CS 15-851: Algorithms for Big Data Spring 2024

Lecture 5 (Part one) — Feb 15

Prof. David Woodruff *Scribe: Aashiq Muhamed*

1 Distributed Low Rank Approximation

In machine learning and data-intensive applications, there's a need for algorithms optimized for distributed environments and capable of efficiently processing large datasets. These settings frequently present significant constraints on computational resources, including storage capacity, communication bandwidth, and processing time. Consequently, our focus shifts toward exploring the efficacy and implementation of sketching-based algorithms within these distributed frameworks.

1.1 Communication Models of Low-Rank Approximation

Sketching-based algorithms have been effective for computing rank-k approximations of an input matrix $A \in \mathbb{R}^{n\times d}$. In distributed settings, it is assumed that A is partitioned among s servers, with each server handling a local matrix A^t for $t = 1, \ldots, s$. Within this context, various models are considered.

The *arbitrary partition model* distributes the matrix $A \in \mathbb{R}^{n\times d}$ among s servers as follows:

$$A = A^1 + A^2 + \cdots + A^s \tag{1}$$

We are interested in the cumulative customer product matrix A, which is the summation of matrices from s distinct outlets. It captures scenarios where a customer might purchase a product at one outlet t and the same product at another outlet t', corresponding to the same entry in each outlet's local matrix. To determine the total number of times a product was purchased, the matrices from all outlets are summed.

Alternatively, the less general form, the *row partition model*, is considered with s servers, each holding a subset of rows of A. This model is applicable in scenarios where a product can be purchased exclusively at a particular outlet.

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

The communication model entails that each server communicates with a Coordinator through two-way communication as illustrated in Figure 1. This setup allows for simulating arbitrary point-to-point communication with a factor of 2 increase in the number of communication bits (and an additional $O(\log s)$ bits per message to identify the server). If server t wishes to send a message to server t', it is relayed via the Coordinator, who then forwards it to server t'.

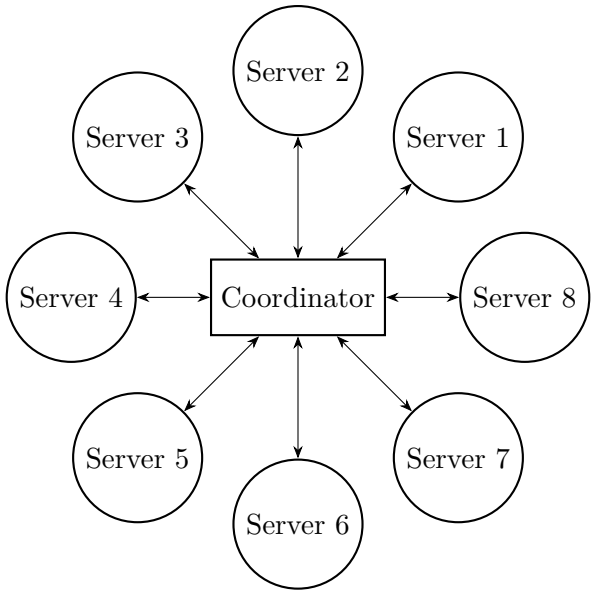

Figure 1: Illustration of the communication model with a coordinator and servers.

1.2 Communication Cost of Low-Rank Matrix Approximation

This section outlines the computational and communication complexities associated with approximating low-rank matrices in a distributed environment.

- **Input:** A matrix $A \in \mathbb{R}^{n \times d}$ is distributed across s servers following the arbitrary partition model. Each server t holds a matrix $A^t \in \mathbb{R}^{n \times d}$, with the matrix entries being $O(\log(nd))$ bit integers. To ensure feasible communication of matrix entries, a bit complexity upper bound is applied to the entries of the matrix.
- **Output:** The servers collaboratively output a k-dimensional subspace W. Let P_W represent the projection matrix onto W, the collective output then is

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

  The goal is to achieve a high-quality low-rank approximation of A. Having the subspace W allows each server to locally project its matrix onto W and represent it with a reduced number of parameters. This reduction is particularly beneficial for applications such as k-means clustering among others.
- **Objective:** The primary objectives include minimizing the total communication overhead, measured in bits, and reducing the communication cost. An additional goal is to decrease the round complexity, or the number of communication rounds required for back-and-forth communication.

1.3 Related Work on Distributed Low-Rank Approximation

We will look at three existing approaches to efficient distributed low-rank approximation:

- The FSS protocol, tailored for the row partition model (3), communicates using $O(skd/\varepsilon)$ real numbers. Given that arbitrary real numbers can encapsulate an indefinite amount of information, this approach raises concerns regarding its practicality. The protocol may suffer from excessive bit complexity and might require considerable computational resources, as it requires performing Singular Value Decomposition (SVD) at each server and an expensive SVD at the coordinator's end.
- The KVW protocol, is applicable to the arbitrary partition model, requires$O(skd/\varepsilon)$ in communication and offers improved computational efficiency.
- The BWZ protocol, also designed for the arbitrary partition model, limits communication to $O(skd) + \text{poly}(sk/\varepsilon)$ words, with the advantage that computations that can be done in input sparsity time.

Remark 1. BWZ protocol (2) achieves the lower bound for the communication cost, at $\Omega(skd)$ words. This bound arises from the need for all s servers to communicate a k-dimensional space, specified by kd words.

Remark 2. Further works have introduced variants of these protocols, with applications extending to kernel low-rank approximation (1), implicit matrices (4), and enhancements for sparsity (2). These adaptations provide a framework for distributed low-rank approximation, catering to distinct data characteristics and computational limitations.

2 The [FSS] Protocol

This section delves into the distributed protocol developed by Feldman, Schmidt, and Sohler (3), hereafter referred to as [FSS]. The key idea in their methodology is the innovative use of coresets.

Definition (Coreset): Given a matrix $A \in \mathbb{R}^{n\times d}$ and its Singular Value Decomposition (SVD) $A = U\Sigma V^T$, for a rank parameter $m = k + \frac{k}{\epsilon}$, let Σ_m match Σ in the top m diagonal entries (representing the highest m singular values) and be zero elsewhere. A coreset in this context is defined as the matrix $\Sigma_m V^T$.

Claim 1: For a matrix A and its coreset $\Sigma_m V^T$, and for all projection matrices $Y = I - X$ onto $(d-k)$-dimensional subspaces,

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

where $c = \|A - A_m\|_F^2$ and is independent of Y.

We can see that X projects onto a k-dimensional subspace and Y onto its complementary space. Given the structure of Σ_m, we can write $\Sigma_m V^T = \Sigma_m V_m^T$ where V_m^T has all but the first m rows zeroed. Therefore, maintaining $\Sigma_m V_m$ is sufficient, which has $md \leq nd$ elements while preserving cost across every k-dimensional space. Drawing a parallel to sketching, envisioning S as U_m^T yields $SA = U_m^T U\Sigma V^T = \Sigma_m V^T$, forming a deterministic sketch.

Theorem: Let $\tilde{Y}$ minimize $\|\Sigma_m V^T Y\|_F^2$ and Y^* minimize $\|AY\|_F^2$. It follows that

$$\|A\tilde{Y}\|_F^2 \leq \|\Sigma_m V^T \tilde{Y}\|_F^2 + c \leq \|\Sigma_m V^T Y^*\|_F^2 + c \leq (1+\epsilon)\|A - A_k\|_F^2.$$

Lemma 1: A projection matrix P does not increase lengths; thus, for a matrix A,

$$\|AP\|_F^2 \leq \|A\|_F^2.$$

Proof: Demonstrating that $\|AY\|_F^2 \leq \|\Sigma_m V^T Y\|_F^2 + c$ requires the following expansion:

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

Here, U's first m columns are considered in the first term, with the remainder in the second. The orthonormality of U's columns means both terms are orthogonal, and we can apply the Pythagorean theorem column-wise. Additionally, by applying Lemma 1 for Y and using the fact that U's orthonormal columns preserve the Frobenius norm, we get

$$\|A\|_F^2 \leq \|\Sigma_m V^T Y\|_F^2 + \|U(\Sigma - \Sigma_m)V^T\|_F^2 = \|\Sigma_m V^T Y\|_F^2 + \|A - A_m\|_F^2 = \|\Sigma_m V^T Y\|_F^2 + c.$$

Here $c = \|A - A_m\|_F^2$.

For the second inequality, showing $\|\Sigma_m V^T Y\|_F^2 + c \leq (1+\epsilon)\|AY\|_F^2$ involves proving:

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

by subtracting $\|AY\|_F^2$ from both sides. With $Y = I - X$, then $\Sigma_m V^T Y + \Sigma_m V^T X = \Sigma_m V^T$, and the orthogonality of X and Y allows row-wise application of the Pythagorean theorem, leading to:

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

Given U's orthonormal columns and the definition of A_m, this simplifies to:

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

Rearranging and recognizing orthogonality between A_m and $(A - A_m)$ yields:

$$\|AX\|_F^2 - \|\Sigma_m V^T X\|_F^2 = \|U(\Sigma - \Sigma_m)V^T X\|_F^2 \leq \|U(\Sigma - \Sigma_m)V^T\|_2^2 \|X\|_F^2,$$

The last inequality follows from submultiplicativity. The first term is in SVD form, so its maximum singular value is σ_{m+1}, equal to its operator norm, and X is a rank-k projection matrix, with k singular values of 1. It follows that

$$= \sigma_{m+1}^2 \sum_{i=1}^{k} 1 = \sigma_{m+1}^2 k = \epsilon\sigma_{m+1}^2(m-k) \leq \epsilon \sum_{i=k+1}^{m} \sigma_i^2 \leq \epsilon \sum_{i=k+1}^{d} \sigma_i^2 = \epsilon\|A - A_k\|_F^2.$$

As $\|A - A_k\|_F^2 = \|AY^*\|_F^2$, we have

$$\epsilon\|AY^*\|_F^2 \leq \epsilon\|AY\|_F^2.$$

which completes the proof. ■

Union of Coresets

One property of coresets that we will use is their composability; specifically, the union of multiple coresets also forms a valid coreset.

Consider a scenario within the row partition model involving matrices $A^1, \ldots, A^s$ and corresponding constructions $\Sigma_m^1 V^{T,1}, \ldots, \Sigma_m^s V^{T,s}$, alongside constants $c_1, \ldots, c_s$. A unified coreset can be derived by merging the rows from $A^1, \ldots, A^s$ and applying coreset inequalities on a row-by-row basis, yielding:

$$\sum_i (\|\Sigma_m^i V^{T,i} Y\|_F^2 + c_i) = (1 \pm \epsilon) \sum_{i=1}^{s} (\|A^i Y\|_F^2) = (1 \pm \epsilon) \|AY\|_F^2.$$

Furthermore, let B denote the matrix formed by concatenating the rows of $\Sigma_m^1 V^{T,1}, \ldots, \Sigma_m^s V^{T,s}$. Upon computing $B = U\Sigma V^T$ and determining a coreset for B, $\Sigma_m V^T$, with $c = \|B - B_m\|_F^2$, it follows that:

$$\|\Sigma_m V^T Y\|_F^2 + c + \sum_i c_i = (1 \pm \epsilon) \|BY\|_F^2 + \sum_i c_i = (1 \pm O(\epsilon)) \|AY\|_F^2.$$

Hence, $\Sigma_m V^T$ constitutes a coreset for A, parameterized by $c + \sum_i c_i$.

[FSS] Protocol

We can now define the [FSS] protocol for the row-partition model, which operates as follows:

- Each server t transmits the top $m = \frac{k}{\epsilon} + k$ principal components of A^t, adjusted by their singular values, along with $c_t = \|A - A_m\|_F^2$.
- The Coordinator aggregates these components and $c + \sum_{t=1}^{s} c_t$, disseminating the collective top m components of $[\Sigma^1 V^1; \Sigma^2 V^2; \ldots; \Sigma^s V^s]$ back to the servers.

However, this protocol suffers from several limitations: Firstly, it requires the communication of sdk/ϵ real numbers, and real numbers have high bit complexity. Second, the protocol mandates the execution of SVD operations on each server and an expensive SVD at the Coordinator's end. Lastly, its design is incompatible with the arbitrary partition model.

Leveraging random matrix techniques within this SVD-based protocol offers a pathway to further reducing computational complexity that we look at next.

3 KVW Protocol

In response to the limitations identified in the FSS protocol, we explore a solution proposed by Kannan, Vempala, and Woodruff (5) (KVW). This protocol addresses nearly all the previously mentioned concerns, except the optimal communication complexity issue which will be addressed by the BWZ protocol. The KVW protocol emerges from the realization that the randomized sketching techniques discussed in our lectures, particularly those related to SVD, hold promise for mitigating these challenges. Our analysis aims to formalize and expand this preliminary observation.

We study arbitrary partition model, where we aim to leverage matrix sketching algorithms to circumvent the need for expensive SVD computations. Let S be a random sketching matrix of dimensions $\frac{k}{\epsilon} \times n$. The protocol encompasses the following steps:

1. The coordinator selects a random seed and disseminates it to all servers, enabling the pseudo-random generation of the sketching matrix S.

2. Each server t calculates SA^t and forwards this to the coordinator.

3. Upon receiving SA^t from each server, the coordinator aggregates these to compute $\sum_{t=1}^{s} SA^t$, equivalent to SA under the arbitrary-partition model. This operation exploits the linearity of sketch matrices to identify a suitable k-dimensional subspace W within the row span of SA. Ideally, each server t would then project A^t onto W, outputting this projection matrix.

However, isolating the k-dimensional subspace applicable to all A proves challenging. Direct server output of A^t projections onto SA's row span does not meet the required rank-k constraint. While the coordinator could theoretically identify and communicate a k-dimensional subspace within these projections, doing so would undesirably link communication complexity to n.

To address this, consider a matrix $Z \in \mathbb{R}^{k \times k/\epsilon}$ that transforms ZSA into a rank-k orthonormal matrix, enabling the definition of projection $P = (ZSA)^T(ZSA) = (SA)^T Z^T ZSA$ by setting $Z^T Z = X$. The search for a k-dimensional subspace within SA is reformulated as minimizing

$$\min_{\text{rank-k } X} \|A(SA)^T X(SA) - A_k\|_F^2 \leq (1+\epsilon)\|A - A_k\|_F^2,$$

where X is a matrix of polynomial dimensions in terms of k/ϵ.

To solve the problem approximately, we employ affine embeddings T_1 and T_2, introduced in a previous lecture, each with $\text{poly}(k/\varepsilon)$ rows and columns respectively. This transforms our initial problem into:

$$\min_{\text{rank-k } X} \|T_1 A(SA)^T X(SA) T_2 - T_1 A T_2\|_F^2.$$

In the arbitrary partition model, the coordinator first sends the seeds for S, T_1, and T_2 to all servers. Each server t computes SA^t and forwards it to the coordinator. The coordinator aggregates these to compute $SA = \sum_t SA^t$ and redistributes SA to all servers. Subsequently, each server t computes $T_1 A^t (SA)^T$ and $T_1 A^t T_2$, sending these back to the coordinator. The coordinator then aggregates these to compute $T_1 A(SA)^T = \sum_t T_1 A^t (SA)^T$ and $T_1 A T_2 = \sum_t T_1 A^t T_2$. The coordinator can then either find $\tilde{X}$ satisfying

$$\tilde{X} = \arg\min_{\text{rank-k } X} \|T_1 A(SA)^T X(SA) T_2 - T_1 A T_2\|_F^2,$$

and communicate $\tilde{X}$ to the servers or communicate the aggregated matrices for the servers to compute $\tilde{X}$. By the properties of affine embedding, it is guaranteed that

$$\|A(SA)^T \tilde{X}(SA) - A\|_F^2 \leq (1 + O(\varepsilon))\|A - A_k\|_F^2.$$

This method maintains bit complexity while achieving an overall communication cost of $O\left(sdk/\epsilon + s \cdot \text{poly}(k/\epsilon)\right)$, where sdk/ϵ accounts for the communication of SA and the rest corresponds to exchanges involving $\text{poly}(k/\epsilon) \times \text{poly}(k/\epsilon)$ matrices.

References

[1] Maria-Florina Balcan, Yingyu Liang, Le Song, David P. Woodruff, and Bo Xie. *Distributed kernel principal component analysis.* arXiv preprint arXiv:1503.06858, 2015.

[2] Christos Boutsidis, David P. Woodruff, and Peilin Zhong. *Optimal principal component analysis in distributed and streaming models*. In Proceedings of the forty-eighth annual ACM symposium on Theory of Computing, pages 236–249, 2016.

[3] Dan Feldman, Melanie Schmidt, and Christian Sohler. *Turning big data into tiny data: Constant-size coresets for k-means, PCA, and projective clustering*. SIAM Journal on Computing, 49(3):601–657, 2020.

[4] David P. Woodruff and Peilin Zhong. *Distributed low rank approximation of implicit functions of a matrix*. In 2016 IEEE 32nd International Conference on Data Engineering (ICDE), pages 847–858. IEEE, 2016.

[5] 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), volume 35 of JMLR Workshop and Conference Proceedings, pages 1040–1057. JMLR.org, 2014.