CS 6220: DATA-SPARSE MATRIX COMPUTATIONS
Lecture 7: Low-dimensional embeddings

Scribes: June Cho and Ke Alexander Wang ({sc782,kaw293}@cornell.edu)

February 13, 2020

1 Introducing to embeddings

Modern scientific computing often requires us to efficently processes high dimensional data. For example, a 256 by 256 RBG image is $256 \times 256 \times 3 = 196,608$ dimensional and a 3 minute song sampled at 44.1kHz is $44,100 \times 3 \times 60 = 7,938,000$ dimensional. In these cases, a naive application of classical algorithms to high dimensional data may requre intractable memory or computation time. However, if we can reduce the data dimensionality, then we can continue to use existing algorithms. Note that any useful dimensionality reduction method must preserve geometric relationships in the data since they contain meaningful information about the inputs. For instance, the relative positions and directions of word embeddings contain semantic and syntactic relationships among words, and the low-dimensional translations must, to some extent, preserve such properties for further analysis and processing.

Here we will introduce low-dimensional embeddings as a method of dimensionality reduction. We will see that such embeddings are cheap to compute using randomness while preserving geometric attributes of the input data. Once computed, such low-dimensional embeddings will only require multiplying the input by the embedding matrix, an operation that can be accelerated with either highly-optimized dense matrix multiplication routines or with sparse embedding matrices.

Definition 1. Let $\mathcal{V} \subseteq \mathbb{R}^n$. $S \in \mathbb{R}^{d \times n}$ is an *embedding* with distortion $\epsilon > 0$ of $\mathcal{V}$ if for all $x \in \mathcal{V}$

$$(1-\epsilon)\|x\|_2 \leq \|Sx\|_2 \leq (1+\epsilon)\|x\|_2 . \tag{1}$$

Note that typically embeddings will have $d < n$. Notice that an embedding must have $Sx \neq 0$ for all $x \neq 0$. Conceptually, this corresponds to a "loss of information" about the original x. Mathematically, it would fail to preserve norms since $(1-\epsilon)\|x\|_2 > 0$.

Definition 2. A $S \in \mathbb{R}^{d \times n}$ that satisfies 1 is a *subspace embedding* of $\mathcal{V}$ if $\mathcal{V}$ is a linear subspace with dimension k of $\mathbb{R}^n$.

In other words, an ϵ-embedding is one that approximately preserves distances and inner products within the set $\mathcal{V}$. It turns out that we can construct embeddings quite easily by letting them be random matrices. Even better, sparse low dimensional embeddings can also be made from random matrices. The existence of low dimensional, sparse embeddings has implications for algorithms that we've seen in this class. For example, when doing randomized least squares, we can use a sparse S instead of a random Gaussian dense matrix to accelerate the algorithm and reduce memory usage.

Definition 3. Given some subset of the unit sphere, $E \subseteq \mathbb{S}^{n-1}$, the minimum and maximum *restricted singular values* of $S \in \mathbb{R}^{d \times n}$ with respect to E are

$$\sigma_{\min}(S, E) := \min_{x \in E} \|Sx\|_2 \tag{2}$$

$$\sigma_{\max}(S, E) := \max_{x \in E} \|Sx\|_2 \tag{3}$$

Restricted singular values are generalizations of ordinary singular values, which we obtain if $E = \mathbb{S}^{n-1}$. These quantities measure how much the embedding S can distort the set E. If we can control the minimum and maximum restricted singular values of S then we can control the approximation quality of the embedding S. The following analysis will be based on choosing S to be a Gaussian embedding.

Definition 4. A *Gaussian embedding* $S = \Gamma \in \mathbb{R}^{d \times n}$ has entries $\Gamma_{ij} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, d^{-1})$. Note that the variance of the distribution is to normalize the norm of x such that

$$\mathbb{E} \|\Gamma x\|_2^2 = \|x\|_2^2, \quad \forall x \in \mathbb{R}^n.$$

Definition 5. Given $E \subseteq \mathbb{R}^n$, the *Gaussian width* of E is defined to be

$$\omega(E) := \mathbb{E}_g \left[\sup_{x \in E} \langle g, x \rangle \right]$$

where $g \in \mathbb{R}^n$ has i.i.d. $\mathcal{N}(0, 1)$ entries.

The Gaussian width can be thought of as another fundamental geometric property of any $E \subseteq \mathbb{R}$ alongside volume and surface area. For example, the Gaussian width of a unit $n-1$-sphere $\mathbb{S}^{n-1}$ is asymptotically $\sqrt{n}$. A great reference for the Gaussian width and other probabilistic methods can be found in Section 7.5 of Vershynin (2018) which is also freely available online.

The Gaussian width has some nice properties:

- $\omega(QE) = \omega(E)$ for any orthogonal Q where $QE := \{Qx \mid x \in E\}$
- $E \subseteq F$ implies $\omega(E) \leq \omega(F)$
- $0 \leq \omega(E) \leq \sqrt{n}$

Proposition 1. *If $\mathcal{V}$ is a k dimensional subspace and $E = \mathcal{V} \cap \mathbb{S}^{n-1}$, we have*

$$\sqrt{k-1} \leq \omega(E) \leq \sqrt{k}.$$

Thus the squared Gaussian width in some sense captures the "dimensionality" of E even though E may not be a subspace.

The following theorem from Gordon (1988) shows that the Gaussian width can be used to control minimum and maximum restricted singular values with respect to a subset on the unit sphere.

Theorem 2. *Fix $E \subseteq \mathbb{S}^{n-1}(\mathbb{R})$ and let $\Gamma \in \mathbb{R}^{d \times n}$ have i.i.d. $\mathcal{N}(0, d^{-1})$ entries. Then*

$$\mathbb{P}\{\sigma_{\min}(\Gamma, E) \leq 1 - \frac{\omega(E) + 1}{\sqrt{d}} - t\} \leq e^{-dt^2/2} \tag{4}$$

$$\mathbb{P}\{\sigma_{\max}(\Gamma, E) \geq 1 + \frac{\omega(E)}{\sqrt{d}} + t\} \leq e^{-dt^2/2}. \tag{5}$$

Remark 1. Notice that neither the bound nor the probability depends on the input dimension n. These theorems should be thought of as giving high probability upper bounds on $\sigma_{\max}$ and lower bounds on $\sigma_{\min}$ which you may recall correspond to how much Γ perturbs E.

Remark 2. The theorem also leads to the following relation.

$$1 - \frac{\omega(E)+1}{\sqrt{d}} \lessapprox \sigma_{\min}(\Gamma, E) \leq \sigma_{\max}(\Gamma, E) \lessapprox 1 + \frac{\omega(E)}{\sqrt{d}} \tag{6}$$

This implies that setting $d > (\omega(E)+1)^2$ ensures that the embedding Γ annihilates a point in E with low probability.

Lastly, a generalization of the Bai-Yin Law in Section 5.2 of Bai and Silverstein (2010) implies that the bounds in Equation 6 are nearly optimal.

Theorem 3. *Let $E = \mathbb{S}^{n-1}$ and let $\Gamma \in \mathbb{R}^{d\times n}$ have i.i.d. $\mathcal{N}(0, d^{-1})$ entries. Then*

$$1 - \sqrt{\frac{n}{d}} \lessapprox \mathbb{E}\sigma_{\min}(\Gamma, E) \leq \mathbb{E}\sigma_{\max}(\Gamma, E) \leq 1 + \sqrt{\frac{n}{d}} \tag{7}$$

as $n, d \to \infty$ with n/d fixed in $[0, 1]$.

Note that this wouldn't apply if $d < n$ which is the typical setting for low dimensional embeddings.

2 Johnson-Lindenstrauss as a random matrix embedding

The Johnson-Lindenstrauss theorem shows that random projections of a set of points to a lower dimension will approximately preserve pairwise distances. The classic dimension reduction problem introduced in Johnson and Lindenstrauss (1984) is a special case of low dimensional embeddings with a specific E which turns out to be a Gaussian embedding.

Thus the setting is: given $\epsilon > 0$ and a discrete set of N points, $\{a_1, \ldots, a_N\} \subset \mathbb{R}^n$, does a Gaussian embedding $\Gamma \in \mathbb{R}^{d\times n}$ preserve all pairwise distances? In other words, is

$$(1-\epsilon) \leq \frac{\|\Gamma(a_i - a_j)\|_2}{\|a_i - a_j\|_2} \leq (1+\epsilon) \qquad \forall i \neq j \quad ? \tag{8}$$

Consider $E = \{\frac{a_i - a_j}{\|a_i - a_j\|_2} \mid i, j = 1, \ldots, N, i \neq j\}$ so that $\omega(E) \leq \sqrt{2\log|E|} < 2\sqrt{\log N/2}$. Plugging this upper bound on $\omega(E)$ into Theorem 2, it can be shown that

$$\mathbb{P}\{\sigma_{\min}(\Gamma, E) \leq 1 - (1 + 2\sqrt{\log N/2})/\sqrt{d} - t\} \leq e^{-dt^2/2} \tag{9}$$

$$\mathbb{P}\{\sigma_{\max}(\Gamma, E) \geq 1 + 2\sqrt{\log N/2}/\sqrt{d} + t\} \leq e^{-dt^2/2} \tag{10}$$

Remark 3. Given a fixed accuracy ϵ, choosing $d \geq 8\epsilon^{-2}\log N$ achieves Equation 8 with high probability. While the sufficient embedding dimension is logarithmic with respect to the number of points N, it has a quadratic dependence on ϵ^{-1}. This leads to requiring a relatively large number of embedding dimensions even for ϵ close to 1, and thus more practical approaches have focused more towards bounding the minimum restricted singular value to avoid coalescing any of the points in E.

3 How general is this phenomenon for random matrices?

It can be shown that this low dimensional embedding phenomenon for random matrices is "universal": a large family of randomized dimension reduction embeddings including Rademacher matrices and random sparse matrices share the same high probability bound on the minimum restricted singular value. All that is required is that certain moments of their entries are bounded in expectation. Below is a simplified version of Theorem 9.1 from Oymak and Tropp (2017).

Theorem 4. *Fix $E \subseteq \mathbb{S}^{n-1}(\mathbb{R})$ and let $S \in \mathbb{R}^{d\times n}$ be random with independent entries satisfying:*

1. $\mathbb{E}[S_{ij}] = 0$
2. $\mathbb{E}[S_{ij}^2] = d^{-1}$
3. $\mathbb{E}[S_{ij}^5] \leq R$.

When $d \leq n$, we have with high probability,

$$\sigma_{\min}(S, E) \geq 1 - \frac{\omega(E)}{\sqrt{d}} - o\left(\sqrt{\frac{n}{d}}\right) \tag{11}$$

where the constant in the little-o depends only on R.

Note that the entries of S don't need to be from the same distribution. The fact that R is uniform over i, j is key to the proof of this theorem.

4 Sparse random matrix embeddings

It would be great if we can have a embedding S that allows us to compute Sx quickly. If $S \in \mathbb{R}^{d\times n}$ is dense, then multiplying a vector by S is $O(dn)$. One way to reduce the time complexity would be to have an embedding S that is sparse which would take only $O(\mathrm{nnz}(S))$ time to compute a matrix-vector multiplication Sx. Indeed, the following theorem from Cohen (2016) shows that sparse subspace embeddings do exist with the target dimension having a sub-quadratic dependence on the subspace dimension.

Theorem 5. *Let $S \in \mathbb{R}^{d\times n}$. For $2 \leq \gamma \leq d$ where γ is an integer, let*

$$S = \sqrt{\frac{n}{\gamma}}[s_1, \ldots, s_n]$$

where $s_i \in \mathbb{R}^d$ are i.i.d. vectors uniformly sampled from vectors with γ ones in $\mathbb{R}^d$. For a constant ϵ, S is an accurate embedding for subspace $\mathcal{V}$ with dimension k as long as $d = O(k \log k)$ and $\gamma = O(\log k)$ where ϵ is hidden in the asymptotic notation.

References

Z. Bai and J. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. 01 2010. doi: 10.1007/978-1-4419-0661-8.

M. B. Cohen. Nearly tight oblivious subspace embeddings by trace inequalities. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '16, page 278–287, USA, 2016. Society for Industrial and Applied Mathematics. ISBN 9781611974331.

Y. Gordon. On milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$. In J. Lindenstrauss and V. D. Milman, editors, *Geometric Aspects of Functional Analysis*, pages 84–106, Berlin, Heidelberg, 1988. Springer Berlin Heidelberg. ISBN 978-3-540-39235-4.

W. Johnson and J. Lindenstrauss. Extensions of lipschitz maps into a hilbert space. *Contemporary Mathematics*, 26:189–206, 01 1984. doi: 10.1090/conm/026/737400.

P.-G. Martinsson and J. Tropp. Randomized numerical linear algebra: Foundations & algorithms, 2020.

S. Oymak and J. A. Tropp. Universality laws for randomized dimension reduction, with applications. *Information and Inference: A Journal of the IMA*, 7(3):337–446, 11 2017. ISSN 2049-8764. doi: 10.1093/imaiai/iax011. URL https://doi.org/10.1093/imaiai/iax011.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.