CS 15-859: Algorithms for Big Data Fall 2020

Lecture 5.1 — October 8

Prof. David Woodruff *Scribe: Chris Seiler*

1 Distributed Low Rank Approximation: Introduction

So far, we have seen sketching-based algorithms for computing rank-k approximations to an input matrix $A \in \mathbb{R}^{n\times d}$. However, there are some instances for which A is too large to store on a single server; instead, A is distributed among s servers. In this case, we would like to solve the same low rank approximation problem, but while modelling the cost of both computation *and* communication.

Definition. In the *arbitrary partition model*, the matrix $A \in \mathbb{R}^{n\times d}$ is distributed among s servers as

$$A = A^1 + A^2 + \cdots + A^s$$

where A^t denotes the t^{th} server's matrix.

Example 1. There are s shops, each serving a set of n customers and d products. The t^{th} shop tracks a matrix A^t, where A^t_{ij} contains the number of times customer i bought product j at this particular shop. Then the customer product matrix A, in the arbitrary partition model, has entries which indicate the total number of times each customer purchased each product.

The arbitrary partition model is not the only setting under which we consider the problem at hand; the following (less general) setting is often a useful model for distributed matrices.

Definition. The *row partition model* contains s servers, each containing a subset of rows in A. Typically, A is given by the block matrix

$$A = \begin{bmatrix} A^1 \\ \hline A^2 \\ \hline \vdots \\ \hline A^s \end{bmatrix}$$

1.1 The Communication Model

When solving distributed low rank approximation, we need a way to account for communication costs. Typically, we assume the following model.

Definition. The *coordinator model* assumes that each of the s servers is allowed 2-way communication with a coordinator node, and no more.

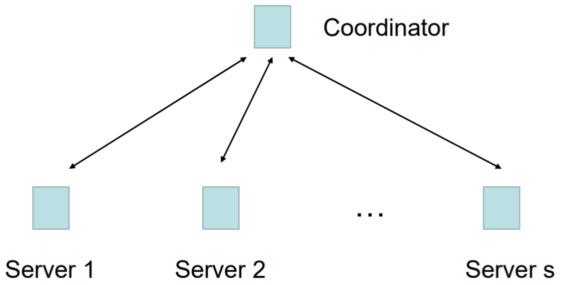

Figure 1: An illustration of communication allowed in the coordinator model.

Remark 1. A similarly-defined peer-to-peer model allows for communication between any two arbitrary nodes. However, there are two reasons we only consider the coordinator model. First, it is difficult to analyze communications in a peer-to-peer setting; there are $\binom{s}{2}$ links, and communication is not necessarily synchronized. Second, we do not lose much power in the coordinator model; we can simulate the peer-to-peer model up to a factor of 2 messages, as well as an additive $O(\log s)$ factor per message. Each server simply needs to send a message to the coordinator, specifying the target server in $O(\log s)$ bits. Then the coordinator can forward the message appropriately.

1.2 Communication Cost of Low Rank Approximation

- Input: A matrix $A \in \mathbb{R}^{n \times d}$ stored on s servers in the arbitrary partition model. Server t has a matrix $A^t \in \mathbb{R}^{n \times d}$, and the entries of each A^t are $O(\log(nd))$-bit integers.
- Output: Each server must compute the same k-dimensional space W. If P_W denotes the projection matrix onto W, then the output is

  $$C = A^1 P_W + A^2 P_W + \cdots + A^s P_W = AP_W$$

  which is a projection of the rows of A onto W. We want C to be a good estimate for the best rank-k approximate matrix, which we denote A_k; more precisely, we want

  $$\|A - C\|_F \leq (1 + \epsilon)\|A - A_k\|_F$$

  for some parameter ϵ.
- Resources: We want to minimize the total communication, in bits, as well as the total computational cost incurred. Additionally, we often want to devise a protocol with $O(1)$ "round complexity", which is the number of back-and-forth messages required to compute the answer.

1.3 Work on Distributed Low Rank Approximation

There are several existing approaches to the problem we will consider.

- The FSS protocol [3] applies to the row partition model, requiring $O(sdk/\epsilon)$ *real numbers* of communication. Although the communication does not depend on n, note that the use of real numbers in communication is generally avoided, as arbitrary real numbers can encode an unbounded amount of information in their representation. Assuming a fixed precision, the bit complexity can be large; furthermore, FSS requires SVD running time on the input A.

- The KVW protocol [4] requires $O(skd/\epsilon)$ communication in the arbitrary partition model, and has a faster running time with respect to n. However, KVW is still not optimal in terms of communication; we would like to remove the ϵ from the denominator in the high-order term skd.
- The BWZ protocol [2] uses $O(skd)+\text{poly}(sk/\epsilon)$ words of communication, again in the arbitrary partition model, improving on the higher-order term from KVW. Furthermore, the computation can be done in input sparsity time.

Remark 2. The BWZ protocol has a matching lower bound on the higher-order term of communication cost: $\Omega(skd)$ words of communication are required in the general case. Intuitively, if the subspace W is represented by a $k \times d$ matrix with orthonormal rows (which we will also call W), then it is possible that most servers have little knowledge of W. We require that each server learns W, so loosely speaking, the kd entries of W have to be sent to each of the s servers. This idea can be formalized to show the desired bound.

Remark 3. Some variants on low rank approximation are considered in literature. The BLSWX protocol [1] tackles kernel low-rank approximation, which seeks a rank-k approximation to A after applying a "kernel mapping" $f : \mathbb{R}^d \to \mathbb{R}^{d'}$ to each row of A. Other variants are for implicit matrices, using the WZ protocol [5], and sparsity, using the BWZ protocol [2].

2 Constructing a Coreset [FSS]

We now examine the first protocol, FSS, which will require a definition.

Definition. Let $A \in \mathbb{R}^{n\times d}$ and let $A = U\Sigma V^T$ be its SVD. For some m, let Σ_m agree with Σ on the first m diagonal entries (i.e. the highest m singular values), and be 0 otherwise. We define a *coreset* to be the matrix

$$\Sigma_m V^T$$

Claim 1. For a matrix A and a coreset $\Sigma_m V^T$, where $m = k + k/\epsilon$, and for all projection matrices $Y = I - X$ onto a $(d-k)$-dimensional subspace,

$$\|AY\|_F^2 \leq \|\Sigma_m V^T Y\|_F^2 + c \leq (1+\epsilon)\|AY\|_F^2$$

where $c = \|A - A_m\|_F^2$ does not depend on n.

Before proving this claim, we make a few observations. Here, X is a projection onto a k-dimensional space, and $Y = I - X$ is a projection onto the complement space of X. The coreset $\Sigma_m V^T = \Sigma_m V_m^T$, where V_m^T has all but the first m rows of V^T zeroed out, since Σ_m only has m diagonal entries.

A consequence of claim 1 is that it suffices to know $\Sigma_m V_m^T$, which only has $md \ll nd$ parameters, to accurately answer queries about $\|AY\|_F^2$ up to a $1+\epsilon$ factor. Additionally, the provided guarantee is similar to what we have seen in sketching matrices; indeed, we can think of the matrix S as U_m^T. Then, $SA = U_m^T U\Sigma V^T = \Sigma_m V^T$ is the sketch, which is exactly the coreset.

One final observation we will make is that, if $\tilde{Y}$ is the appropriate projection minimizing $\|\Sigma_m V^T Y\|_F^2$ and Y^* is similarly a projection which minimizes $\|AY\|_F^2$, then

$$\begin{aligned}\|A\tilde{Y}\|_F^2 &\leq \|\Sigma_m V^T \tilde{Y}\|_F^2 + c && (1)\\ &\leq \|\Sigma_m V^T Y^*\|_F^2 + c && (2)\\ &\leq (1+\epsilon)\|AY^*\|_F^2 && (3)\\ &= (1+\epsilon)\|A - A_k\|_F^2 && (4)\end{aligned}$$

(1) and (3) follow from claim 1; (2) follows from the definition of $\tilde{Y}$ being the minimizer. (4) follows since $Y^* = I - X^*$ is the projection onto the complement space of some X^*; since X^* is a k-dimensional projection minimizing $\|A - AX\|_F^2 = \|AY\|_F^2$, we have $AX^* = A_k$.

The following lemma will be helpful for proving claim 1.

Lemma 1. *Any projection matrix P will not increase lengths; that is, for any matrix A, $\|AP\|_F^2 \leq \|A\|_F^2$.*

Proof. Recall that a projection matrix P must have all singular values be either 0 or 1. Then, for any A and B, note that

$$\begin{aligned}\|AB\|_F^2 &= \sum_{\text{rows } i} \|A_{i,*}B\|_2^2\\ &\leq \sum_{\text{rows } i} \|A_{i,*}\|_2^2 \|B\|_2^2 && \text{def. of operator norm}\\ &= \|A\|_F^2 \|B\|_2^2\end{aligned}$$

and, for $B = P$, the operator norm is $\|P\|_2^2 = \sigma_{\max}^2 \leq 1$, so the lemma follows. ■

Finally, we prove claim 1.

Proof of Claim 1. First, we want to show the direction $\|AY\|_F^2 \leq \|\Sigma_m V^T Y\|_F^2 + c$, as follows:

$$\|AY\|_F^2 = \|U\Sigma_m V^T Y + U(\Sigma - \Sigma_m)V^T Y\|_F^2$$

Note that the first m columns of U are selected in the first term, and all but these columns are selected in the second term. Since U has orthonormal columns, it follows that the columns in both terms are orthogonal; by the Pythagorean theorem,

$$= \|U\Sigma_m V^T Y\|_F^2 + \|U(\Sigma - \Sigma_m)V^T Y\|_F^2$$

Since U has orthonormal columns, it does not change the Frobenius norm. In the second term, Y is a projection matrix, so by Lemma 1 we can continue as follows:

$$\begin{aligned}&\leq \|\Sigma_m V^T Y\|_F^2 + \|U(\Sigma - \Sigma_m)V^T\|_F^2\\ &= \|\Sigma_m V^T Y\|_F^2 + \|A - A_m\|_F^2\\ &= \|\Sigma_m V^T Y\|_F^2 + c\end{aligned}$$

as desired.

We now prove the second direction: $\|\Sigma_m V^T Y\|_F^2 + c \le (1+\epsilon)\|AY\|_F^2$. Subtracting $\|AY\|_F^2$ on both sides and using the definition of c, it suffices to show

$$\|\Sigma_m V^T Y\|_F^2 + \|A - A_m\|_F^2 - \|AY\|_F^2 \le \epsilon \|AY\|_F^2$$

If $Y = I - X$, then $\Sigma_m V^T Y + \Sigma_m V^T X = \Sigma_m V^T$. Furthermore, X and Y project onto complementary spaces; this allows us to show that X and Y have orthogonal rows, so we can apply the Pythagorean theorem:

$$\|\Sigma_m V^T Y\|_F^2 + \|A - A_m\|_F^2 - \|AY\|_F^2 = \|\Sigma_m V^T\|_F^2 - \|\Sigma_m V^T X\|_F^2 + \|A - A_m\|_F^2 - \|A\|_F^2 + \|AX\|_F^2$$

Since U has orthonormal columns,

$$= \|U\Sigma_m V^T\|_F^2 - \|\Sigma_m V^T X\|_F^2 + \|A - A_m\|_F^2 - \|A\|_F^2 + \|AX\|_F^2$$

By definition of A_m,

$$= \|A_m\|_F^2 - \|\Sigma_m V^T X\|_F^2 + \|A - A_m\|_F^2 - \|A\|_F^2 + \|AX\|_F^2$$

We apply the Pythagorean theorem again, this time noticing that $A_m + (A - A_m) = A$ and that A_m and $(A - A_m)$ have orthogonal columns:

$$\begin{aligned} &= \|AX\|_F^2 - \|\Sigma_m V^T X\|_F^2 \\ &= \|U\Sigma V^T X\|_F^2 - \|U\Sigma_m V^T X\|_F^2 \\ &= \|U(\Sigma - \Sigma_m)V^T X\|_F^2 \end{aligned}$$

where we once again use orthogonality and that U has orthonormal columns. Using $\|CD\|_F^2 \le \|C\|_F^2 \|D\|_2^2$ from within Lemma 1,

$$\le \|U(\Sigma - \Sigma_m)V^T\|_2^2 \|X\|_F^2$$

The first term is in SVD form, so its maximum singular value is σ_{m+1} and is equal to its operator norm. In the second term, X is a rank-k projection matrix, which has k singular values of exactly 1, so

$$\begin{aligned} &= \sigma_{m+1}^2 \sum_{i=1}^{k} 1^2 = \sigma_{m+1}^2 k \\ &= \epsilon \sigma_{m+1}^2 (m - k) \\ &\le \epsilon \sum_{i=k+1}^{m} \sigma_i^2 \\ &\le \epsilon \sum_{i=k+1}^{d} \sigma_i^2 = \epsilon \|A - A_k\|_F^2 \\ &\le \epsilon \|AY\|_F^2 \end{aligned}$$

as desired, where the last line follows from that $\|A - A_k\|_F^2 = \|AY^*\|_F^2$, where Y^* is a $(d-k)$-dimensional projection which minimizes $\|AY^*\|_F^2$. Other lines above follow from the definitions of m, decreasing singular values down the diagonal, and properties of the Frobenius norm. ■

References

[1] Maria-Florina Balcan et al. “Communication Efficient Distributed Kernel Principal Component Analysis”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. 2016, pp. 725–734. DOI: 10.1145/2939672.2939796. URL: https://doi.org/10.1145/2939672.2939796.

[2] Christos Boutsidis, David P. Woodruff, and Peilin Zhong. “Optimal principal component analysis in distributed and streaming models”. In: *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*. 2016, pp. 236–249. DOI: 10.1145/2897518.2897646. URL: https://doi.org/10.1145/2897518.2897646.

[3] Dan Feldman, Melanie Schmidt, and Christian Sohler. “Turning Big Data Into Tiny Data: Constant-Size Coresets for k-Means, PCA, and Projective Clustering”. In: *SIAM J. Comput.* 49.3 (2020), pp. 601–657. DOI: 10.1137/18M1209854. URL: https://doi.org/10.1137/18M1209854.

[4] Ravi Kannan, Santosh S. Vempala, and David P. Woodruff. “Principal Component Analysis and Higher Correlations for Distributed Data”. In: *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*. 2014, pp. 1040–1057. URL: http://proceedings.mlr.press/v35/kannan14.html.

[5] David P. Woodruff and Peilin Zhong. “Distributed low rank approximation of implicit functions of a matrix”. In: *32nd IEEE International Conference on Data Engineering, ICDE 2016, Helsinki, Finland, May 16-20, 2016*. 2016, pp. 847–858. DOI: 10.1109/ICDE.2016.7498295. URL: https://doi.org/10.1109/ICDE.2016.7498295.