CS 15-859: Algorithms for Big Data Fall 2019

Lecture 5 - Part 2 — 10/3/2019

Prof. David Woodruff *Scribe: Neil Xu*

1 A More Efficient Distributed Low Rank Approximation Protocol [BWZ]

In the previous protocol [KVW] we achieved communication complexity of $O(skd/\epsilon) + \text{poly}(sk/\epsilon)$. The skd/ϵ term of this complexity results from from each server transmitting SA_t, which is of dimension $k/\epsilon \times d$.

The protocol from [BWZ] improves on [KVW] by sketching the SA_t matrix again before transmission, which reduces the communication complexity to $O(skd + \text{poly}(k/\epsilon))$.

1.1 Projection-cost preserving sketches

We introduce a new class of sketching matrices (which all of the sampling matrices we've discussed are a part of e.g. CountSketch, ind. Gaussian, subsampled randomized Haadamard etc.).

Definition. Let A be a $n \times d$ matrix, S be a random $k/\epsilon^2 \times n$ matrix.

S is a **projection-cost preserving (PCP) sketch** if there is a scalar $c \geq 0$ s.t. for all k-dimensional projection matrices P:

$$|SA(I-P)|_F^2 + c = (1 \pm \epsilon)\, |A(I-P)|_F^2$$

For more information on what sketching matrices qualifies as a projection-cost preserving sketch, refer to [CEM$^+$]

1.2 Protocol

To rehash the goal, we want to find a k-dimensional space W s.t. for it's projection matrix P_W:

$$\left| A - \sum_t A^t P_W \right|_F^2 \leq (1+\epsilon)\, |A - A_k|_F^2$$

where

$$A_k = \underset{A' \in \text{k-rank matrices}}{\text{argmin}} |A - A'|_F$$

Thus, our protocol is:

Let S and T be $k/\epsilon^2 \times n$ and $d \times k/\epsilon^2$ PCP sketches, respectively.
Server t sends SA^tT to the coordinator.
Coordinator sends back $SAT = \sum_t SA^tT$
Each server computes $k/\epsilon^2 \times k$ matrix U of top k left singular vectors of SAT.
Server t sends $U^T SA^t$ to coordinator.
Coordinator returns the space $W = U^T SA = \sum_t U^T SA^t$

Note that in this algorithm SA^tT, SAT are $k/\epsilon^2 \times k/\epsilon^2$, and $U^T SA^t, U^T SA$ are $k \times d$. Since these matrices encompass all of the communication, this gives us the communication complexity bound of $O(sdk) + \text{poly}(k/\epsilon^2)$.

Intuitively, what we're doing with U is selecting the top k left singular values of SA, and then $U^T SA$ would analogously be the top k right singular vectors of SA.

Of course, U isn't the top left singular values of SA, but rather of SAT. We will show how the PCP property of our sketching matrices account for U not quite being the top k left singular vectors of SA.T

1.3 Approximation error

Let W be the row span of $U^T SA$ and P be the corresponding projection matrix.

We want to show that

$$|A - AP|_F^2 \leq (1 + O(\epsilon)) \, |A - A_k|_F^2$$

We will do this by justifying the following inequality

$$\begin{aligned}
(1-\epsilon) \, |A - AP|_F^2 &\leq |SA - SAP|_F^2 + c && (1) \\
&\leq \left| SA - UU^T SA \right|_F^2 + c && (2) \\
&\leq \frac{1+\epsilon}{1-\epsilon} \left(|SA - [SA]_k|_F^2 + c \right) && (3) \\
&\leq \frac{1+\epsilon}{1-\epsilon} \left(\left| SA - SAP_{[A]_k} \right|_F^2 + c \right) && (4) \\
&\leq \frac{(1+\epsilon)^2}{1-\epsilon} \, |A - [A]_k|_F^2 && (5)
\end{aligned}$$

where $P_{[A]_k}$ is the projection matrix s.t. $P_{[A]_k} A = [A]_k$.

(1) S is a PCP matrix.

(2) SAP is the matrix in the space W that is the shortest distance to SA, since P is the projection matrix for W. The row space of $U^T SA$ is W by definition, so $U(U^T SA)$ is a matrix in W.

Thus, the distance between SA and $U(U^T SA)$ must at least be the distance between SA and SAP.

(3) We can first observe that UU^T is the optimal k-rank projection for SAT since U is the top k left singular vectors of SAT. Thus,

$$\left|SAT - UU^T SAT\right|_F^2 = |SAT - [SAT]_k|_F^2$$

By T being a PCP matrix:

$$(1-\epsilon)\left|SA - UU^T SA\right|_F^2 \le \left|SAT - UU^T SAT\right|_F^2 + c_T$$
$$\left|(I - P_{[SA]_k})SAT\right|_F^2 + c \le (1+\epsilon)\,|SA - [SA]_k|_F^2$$

where $P_{[SA]_k}$ is the projection matrix s.t. $P_{[SA]_k} SA = [SA]_k$.
We also know by definition of $[SAT]_k$:

$$|SAT - [SAT]_k|_F^2 = \min_{P' \text{ is rank } k} |(I - P')SAT)|_F^2$$

that:

$$|SAT - [SAT]_k|_F^2 \le \left|(I - P_{[SA]_k})SAT\right|_F^2$$

Thus, we can string our inequalities together to get:

$$\begin{aligned}(1-\epsilon)\left|SA - UU^T SA\right|_F^2 &\le \left|SAT - UU^T SAT\right|_F^2 + c_T\\ &= |SAT - [SAT]_k|_F^2 + c_T\\ &\le \left|(I - P_{[SA]_k})SAT\right|_F^2 + c_T\\ &\le (1+\epsilon)\,|SA - [SA]_k|_F^2\end{aligned}$$

Thus

$$\left|SA - UU^T SA\right|_F^2 \le \frac{1+\epsilon}{1-\epsilon}\,|SA - [SA]_k|_F^2$$

and we can push in the c term.

$$\left|SA - UU^T SA\right|_F^2 + c \le \frac{1+\epsilon}{1-\epsilon}\,|SA - [SA]_k|_F^2 + c \le \frac{1+\epsilon}{1-\epsilon}\left(|SA - [SA]_k|_F^2 + c\right)$$

(4) Similar to (2), we know that by definition:

$$|SA - [SA]_k|_F^2 = \min_{P' \text{ of rank } k} |SA - SAP'|_F^2$$

Thus,

$$|SA - [SA]_k|_F^2 \le \left|SA - SAP_{[A]_k}\right|_F^2$$

(5) S is still a PCP matrix.

Thus

$$|A - AP|_2^F \leq \frac{(1+\epsilon)^2}{(1-\epsilon)^2} |A - [A]_k|_F^2$$

With sufficiently small ϵ, $\frac{(1+\epsilon)^2}{(1-\epsilon)^2} = 1 + O(\epsilon)$, so we have achieved our desired approximation bound.

1.4 Conclusions

The above analysis of the protocol from [BWZ] does not account for the bit complexity of transmitted matrices, but they can bit complexity can be bounded appropriately by adding noise to the true values. Another implicit advantage of this protocol is that matrices can be computed in input sparsity time (e.g. when using a CountSketch matrix). This protocol also closes the gap with the lower bound on the number of rounds of communication, 2, where each round of communication consists of the server sending data to the coordinator and the coordinator sending data back to the servers.

In addition to the low rank approximation problem in the distributed setting, there are other problems that can also be explored e.g.

- Computing the rank of an n x n matrix over the reals
- Linear programming
- Graph problems: matching
- etc.

2 L1 Regression

The L1 regression is defined as finding:

$$x^* = \underset{x}{\operatorname{argmin}}\ |Ax - b|_1 = \underset{x}{\operatorname{argmin}}\ \sum_{i=1}^{n} |b_i - \langle A_i, x \rangle|$$

If we imagine the hyperplane that contains $(A_i, \langle A_i, x \rangle)$, the L1 distance is simply the distance of each of the points to the corresponding (A_i, b_i). The 1 norm on this penalizes further points less relatively when compared to the 2 norm, since it doesn't square distances.

2.1 Linear Programming

We can solve the L1 problem with a LP of the form:

$$\min \mathbf{1}^T(\alpha^+ + \alpha^-)$$
$$Ax + \alpha^+ - \alpha^- = b$$
$$\alpha^+, \alpha^- \geq \mathbf{0}$$

Consider row i. If $\langle A_i, x\rangle - b_i < 0$, $\alpha_i^+ + \alpha_i^-$ is minimized if $\alpha_i^+ = 0, \alpha_i^- = b_i - \langle A_i, x\rangle$, and vice versa.

Thus, for whatever x is $(\alpha^+ + \alpha^-)_i = |\langle A_i, x\rangle - b_i|$.

This formulation. however, has poly(nd) runtime.

2.2 Sketching and Well-Conditioned Bases

Motivation Using sketching, we want to achieve a faster runtime. In fact, we've solved regression problem already with sketching albeit on the 2 norm. In proving the subspace embedding property, we decomposed $A = UW$ such that for all x, $|Ux|_2 = |x|$ i.e. U is an orthonormal basis for the column space of A. Our goal is to derive something similar for L1 Regression i.e. find $A = UW$ such that for all x, $|Ux|_1 \approx |x|$.

Well-conditioned bases Let $A = QW$, where Q has full column rank.

Let

$$|z|_{Q,1} = |Qz|_1$$

Note that $|z|_{Q,1}$ is a norm since:

- Positive definite: Q is full rank so $Qz = 0 \implies z = 0$. So $|z|_{Q,1} = 0 \implies z = 0$.
- Absolute homogeneity: $|az|_{Q,1} = |Qaz|_1 = |a|_1 |Qz|_1 = |a|_1 |z|_{Q,1}$.
- Subadditivity: $|z_1 + z_2|_{Q,1} = |Q(z_1 + z_2)|_1 = |Qz_1 + Qz_2|_1$. By triangle inequality of the 1 norm, we get $|Qz_1 + Qz_2|_1 \leq |Qz_1|_1 + |Qz_2|_1 = |z_1|_{Q,1} + |z_2|_{Q,2}$.

Define

$$C = \{z \in \mathbb{R}^d : |z|_{Q,1} \leq 1\}$$

i.e. the unit ball under the $Q, 1$ norm.

Note that C is a convex set that symmetric around the origin. The symmetry arises from the fact absolute homogeneity of the norm i.e. if $|z|_{Q,1} \leq 1$ then $|-z|_{Q,1} = |z|_{Q,1} \leq 1$.

Convexity arises from the fact that for $z_1, z_2 \in C$, we can observe that for $t \in [0, 1]$ that $|tz_1 + (1-t)z_2|_{Q,1} \leq |tz_1|_{Q,1} + |(1-t)z_2|_{Q,1} = t\,|z_1|_{Q,1} + (1-t)\,|z_2|_{Q,1} \leq t + (1-t) = 1$ by using the norm properties.

Theorem 1. ***Lowner-John theorem*** *states that for C that is convex and symmetric around the origin, there exists ellipsoid*

$$E = \{z \in \mathbb{R}^d : z^T F z \leq 1\}$$

where $F = GG^T$ is a symmetric matrix, such that $E \subseteq C \subseteq \sqrt{d}E$.

Corollary 1. *For all $z \in \mathbb{R}^d$: $\sqrt{z^T F z} \leq |z|_{Q,1} \leq \sqrt{d z^T F z}$.*

Using this notion of ellipsoid, we can define $U = QG^{-1}$ as our **well-conditioned basis**.

Thus,

$$|Ux|_1 = \left|QG^{-1}x\right|_1$$

and set $z = G^{-1}x$ which means

$$|Ux|_1 = |Qz|_1 = |z|_{Q,1}$$

We note that

$$\begin{aligned} z^T F z = x^T (G^{-1})^T G^T G G^{-1} x = x^T x = |x|_2 \\ |z|_{Q,1} = \left|QG^{-1}x\right|_1 = |Ux|_1 \end{aligned}$$

and when combined with Corollary 1, we get that

$$|x|_2 \leq |Ux|_1 \leq \sqrt{d}\,|x|_2$$

By triangle inequality of Euclidean space, we know that $|x|_1 \geq |x|_2$.

Also note for arbitrary $a, b \in \mathbb{R}$ we see that $a^2 + b^2 \geq 2ab$ since $a^2 + b^2 - 2ab = (a-b)^2 \geq 0$.

$$\begin{aligned} d\,|x|_2^2 = d\sum_{i=1}^{d} x_i^2 = \sum_{i=1}^{d} x_i^2 + \sum_{i=1}^{d}\sum_{j=i+1}^{d}\left(x_i^2 + x_j^2\right) \\ |x|_1^2 = \left(\sum_{i=1}^{d}|x_i|\right)^2 = \sum_{i=1}^{d} x_i^2 + \sum_{i=1}^{d}\sum_{j=i+1}^{d} 2\,|x_i|\,|x_j| \end{aligned}$$

Thus, $d\,|x|_2^2 \geq |x|_1^2$, which gives us the following relationship between 1 and 2 norms:

$$\frac{|x|_1}{\sqrt{d}} \leq |x|_2 \leq |Ux|_1 \leq \sqrt{d}\,|x|_2 \leq \sqrt{d}\,|x|_1$$

References

[BWZ] Christos Boutsidis, David P. Woodruff, and Peilin Zhong. Optimal principal component analysis in distributed and streaming models.

[CEM+] Michael B. Cohen, Sam Elder, Cameron Musco, Christopher Musco, and Madalina Persu. Dimensionality Reduction for k-Means Clustering and Low Rank Approximation.

[KVW] Ravindran Kannan, Santosh S Vempala, and David P Woodruff. Principal Component Analysis and Higher Correlations for Distributed Data. page 17.