the range of Y, thus

$$\|A - U\Sigma_{2k}V^\top\| = \|A - QQ^\top A\| = \|(I - P_Y)A\| \tag{2.6}$$

where P_Y denotes the projection matrix to the range of Y.
Perturbation analysis on $\|(I - P_Y)A\|$. To set up the theorem and proof based on perturbation analysis, let's consider the block form (here we only consider the $n \leq d$ case, the $n > d$ case can be derived similarly)

TL: For $n > d$ case, one needs to stack an extra row of 0 below Σ_1, Σ_2 and proceed with the proof with $\Sigma_1 \in \mathbb{R}^{k\times k}, \Sigma_2 \in \mathbb{R}^{(d-k)\times(d-k)}$

$$A = U \begin{bmatrix} \Sigma_1 & \\ & \Sigma_2 \end{bmatrix} \begin{bmatrix} V_1^\top \\ V_2^\top \end{bmatrix}$$

$U \in \mathbb{R}^{n\times n}$, and $\Sigma_1 \in \mathbb{R}^{k\times k}, \Sigma_2 \in \mathbb{R}^{(n-k)\times(n-k)}$, and $V_1 \in \mathbb{R}^{d\times k}, V_2 \in \mathbb{R}^{d\times(n-k)}$. Recall here $UU^\top = U^\top U = I_n$, and $V^\top V = I_n$. Define

$$G_1 := V_1^\top G \in \mathbb{R}^{k\times \ell}\ , \tag{2.7}$$

$$G_2 := V_2^\top G \in \mathbb{R}^{(n-k)\times \ell}\ . \tag{2.8}$$

Denote

$$Y = AG = U \begin{bmatrix} \Sigma_1 G_1 \\ \Sigma_2 G_2 \end{bmatrix} \tag{2.9}$$

and we will show that

$$\|(I - P_Y)A\|^2 \leq \|\Sigma_2\|^2 + \|\Sigma_2 G_2 G_1^\dagger\|^2 \tag{2.10}$$

By unitary invariance, we first show that $U \in \mathbb{R}^{n\times n}$ does not play a role here

$$\|(I - P_Y)A\| = \|U^\top (I - P_Y) U\Sigma V^\top\| \tag{2.11}$$

$$= \|(I - P_{U^\top Y})\Sigma V^\top\| \tag{2.12}$$

note that

$$U^\top Y = U^\top AG = U^\top U\Sigma V^\top G = \begin{bmatrix} \Sigma_1 G_1 \\ \Sigma_2 G_2 \end{bmatrix}$$

Define a matrix Z whose range$(Z) \subset$ range$(U^\top Y)$

$$Z = U^\top Y \cdot G_1^\dagger \Sigma_1^{-1} = \begin{bmatrix} I_k \\ \Sigma_2 G_2 G_1^\dagger \Sigma_1^{-1} \end{bmatrix} =: \begin{bmatrix} I_k \\ F \end{bmatrix} \tag{2.13}$$

where F acts as the perturbation to the range.

TL: Ideally, we have in mind $F \approx 0$

Then

$$\|(I - P_{U^\top Y})\Sigma V^\top\| \leq \|(I - P_Z)\Sigma V^\top\| \tag{2.14}$$

and then by unitary invariance, we can show that V does not matter here since

TL: Recall $V^\top V = I_n$

$$\begin{aligned} \|(I - P_Z)\Sigma V^\top\|^2 &= \|(I - P_Z)\Sigma V^\top V \Sigma (I - P_Z)\| \\ &= \|\Sigma(I - P_Z)\Sigma\| \\ &= \|\Sigma \begin{bmatrix} I - (I + F^\top F)^{-1} & -(I + F^\top F)^{-1}F^\top \\ -F(I + F^\top F)^{-1} & I - F(I + F^\top F)^{-1}F^\top \end{bmatrix} \Sigma\| \end{aligned}$$

Note that

$$I - (I + F^\top F)^{-1} \preceq F^\top F$$
$$I - F(I + F^\top F)^{-1}F^\top \preceq I$$

and by Lemma 3, we have

$$\|(I-P_Y)A\|^2 \le \|F\Sigma_1\|^2 + \|\Sigma_2\|^2 \tag{2.15}$$
$$= \|\Sigma_2 G_2 G_1^\dagger\|^2 + \|\Sigma_2\|^2 \tag{2.16}$$

Probabilistic bounds on $\mathbb{E}\,\|\Sigma_2 G_2 G_1^\dagger\|^2$**.** Note that

$$\mathbb{E}[\|(I-P_Y)A\|] = \mathbb{E}\left(\|(I-P_Y)A\|^2\right)^{1/2} \le \mathbb{E}\left(\|\Sigma_2 G_2 G_1^\dagger\|^2 + \|\Sigma_2\|^2\right)^{1/2} \tag{2.17}$$
$$\le \mathbb{E}\,\|\Sigma_2 G_2 G_1^\dagger\| + \|\Sigma_2\| \tag{2.18}$$

Now we apply Lemma 4, we have

$$\mathbb{E}\,\|\Sigma_2 G_2 G_1^\dagger\| \le \|\Sigma_2\|_{\mathrm{F}}\,\mathbb{E}\,\|G_1^\dagger\| + \|\Sigma_2\|\,\mathbb{E}\,\|G_1^\dagger\|_{\mathrm{F}} \tag{2.19}$$
$$\le \big(\sqrt{\min\{n,d\}-k}\frac{e\sqrt{\ell}}{\ell-k} + \sqrt{\frac{k}{\ell-k-1}}\big)\sigma_{k+1} \tag{2.20}$$
$$\le \big[1+4\sqrt{\frac{2\min\{n,d\}}{k-1}}\big]\sigma_{k+1}\,. \tag{2.21}$$

where the last step plugs in $\ell = 2k$.
Power scheme for $q>0$**.** With $q>0$, we can define a matrix

$$A^{(2q+1)} := (AA^\top)^q A = U\Sigma^{2q+1}V^\top \tag{2.22}$$

The algorithm with general $q > 0$ thus becomes: defining $Y := A^{(2q+1)}G$ and then projecting to the range of Y, by Lemma 2

$$\|(I-P_Y)A\| \le \|(I-P_Y)A^{(2q+1)}\|^{1/(2q+1)} \tag{2.23}$$

The perturbation analysis can be repeated, replacing A by $A^{(2q+1)}$ and Σ by Σ^{2q+1}.

$$\begin{aligned}
\mathbb{E}[\|(I-P_Y)A\|] &\le \mathbb{E}\left(\|(I-P_Y)A^{(2q+1)}\|^2\right)^{1/2(2q+1)} \\
&\le \mathbb{E}\left(\|\Sigma_2^{2q+1}G_2G_1^\dagger\|^2 + \|\Sigma_2^{2q+1}\|^2\right)^{1/2(2q+1)} \\
&\le \mathbb{E}\,\|\Sigma_2^{2q+1}G_2G_1^\dagger\|^{1/(2q+1)} + \sigma_{k+1} \\
&\le \left(\mathbb{E}\,\|\Sigma_2^{2q+1}G_2G_1^\dagger\|\right)^{1/(2q+1)} + \sigma_{k+1} \\
&\le \left\{[1+4\sqrt{\frac{2\min\{n,d\}}{k-1}}]\sigma_{k+1}^{2q+1}\right\}^{1/(2q+1)} + \sigma_{k+1}\,.
\end{aligned}$$

□

We use the following results in the analysis.

Lemma 2 (Proposition 8.6 in Halko, Martinsson, and Tropp (2010))**.** *Let P be an orthogonal projector, and let M be a matrix. For each positive number q,*

$$\|PM\| \le \|P(MM^\top)^q M\|^{1/(2q+1)} \tag{2.24}$$

Lemma 3 (Proposition 8.3 in Halko, Martinsson, and Tropp (2010))**.** *We have $\|M\| \le \|A\| + \|C\|$ for each positive semi-definite block matrix M*

$$M = \begin{bmatrix} A & B \\ B^\top & C \end{bmatrix}$$

Lemma 4 (Proposition 10.1 in Halko, Martinsson, and Tropp (2010))**.** *Fix matrix S and T, and we draw a standard Gaussian matrix G*

$$\mathbb{E}\,\|SGT\| \le \|S\|\|T\|_{\mathrm{F}} + \|S\|_{\mathrm{F}}\|T\|\ . \tag{2.25}$$

Lemma 5 (Proposition 10.2 in Halko, Martinsson, and Tropp (2010))**.** *Draw a standard Gaussian matrix $G \in \mathbb{R}^{k\times\ell}$ with $k+1 < \ell$*

$$\mathbb{E}\,\|G^\dagger\|_{\mathrm{F}}^2 = \frac{k}{\ell-k-1} \tag{2.26}$$

$$\mathbb{E}\,\|G^\dagger\| \le \frac{e\sqrt{\ell}}{\ell-k} \tag{2.27}$$

TL: The typical behavior of minimal singular value of G is $\frac{1}{\sqrt{\ell}-\sqrt{k}} \le \frac{2\sqrt{\ell}}{\ell-k}$. The above theorem is just a stronger version of this intuition. The first is due to standard facts in multivariate analysis, or one can also derive from the spectral density of Gaussian matrices.

3 *Matrix Sampling and Dimension Reduction*

References

Avrim Blum, John E. Hopcroft, and Ravindran Kannan. *Foundations of Data Science*. Cambridge University Press, 2020.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions, December 2010.