CS369G: Algorithmic Techniques for Big Data Spring 2015-2016

Lecture 19: Sparse Subspace Embeddings

Prof. Moses Charikar *Scribes: Colin Wei*

1 Overview

In this lecture, we discuss algorithms to produce a subspace embedding for the column space of a matrix A. The algorithm given by Clarkson and Woodruff [3] uses the count sketch matrix to produce a subspace embedding that runs in time $O(\mathrm{nnZ}(A) + \mathrm{poly}(d/\epsilon))$. We present a proof that the algorithm works with high probability.

2 Preliminaries

Definition 1. *Let A be a n by d matrix. A $(1 \pm \epsilon) - l_2$ subspace embedding for the column space of A is S such that $\forall x \in \mathbb{R}^d$*

$$(1-\epsilon)\|Ax\|_2^2 \leq \|SAx\|_2^2 \leq (1+\epsilon)\|Ax\|_2^2$$

We can let U be a matrix with orthonormal columns such that $\mathrm{colspace}(U) = \mathrm{colspace}(A)$. Then the requirement for $(1 \pm \epsilon) - l_2$ subspace embedding becomes:

$$\|SUy\|_2^2 \in [(1-\epsilon)\|Uy\|_2^2, (1+\epsilon)\|Uy\|_2^2] = [(1-\epsilon)\|y\|_2^2, (1+\epsilon)\|y\|_2^2]$$

Equivalently, we could also require $\|I_d - U^T S^T SU\|_2 \leq \epsilon$.

Definition 2. *Let π be a distribution on r by n matrices S, where $r = f(n, d, \epsilon, \delta)$ for some function f. Suppose that with probability $\geq 1 - \delta$ and any fixed n by d matrix A, $S \sim \pi$ is a $(1 \pm \epsilon) - l_2$ subspace embedding for A. Then π is called an (ϵ, δ)-oblivious subspace embedding.*

Examples of oblivious subspace embeddings include when the entries of S are i.i.d. Gaussian, S is a FJLT matrix, or when S is a P.H.D. matrix.

3 Sparse Embedding Matrix

In the setting where the matrix A is sparse, [3] provide an embedding which can be computed in time $O(\mathrm{nnZ}(A))$, the number of nonzero elements in the matrix A. The embedding can be computed by the count-sketch or sparse-embedding matrix, which is a r by n matrix constructed as follows: let $h : [n] \to [r]$ and $\sigma : [n] \to \{-1, 1\}$ be hash functions. Then the i-th column of the sparse embedding matrix S is nonzero only in the $h(i)$-th row. This nonzero entry has value $\sigma(i)$. We can see from this construction that the product SA can be computed in $O(\mathrm{nnZ}(A))$ time because each non-zero entry in A is multiplied by at most one nonzero entry in S. The following theorem holds:

Theorem 3. *Let S be the sparse embedding matrix of dimension r by n, where $r = O\left(\frac{d}{\epsilon^2} polylog\left(\frac{d}{\epsilon}\right)\right)$. Then for any fixed A, S is a $(1 \pm \epsilon) - l_2$ subspace embedding for A with constant probability.*

We discuss the following slightly different result:

Theorem 4. *Let S be the sparse embedding matrix with $r = O\left(\frac{d^2}{\epsilon'^2 \delta}\right)$ rows. Then with probability $1 - \delta$ for any fixed A, S is a $(1 \pm \epsilon') - l_2$ subspace embedding for the columns of A.*

For this theorem to hold, h needs to be a 2-wise independent hash function, and σ needs to be a 4-wise independent hash function.

Proof Sketch due to [3]. The proof in [3] proceeds by bounding

$$P(\|I_d - U^T S^T S U\|_2 > \epsilon) = P(\|I_d - U^T S^T S U\|_2^l > \epsilon^l)$$

using trace inequalities. □

We present a different, simpler proof by [2], which leverages the machinery of approximate matrix multiplication.

Definition 5. *We say that C is an ϵ-approximate matrix product of A, B if it satisfies*

$$\|A^T B - C\|_F \leq \epsilon \|A\|_F \|B\|_F$$

The idea to compute a approximate matrix product is to maintain sketches SA, SB of the original matrices, where we want $E[A^T S^T S B] = A^T B$. S is an r by n matrix, and we want to bound the size of r needed to get a good approximation with high probability.

Definition 6. *[4] A distribution $\mathcal{D}$ on $S \in \mathbb{R}^{kxd}$ is said to satisfy the (ϵ, δ, l)-JL moment property if $\forall x \in \mathcal{R}^d$ where $\|x\|_2 = 1$, $E[(\|Sx\|_2^2 - 1)^l] \leq \epsilon^l \delta$.*

Definition 7. *For a scalar random variable X, let $\|X\|_p = E[|X|^p]^{1/p}$. $\|\cdot\|_p$ is a metric, so $\|X + Y\|_p \leq \|X\|_p + \|Y\|_p$.*

Lemma 8. *Let $l \geq 2$, $\epsilon, \delta \in (0, 1/2)$, and $\mathcal{D}$ be a distribution that satisfies the (ϵ, δ, l)-JL moment property. Then for A, B with d rows,*

$$P_{S \sim D}\left[\|A^T S^T S B - A^T B\|_F > 3\epsilon \|A\|_F \|B\|_F\right] \leq \delta$$

Proof. We first note that for $x, y \in \mathbb{R}^d$, $\langle Sx, Sy \rangle = \frac{1}{2}\left(\|Sx\|_2^2 + \|Sy\|_2^2 - \|S(x-y)\|_2^2\right)$. Thus,

$$\begin{aligned}
\|\langle Sx, Sy \rangle - \langle x, y \rangle\|_l &= \frac{1}{2}\|(\|Sx\|_2^2 - 1) + (\|Sy\|_2^2 - 1) - (\|S(x-y)\|_2^2 - \|x-y\|_2^2)\|_l \\
&\leq \frac{1}{2}\left(\|\|Sx\|_2^2 - 1\|_l + \|\|Sy\|_2^2 - 1\|_l + \|\|S(x-y)\|_2^2 - \|x-y\|_2^2\|_l\right) \\
&\leq \frac{1}{2}(\epsilon\delta^{1/l} + \epsilon\delta^{1/l} + \|x-y\|_2^2 \epsilon\delta^{1/l} \\
&\leq 3\epsilon\delta^{1/l}
\end{aligned}$$

where we first apply triangle inequality and then apply the JL moment property. From this, we can conclude that for arbitrary x, y,

$$\|\langle Sx, Sy\rangle - \langle x, y\rangle\|_l \le 3\epsilon\delta^{1/l}\|x\|_2\|y\|_2$$

Now since the ij-th entry of $A^T B$ is given by $\langle A^i, B^j\rangle$, the inner product of the i-th column of A and the j-th column of B, we have that

$$\begin{aligned}\|\|A^T S^T SB - A^T B\|_F^2\|_{l/2} &\le \sum_{ij} \|(\langle SA^i, SB^j\rangle - \langle A^i, B^j\rangle)^2\|_{l/2} \\ &\le (3\epsilon\delta^{1/l})^2 \sum_{ij} \|A^i\|_2^2\|B^j\|_2^2 \\ &= (3\epsilon\delta^{1/l})^2\|A\|_F^2\|B\|_F^2\end{aligned}$$

where the first line follows from triangle inequality, and the second from plugging in the inequality derived previously. Now we plug this into Markov's inequality to get that

$$\begin{aligned}P[\|A^T S^T SB - A^T B\|_F^l > (3\epsilon)^l\|A\|_F^l\|B\|_F^l] &\le \frac{1}{(3\epsilon\|A\|_F\|B\|_F)^l} E[\|A^T S^T SB - A^T B\|_F^l] \\ &\le \delta\end{aligned}$$

□

Now we are ready to prove Theorem 4.

Proof of Theorem 4. We want to show that if S is the sparse embedding matrix with at least $\frac{2}{\epsilon^2\delta}$ rows, S satisfies the $(\epsilon, \delta, 2)$-JL moment property. We need to show that for a unit vector x with $\|x\|_2 = 1$, $E[(\|Sx\|_2^2 - 1)^2] \le \epsilon^2\delta$. We do this by expanding to get $E[\|Sx\|_2^4] - 2E[\|Sx\|_2^2] + 1$; the middle term is 1 and from expansion we can show that $E[\|Sx\|_2^4] \le 1 + \frac{2}{r}$, so $E[(\|Sx\|_2^2 - 1)^2] \le \frac{2}{r}$. Thus, if $r > \frac{2}{\epsilon^2\delta}$, the $(\epsilon, \delta, 2)$-JL moment property holds.

Let U be an orthonormal basis for the columns of A. Now since S satisfies the $(\epsilon, \delta, 2)$-JL moment property,

$$\begin{aligned}P[\|U^T S^T SU - U^T U\|_F > 3\epsilon\|U\|_F^l\|U\|_F] &\le \delta \\ \implies P[\|U^T S^T SU - I_d\|_F > 3\epsilon d] &\le \delta\end{aligned}$$

So with $\epsilon = \frac{\epsilon'}{d}$, we get $r = O\left(\frac{d^2}{\epsilon'^2\delta}\right)$ rows needed. □

References

[1] Sketching as a Tool for Numerical Linear Algebra, David P. Woodruff, Foundations and Trends in Theoretical Computer Science, 2014.

[2] OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings, J. Nelson and H.L. Nguyen, FOCS 2013.

[3] Low Rank Approximation and Regression in Input Sparsity Time, K.L. Clarkson and D.P. Woodruff, STOC 2013.

[4] A Sparser Johnson-Lindenstrauss Transform, Daniel M. Kane and Jelani Nelson, 2010.