THEORY OF COMPUTING, Volume 16 (15), 2020, pp. 1–25
www.theoryofcomputing.org

SPECIAL ISSUE: APPROX-RANDOM 2016

Online Row Sampling

Michael B. Cohen* Cameron Musco† Jakub Pachocki‡

Received April 14, 2017; Revised April 10, 2020; Published December 11, 2020

Abstract. Finding a small spectral approximation for a tall $n \times d$ matrix $\mathbf{A}$ is a fundamental numerical primitive. For a number of reasons, one often seeks an approximation whose rows are sampled from those of $\mathbf{A}$. Row sampling improves interpretability, saves space when $\mathbf{A}$ is sparse, and preserves structure, which is important, e. g., when $\mathbf{A}$ represents a graph.

However, correctly sampling rows from $\mathbf{A}$ can be costly when the matrix is large and cannot be stored and processed in memory. Hence, a number of recent publications focus on row sampling in the streaming setting, using little more space than what is required to store the returned approximation (Kelner–Levin, *Theory Comput. Sys.* 2013, Kapralov et al., *SIAM J. Comp.* 2017).

Inspired by a growing body of work on online algorithms for machine learning and data analysis, we extend this work to a more restrictive *online* setting: we read rows of $\mathbf{A}$ one by one and immediately decide whether each row should be kept in the spectral approximation or discarded, without ever retracting these decisions. We present an extremely

A preliminary version of this paper appeared in the Proceedings of the 19th International Workshop on Approximation Algorithms for Combinatorial Optimization Problems (APPROX 2016).

*Work supported in part by NSF grant CCF-1111109.

†Work supported by NSF Graduate Research Fellowship No. 1122374, AFOSR grant FA9550-13-1-0042 and the NSF Center for Science of Information.

‡Work supported by NSF grant CCF-1065106.

ACM Classification: F.2.1, F.1.2, G.2.2

AMS Classification: 68W25, 68W20, 68R10

Key words and phrases: spectral sparsification, leverage scores, online algorithms, matrix approximation

© 2020 Michael B. Cohen, Cameron Musco, and Jakub Pachocki
Licensed under a Creative Commons Attribution License (CC-BY)
DOI: 10.4086/toc.2020.v016a015

simple algorithm that approximates $\mathbf{A}$ up to multiplicative error $1+\varepsilon$ and additive error δ using $\mathcal{O}(d \log d \log(\varepsilon \|\mathbf{A}\|_2^2/\delta)/\varepsilon^2)$ online samples, with memory overhead proportional to the cost of storing the spectral approximation. We also present an algorithm that uses $\mathcal{O}(d^2)$ memory but only requires $\mathcal{O}(d \log(\varepsilon \|\mathbf{A}\|_2^2/\delta)/\varepsilon^2)$ samples, which we show is optimal.

Our methods are clean and intuitive, allow for lower memory usage than prior work, and expose new theoretical properties of leverage score based matrix approximation.

1 Introduction

1.1 Background

A spectral approximation to a tall $n \times d$ matrix $\mathbf{A}$ is a smaller, typically $\tilde{\mathcal{O}}(d) \times d$ matrix $\tilde{\mathbf{A}}$ such that $\|\tilde{\mathbf{A}}\mathbf{x}\|_2 \approx \|\mathbf{A}\mathbf{x}\|_2$ for all $\mathbf{x}$. Typically one asks for a multiplicative approximation, which guarantees that $(1-\varepsilon)\|\mathbf{A}\mathbf{x}\|_2^2 \le \|\tilde{\mathbf{A}}\mathbf{x}\|_2^2 \le (1+\varepsilon)\|\mathbf{A}\mathbf{x}\|_2^2$. In other notation,

$$(1-\varepsilon)\mathbf{A} \preceq \tilde{\mathbf{A}} \preceq (1+\varepsilon)\mathbf{A}.$$

Such approximations have many applications, most notably for solving least squares regression over $\mathbf{A}$ [9, 11]. If $\mathbf{A}$ is the vertex-edge incidence matrix of a graph, $\tilde{\mathbf{A}}$ is a *spectral sparsifier* [26]. It can be used to approximate effective resistances, spectral clustering, mixing time and random walk properties, and many other computations.

A number of recent papers focus on fast algorithms for spectral approximation. Using sparse random subspace embeddings [9, 23, 22], it is possible to find $\tilde{\mathbf{A}}$ in *input sparsity time*—i. e., running time scaling linearly in the number of nonzero entries in $\mathbf{A}$. These methods produce $\tilde{\mathbf{A}}$ by randomly recombining the rows of $\mathbf{A}$ into a smaller number of rows. In some cases these embeddings are not enough, as it is desirable for the rows of $\tilde{\mathbf{A}}$ to be a subset of the rows of $\mathbf{A}$. If $\mathbf{A}$ is sparse, this ensures that $\tilde{\mathbf{A}}$ is also sparse. If $\mathbf{A}$ represents a graph, it ensures that $\tilde{\mathbf{A}}$ is also a graph, specifically a weighted subgraph of the original.

It is well known that sampling $\mathcal{O}(d \log d/\varepsilon^2)$ rows of $\mathbf{A}$ with probabilities proportional to their *leverage scores* yields a $(1 \pm \varepsilon)$-factor spectral approximation to $\mathbf{A}$. Further, this sampling can be done in input sparsity time, either using subspace embeddings to approximate leverage scores, or using iterative sampling techniques [20], some that only work with subsampled versions of the original matrix [11].

1.2 Streaming and online row sampling

When $\mathbf{A}$ is very large, input sparsity running times are not enough—memory restrictions also become important. Hence, recent work has tackled row sampling in a streaming model of computation. [16] presents a simple algorithm for sampling rows from an insertion-only stream, using space approximately proportional to the size of the final approximation. [15] gives a sparse-recovery based algorithm that works in dynamic streams with row insertions and deletions, also using nearly optimal space. Unfortunately, to handle dynamic streams, the algorithm in [15] is complex, requires additional restrictions on the input matrix, and uses significantly suboptimal running time to recover a spectral approximation from its low memory representation of the input stream.

While the algorithm in [16] is simple and efficient, we believe that its proof is incomplete, and do not see an obvious way to fix it. The main idea behind the algorithm is to sample rows by their leverage scores with respect to the stream seen so far. These leverage scores may be coarse overestimates of the true scores. However as more rows are streamed in, better estimates can be obtained and the sampled rows pruned to a smaller set. Unfortunately, the probability of sampling a row becomes dependent on which other rows are sampled. This seems to break the argument in that paper, which essentially claims that their process has the same distribution as would a single round of leverage score sampling.[1]

In this paper we initiate the study of row sampling in an *online setting*. As in an insertion stream, we read rows of $\mathbf{A}$ one by one. However, upon seeing a row, we immediately decide whether it should be kept in the spectral approximation or discarded, without ever retracting these decisions. We present a similar algorithm to [16], however, since we never prune previously sampled rows, the probability of sampling a row only depends on whether previous rows in the stream were sampled. This limited dependency structure allows us to rigorously argue that a spectral approximation is obtained.

In addition to addressing gaps in the literature on streaming spectral approximation, our restricted model extends work on online algorithms for a variety of other machine learning and data analysis problems, including principal component analysis [4], clustering [21], classification [3, 14], and regression [14]. In practice, online algorithms are beneficial since they can be highly computationally and memory efficient. Further, they can be applied in scenarios in which data is produced in a continuous stream and intermediate results must be output as the stream is processed. Spectral approximation is a widely applicable primitive for approximate learning and computation, so studying its implementation in an online setting is a natural direction. Since the initial publication of this work, online row sampling methods have found applications in kernel matrix approximation [7, 8] and sliding window algorithms for streaming matrix approximation [6].

1.3 Our results

Our primary contribution is a very simple algorithm for leverage score sampling in an online manner. The main difficultly with row sampling using leverage scores is that leverage scores themselves are not easy to compute. They are given by $l_i = \mathbf{a}_i^T(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{a}_i$, and so require solving systems in $\mathbf{A}^T\mathbf{A}$ if computed naively. This is not only expensive, but also impossible in an online setting, where we do not have access to all of $\mathbf{A}$.

A critical observation is that it always suffices to sample rows by overestimates of their true leverage scores. The number of rows that must be sampled is proportional to the sum of these overestimates. Since the leverage score of a row can only go up when we remove rows from the matrix, a simple way to obtain an overestimate is to compute leverage score using just a subset of the other rows of $\mathbf{A}$. That is, letting $\mathbf{A}_j$ contain just j of the n rows of $\mathbf{A}$, we can overestimate l_i by $\tilde{l}_i = \mathbf{a}_i^T(\mathbf{A}_j^T\mathbf{A}_j)^{-1}\mathbf{a}_i$

[11] shows that if $\mathbf{A}_j$ is a subset of rows sampled uniformly at random, then the expected leverage score of $\mathbf{a}_i$ is d/j. This simple fact immediately gives a result for online sampling from a *randomly ordered stream*. If we compute the leverage score of the current row $\mathbf{a}_i$ against all previously seen rows (or some approximation to these rows), then the expected sum of our overestimates is bounded by $d + d/2 + \cdots + \cdots + d/n = \mathcal{O}(d\log n)$. So, sampling $\mathcal{O}(d\log d\log n/\varepsilon^2)$ rows is enough obtain a $(1+\varepsilon)$ multiplicative-error spectral approximation.

[1] Since the initial publication this work, the independence issue in [16] has been resolved by [18], which presents a similar algorithm for insertion-only streams that admits a correct proof.

What if we cannot guarantee a randomly ordered input stream? Is there any hope of being able to compute good leverage score estimates in an online manner? Surprisingly the answer to this is yes—we can in fact run nearly the exact same algorithm and be guaranteed that the sum of estimated leverage scores is low, *regardless of stream order*. Roughly, each time we receive a row which has high leverage score with respect to the previous rows, it must compose a significant part of the spectrum of $\mathbf{A}$. If $\mathbf{A}$ does not continue to grow unboundedly, there simply cannot be too many of these significant rows.

Specifically, we show that if we sample by the *ridge leverage scores* [1] over all previously seen rows, which are the leverage scores computed over $\mathbf{A}_i^T\mathbf{A}_i + \lambda\mathbf{I}$ for some small regularizing factor λ, then with just $\mathcal{O}(d\log d\log(\varepsilon\|\mathbf{A}\|_2^2/\delta)/\varepsilon^2)$ samples we obtain a $(1+\varepsilon)$ multiplicative-error, δ additive-error spectral approximation. That is, with high probability we sample a matrix $\tilde{\mathbf{A}}$ with $(1-\varepsilon)\mathbf{A}^T\mathbf{A}-\delta\mathbf{I} \preceq \tilde{\mathbf{A}}^T\tilde{\mathbf{A}} \preceq (1+\varepsilon)\mathbf{A}^T\mathbf{A}+\delta\mathbf{I}$.

To gain intuition behind this bound, note that we can convert it into a multiplicative one by setting $\delta = \varepsilon\sigma_{min}(\mathbf{A})^2$ where $\sigma_{min}(\mathbf{A})$ is the minimum singular value of $\mathbf{A}$ (as long as we have some estimate of $\sigma_{min}(\mathbf{A})$). This setting of δ will require taking $\mathcal{O}(d\log d\log(\kappa(\mathbf{A}))/\varepsilon^2)$ samples, where $\kappa(\mathbf{A}) = \sigma_{max}(\mathbf{A})/\sigma_{min}(\mathbf{A})$ is the condition number of $\mathbf{A}$. If we have a polynomial bound on this condition number, as we do, for instance, for graphs with polynomially bounded edge weights, this becomes $\mathcal{O}(d\log^2 d/\varepsilon^2)$—nearly matching the $\mathcal{O}(d\log d/\varepsilon^2)$ achievable if sampling by true leverage scores.

Our online sampling algorithm is extremely simple. When each row comes in, we compute the online ridge leverage score, or an estimate of it, and then irrevocably either add the row to our approximation or remove it. As mentioned, it is similar in form to the streaming algorithm of [16], except that it does not require pruning previously sampled rows. This allows us to avoid difficult dependency issues. Additionally, without pruning, we do not even need to store all previously sampled rows. As long as we store a constant-factor spectral approximation our previous samples, we can compute good approximations to the online ridge leverage scores. In this way, we can store just $\mathcal{O}(d\log d\log(\varepsilon\|\mathbf{A}\|_2^2/\delta))$ rows in working memory ($\mathcal{O}(d\log^2 d)$ if we want a spectral graph sparsifier), filtering our input stream into an $\mathcal{O}(d\log d\log(\kappa(\mathbf{A}))/\varepsilon^2)$-size output stream. Note that this memory bound in fact *improves* as ε decreases, and regardless, can be significantly smaller than the output size of the algorithm.

In addition to our main sampling result, we use our bounds on online ridge leverage score approximations to show that an algorithm in the style of [2] allows us to remove a $\log d$ factor and sample just $\mathcal{O}(d\log(\varepsilon\|\mathbf{A}\|_2^2/\delta)/\varepsilon^2)$ rows (Theorem 4.1). This algorithm is more complex and can require $\mathcal{O}(d^2)$ working memory. However, in Theorem 5.1 we show that it is asymptotically optimal. The $\log(\varepsilon\|\mathbf{A}\|_2^2/\delta)$ factor is not an artifact of our analysis, but is truly the cost of the restricting ourselves to online sampling. No algorithm can obtain a multiplicative $(1+\varepsilon)$ additive δ spectral approximation taking fewer than $\Omega(d\log(\varepsilon\|\mathbf{A}\|_2^2/\delta)/\varepsilon^2)$ rows in an online manner.

2 Overview

Let $\mathbf{A}$ be an $n\times d$ matrix with rows $\mathbf{a}_1,\ldots,\mathbf{a}_n$. A natural approach to row sampling from $\mathbf{A}$ is picking an *a priori* probability with which each row is kept, and then deciding whether to keep each row independently. A common choice is for the sampling probabilities to be proportional to the *leverage scores* of the rows. The leverage score of the i-th row of $\mathbf{A}$ is defined to be

$$\mathbf{a}_i^T(\mathbf{A}^T\mathbf{A})^{\dagger}\mathbf{a}_i\,,$$

where the dagger symbol denotes the pseudoinverse. In this work, we will be interested in approximating $\mathbf{A}^T\mathbf{A}$ with some (very) small multiple of the identity added. Hence, we will be interested in the *λ-ridge leverage scores* [1]:

$$\mathbf{a}_i^T(\mathbf{A}^T\mathbf{A}+\lambda\mathbf{I})^{-1}\mathbf{a}_i\,,$$

for a parameter $\lambda > 0$.

In many applications, obtaining the (nearly) exact values of $\mathbf{a}_i^T(\mathbf{A}^T\mathbf{A}+\lambda\mathbf{I})^{-1}\mathbf{a}_i$ for sampling is difficult or outright impossible. A key idea is that as long as we have a sequence $l_1,\ldots,l_n$ of *overestimates* of the λ-ridge leverage scores, that is, for $i = 1,\ldots,n$,

$$l_i \geq \mathbf{a}_i^T(\mathbf{A}^T\mathbf{A}+\lambda\mathbf{I})^{-1}\mathbf{a}_i\,,$$

we can sample by these overestimates and obtain rigorous guarantees on the quality of the obtained spectral approximation. This notion is formalized in Theorem 2.1.

Theorem 2.1. *Let $\mathbf{A}$ be an $n\times d$ matrix with rows $\mathbf{a}_1,\ldots,\mathbf{a}_n$. Let $\varepsilon\in(0,1),\delta>0,\lambda:=\delta/\varepsilon,c:=8\log d/\varepsilon^2$. Assume we are given $l_1,\ldots,l_n$ such that for all $i=1,\ldots,n$,*

$$l_i \geq \mathbf{a}_i^T(\mathbf{A}^T\mathbf{A}+\lambda\mathbf{I})^{-1}\mathbf{a}_i\,.$$

For $i=1,\ldots,n$, let $p_i:=\min(cl_i,1)$. Construct $\tilde{\mathbf{A}}$ by independently sampling each row $\mathbf{a}_i$ of $\mathbf{A}$ with probability p_i, and rescaling it by $1/\sqrt{p_i}$ if it is included in the sample. Then, with high probability,

$$(1-\varepsilon)\mathbf{A}^T\mathbf{A}-\delta\mathbf{I}\preceq\tilde{\mathbf{A}}^T\tilde{\mathbf{A}}\preceq(1+\varepsilon)\mathbf{A}^T\mathbf{A}+\delta\mathbf{I},$$

and the number of rows in $\tilde{\mathbf{A}}$ is $\mathcal{O}\left((\sum_{i=1}^n l_i)\log d/\varepsilon^2\right)$.

Proof. This sort of guarantee for leverage score sampling is well known. See for example Lemma 4 of [11]. If we sampled both the rows of $\mathbf{A}$ and the rows of $\sqrt{\lambda}\mathbf{I}$ with the leverage scores over $(\mathbf{A}^T\mathbf{A}+\lambda\mathbf{I})$, we would have $(1-\varepsilon)(\mathbf{A}^T\mathbf{A}+\lambda\mathbf{I})\preceq\tilde{\mathbf{A}}^T\tilde{\mathbf{A}}\preceq(1+\varepsilon)(\mathbf{A}^T\mathbf{A}+\lambda\mathbf{I})$. However, we do not sample the rows of the identity. Since we could have sampled them each with probability 1, we can simply subtract $\lambda\mathbf{I}=(\delta/\varepsilon)\mathbf{I}$ from the multiplicative bound and have $(1-\varepsilon)\mathbf{A}^T\mathbf{A}-\delta\mathbf{I}\preceq\tilde{\mathbf{A}}^T\tilde{\mathbf{A}}\preceq(1+\varepsilon)\mathbf{A}^T\mathbf{A}+\delta\mathbf{I}$. □

The idea of using overestimates of leverage scores to perform row sampling has been applied successfully to various problems (see, e. g., [17, 11]). However, in these applications, access to the entire matrix is required beforehand. In the streaming and online settings, we have to rely on partial data to approximate the true leverage scores. The most natural idea is to just use the portion of the matrix seen thus far as an approximation to $\mathbf{A}$. This leads us to introduce the *online λ-ridge leverage scores*:

$$l_i := \min(\mathbf{a}_i^T(\mathbf{A}_{i-1}^T\mathbf{A}_{i-1}+\lambda\mathbf{I})^{-1}\mathbf{a}_i\,,1),$$

where $\mathbf{A}_i$ $(i=0,\ldots,n)$ is defined as the matrix consisting of the first i rows of $\mathbf{A}$.[2]

Since clearly $\mathbf{A}_i^T\mathbf{A}_i\preceq\mathbf{A}^T\mathbf{A}$ for all i, it is not hard to see that l_i does overestimate the true λ-ridge leverage score for row $\mathbf{a}_i$. A more complex question, however, is establishing an upper bound on $\sum_{i=1}^n l_i$ so that we can bound the number of samples needed by Theorem 2.1.

A core result of this work, stated in Theorem 2.2, is establishing such an upper bound; in fact, this bound is shown to be tight up to constants (Theorem 5.1) and is nearly linear in most cases.

[2] We use the proposed scores l_i for simplicity, however note that the following, perhaps more natural, definition of online leverage scores would also be effective: $l_i' := \mathbf{a}_i^T(\mathbf{A}_i^T\mathbf{A}_i+\lambda\mathbf{I})^{-1}\mathbf{a}_i\,.$

Theorem 2.2. *Let* $\mathbf{A}$ *be an* $n \times d$ *matrix with rows* $\mathbf{a}_1, \ldots, \mathbf{a}_n$. *Let* $\mathbf{A}_i$ *for* $i \in \{0, \ldots, n\}$ *be the matrix consisting of the first* i *rows of* $\mathbf{A}$. *For* $\lambda > 0$, *let*

$$l_i := \min(\mathbf{a}_i^T (\mathbf{A}_{i-1}^T \mathbf{A}_{i-1} + \lambda \mathbf{I})^{-1} \mathbf{a}_i \, , 1)$$

be the online λ*-ridge leverage score of the* i^{th} *row of* $\mathbf{A}$. *Then*

$$\sum_{i=1}^{n} l_i = \mathcal{O}(d \log(\|\mathbf{A}\|_2^2 / \lambda)).$$

Theorems 2.1 and 2.2 suggest a simple algorithm for online row sampling: simply use the online λ-ridge leverage scores, for $\lambda := \delta/\varepsilon$. This gives a spectral approximation with $\mathcal{O}(d \log d \log(\varepsilon \|\mathbf{A}\|_2^2/\delta)/\varepsilon^2)$ rows. Unfortunately, computing each l_i exactly requires us to store *all* the rows we have seen in memory (or alternatively to store the sum of their outer products, $\mathbf{A}_i^T \mathbf{A}_i$). In many cases, such a requirement would defeat the purpose of streaming row sampling.

A natural idea is to use the sample we have kept thus far as an approximation to $\mathbf{A}_i$ when computing l_i. It turns out that the approximate online ridge leverage scores $\tilde{l}_i$ computed in this way will not always be good approximations to l_i; however, we can still prove that they satisfy the requisite bounds and yield the same row sample size! We formalize these results in the algorithm ONLINE-SAMPLE (Figure 1) and Theorem 2.3.

$\tilde{\mathbf{A}} =$ ONLINE-SAMPLE$(\mathbf{A}, \varepsilon, \delta)$, where $\mathbf{A}$ is an $n \times d$ matrix with rows $\mathbf{a}_1, \ldots, \mathbf{a}_n$, $\varepsilon \in (0,1)$, $\delta > 0$.

1. Set $\lambda := \delta/\varepsilon$, $c := 8 \log d/\varepsilon^2$.
2. Let $\tilde{\mathbf{A}}_0$ be a $0 \times d$ matrix.
3. For $i = 1, \ldots, n$:
   (a) Let $\tilde{l}_i := \min((1+\varepsilon)\mathbf{a}_i^T (\tilde{\mathbf{A}}_{i-1}^T \tilde{\mathbf{A}}_{i-1} + \lambda \mathbf{I})^{-1} \mathbf{a}_i \, , 1)$.
   (b) Let $p_i := \min(c\tilde{l}_i \, , 1)$.
   (c) Set $\tilde{\mathbf{A}}_i := \begin{cases} \begin{bmatrix} \tilde{\mathbf{A}}_{i-1} \\ \mathbf{a}_i/\sqrt{p_i} \end{bmatrix} & \text{with probability } p_i, \\ \tilde{\mathbf{A}}_{i-1} & \text{otherwise.} \end{cases}$
4. Return $\tilde{\mathbf{A}} := \tilde{\mathbf{A}}_n$.

Figure 1: The basic online sampling algorithm

Theorem 2.3. *Let* $\tilde{\mathbf{A}}$ *be the matrix returned by* ONLINE-SAMPLE$(\mathbf{A}, \varepsilon, \delta)$. *With high probability,*

$$(1-\varepsilon)\mathbf{A}^T\mathbf{A} - \delta \mathbf{I} \preceq \tilde{\mathbf{A}}^T \tilde{\mathbf{A}} \preceq (1+\varepsilon)\mathbf{A}^T\mathbf{A} + \delta \mathbf{I},$$

and the number of rows in $\tilde{\mathbf{A}}$ *is* $\mathcal{O}(d \log d \log(\varepsilon \|\mathbf{A}\|_2^2/\delta)/\varepsilon^2)$.

To save computation, we note that, with a small modification, we can run ONLINE-SAMPLE with batch processing of rows. Specifically, say we start from the i^{th} position in the stream. we can store the next $b = \mathcal{O}(d)$ rows. We can then compute sampling probabilities for these rows all at once using a system solver for $(\tilde{\mathbf{A}}_{i+b}^T \tilde{\mathbf{A}}_{i+b} + \lambda \mathbf{I})$. Using a trick introduced in [25], by applying a Johnson–Lindenstrauss random projection to the rows whose scores we are computing, we need just $\mathcal{O}(\log(1/\gamma))$ system solves to compute constant-factor approximations to the ridge scores with probability $1-\gamma$. If we set $\gamma = 1/\text{poly}(n)$ then we can union bound over our whole stream, using this trick with each batch of $\mathcal{O}(d)$ input rows. The batch probabilities will only be closer to the true ridge leverage scores than the non-batch probabilities and we will enjoy the same guarantees as ONLINE-SAMPLE.

Additionally, it turns out that with a simple trick, it is possible to reduce the memory usage of the algorithm by a factor of ε^{-2}, bringing it down to $\mathcal{O}(d \log d \log(\varepsilon \|\mathbf{A}\|_2^2/\delta))$ (assuming the row sample is output to an output stream). Note that this expression gets *smaller* with ε; hence we obtain a row sampling algorithm with memory complexity independent of desired multiplicative precision. The basic idea is that, instead of keeping all previously sampled rows in memory, we store a smaller set of rows that give a constant-factor spectral approximation, still enough to give good estimates of the online ridge leverage scores.

This result is presented in the algorithm SLIM-SAMPLE (Figure 2) and Lemma 3.5. A particularly interesting consequence for graphs with polynomially bounded edge weights is

Corollary 2.4. *Let G be a simple graph on d vertices, and $\varepsilon \in (0,1)$. We can construct a $(1+\varepsilon)$-sparsifier of G of size $\mathcal{O}(d\log^2 d/\varepsilon^2)$, using only $\mathcal{O}(d\log^2 d)$ working memory in the online model.*

Proof. Let $\sigma_{min}(\mathbf{A})$ and $\sigma_{max}(\mathbf{A}) = \|\mathbf{A}\|_2$ be the smallest and largest singular values of $\mathbf{A}$ respectively. Let $\kappa(\mathbf{A}) = \sigma_{max}(\mathbf{A})/\sigma_{min}(\mathbf{A})$ be the condition number of $\mathbf{A}$. If we apply Theorem 2.3 with $\delta = \varepsilon/\sigma_{min}^2(\mathbf{A})$. we require sample complexity $O(d\log d \log(\varepsilon\|\mathbf{A}\|_2^2/\delta)/\varepsilon^2) = O(d\log d\log(\kappa(\mathbf{A})^2)/\varepsilon^2)$. For an unweighted graph on d vertices, $\sigma_{max}(\mathbf{A})^2 \leq d$, since d is the largest squared singular value of the complete graph. Combining with Lemma 6.1 of [27], we have that the condition number of a graph on d vertices whose edge weights are within a multiplicative $\text{poly}(d)$ of each other is polynomial in d. So $\log(\kappa^2(\mathbf{A})) = \mathcal{O}(\log d)$, which gives the corollary. □

We remark that the algorithm of Corollary 2.4 can be made to run in nearly linear time in the stream size. We combine SLIM-SAMPLE with the batch processing idea described above. Because $\mathbf{A}$ is a graph, our matrix approximation is always a symmetric diagonally dominant matrix, with $\mathcal{O}(d)$ nonzero entries. We can solve systems in it in time $\tilde{\mathcal{O}}(d)$. Using the Johnson–Lindenstrauss random projection trick of [25], we can compute approximate ridge leverage scores for a batch of $\mathcal{O}(d)$ rows with failure probability polynomially small in n in $\tilde{\mathcal{O}}(d\log n)$ time. Union bounding over the whole stream, we obtain nearly linear running time.

To complement the row sampling results discussed above, we explore the limits of the proposed online setting. In Section 4 we present the algorithm ONLINE-BSS, which obtains spectral approximations with $\mathcal{O}(d\log(\varepsilon\|\mathbf{A}\|_2^2/\delta)/\varepsilon^2)$ rows in the online setting (with larger memory requirements than the simpler sampling algorithms). Its analysis is given in Theorem 4.1. In Section 5, we show that this number of samples is in fact the best achievable, up to constant factors (Theorem 5.1). The $\log(\varepsilon\|\mathbf{A}\|_2^2/\delta)$ factor is truly the cost of requiring rows to be selected in an online manner.

3 Analysis of sampling schemes

We begin by bounding the sum of online λ-ridge leverage scores. The intuition behind the proof of Theorem 2.2 is that whenever we add a row with a large online leverage score to a matrix, we increase its determinant significantly, as follows from the matrix determinant lemma (Lemma 3.1). Thus we can reduce upper bounding the online leverage scores to bounding the matrix determinant.

Lemma 3.1 (Matrix determinant lemma)**.** *Assume* $\mathbf{S}$ *is an invertible square matrix and* $\mathbf{u}$ *is a vector. Then*

$$\det(\mathbf{S}+\mathbf{u}\mathbf{u}^T) = (\det \mathbf{S})(1+\mathbf{u}^T\mathbf{S}^{-1}\mathbf{u}).$$

Proof of Theorem 2.2. By Lemma 3.1, we have

$$\begin{aligned}
\det(\mathbf{A}_{i+1}^T\mathbf{A}_{i+1}+\lambda\mathbf{I}) &= \det(\mathbf{A}_i^T\mathbf{A}_i+\lambda\mathbf{I})\cdot\left(1+\mathbf{a}_{i+1}^T(\mathbf{A}_i^T\mathbf{A}_i+\lambda\mathbf{I})^{-1}\mathbf{a}_{i+1}\right)\\
&\geq \det(\mathbf{A}_i^T\mathbf{A}_i+\lambda\mathbf{I})\cdot(1+l_{i+1})\\
&\geq \det(\mathbf{A}_i^T\mathbf{A}_i+\lambda\mathbf{I})\cdot \mathrm{e}^{l_{i+1}/2}.
\end{aligned}$$

Hence,

$$\begin{aligned}
\det(\mathbf{A}^T\mathbf{A}+\lambda\mathbf{I}) &= \det(\mathbf{A}_n^T\mathbf{A}_n+\lambda\mathbf{I})\\
&\geq \det(\lambda\mathbf{I})\cdot \mathrm{e}^{\sum l_i/2}\\
&= \lambda^d \mathrm{e}^{\sum l_i/2}.
\end{aligned}$$

We have $\det(\mathbf{A}^T\mathbf{A}+\lambda\mathbf{I}) \leq (\|\mathbf{A}\|_2^2+\lambda)^d$. Therefore

$$(\|\mathbf{A}\|_2^2+\lambda)^d \geq \lambda^d \mathrm{e}^{\sum l_i/2}.$$

Taking logarithms of both sides, we obtain

$$\begin{aligned}
d\log(\|\mathbf{A}\|_2^2+\lambda) &\geq d\log\lambda + \sum l_i/2\\
\sum l_i &\leq 2d\log(1+\|\mathbf{A}\|_2^2/\lambda).
\end{aligned}$$

□

We now turn to analyzing the algorithm ONLINE-SAMPLE. Because the samples taken by the algorithm are *not* independent, we are not able to use a standard matrix Chernoff bound like the one in Theorem 2.1. However, we do know that whether we take row i does not depend on later rows; thus, we are able to analyze the process as a martingale. We will use a matrix version of the Freedman inequality given by Tropp.

Theorem 3.2 (Matrix Freedman inequality [28])**.** *Let* $\mathbf{Y}_0,\mathbf{Y}_1,\ldots,\mathbf{Y}_n$ *be a matrix martingale whose values are self-adjoint matrices with dimension* d*, and let* $\mathbf{X}_1,\ldots,\mathbf{X}_n$ *be the difference sequence. Assume that the difference sequence is uniformly bounded in the sense that*

$$\|\mathbf{X}_k\|_2 \leq R \text{ almost surely, for } k=1,\ldots,n.$$

Define the predictable quadratic variation process of the martingale as

$$\mathbf{W}_k := \sum_{j=1}^{k} \mathbb{E}\left[\mathbf{X}_j^2 \mid \mathbf{Y}_{j-1}, \ldots, \mathbf{Y}_0\right], \textit{for } k = 1, \ldots, n.$$

Then, for all $\varepsilon > 0$ *and* $\sigma^2 > 0$,

$$P\left[\|\mathbf{Y}_n\|_2 \geq \varepsilon \textit{ and } \|\mathbf{W}_n\|_2 \leq \sigma^2\right] \leq d \cdot \exp\left(-\frac{\varepsilon^2/2}{\sigma^2 + R\varepsilon/3}\right).$$

We begin by showing that the output of ONLINE-SAMPLE is in fact an approximation of $\mathbf{A}$, and that the approximate online leverage scores are lower bounded by the actual online leverage scores.

Lemma 3.3. *After running* ONLINE-SAMPLE*, it holds with high probability that*

$$(1-\varepsilon)\mathbf{A}^T\mathbf{A} - \delta\mathbf{I} \preceq \tilde{\mathbf{A}}^T\tilde{\mathbf{A}} \preceq (1+\varepsilon)\mathbf{A}^T\mathbf{A} + \delta\mathbf{I},$$

and also

$$\tilde{l}_i \geq \mathbf{a}_i^T(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{a}_i$$

for $i = 1, \ldots, n$.

Proof. Let

$$\mathbf{u}_i := (\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1/2}\mathbf{a}_i\,.$$

We construct a matrix martingale $\mathbf{Y}_0, \mathbf{Y}_1, \ldots, \mathbf{Y}_n \in \mathbb{R}^{d\times d}$ with the difference sequence $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Set $\mathbf{Y}_0 = \mathbf{0}$. If $\|\mathbf{Y}_{i-1}\|_2 \geq \varepsilon$, we set $\mathbf{X}_i := \mathbf{0}$. Otherwise, let

$$\mathbf{X}_i := \begin{cases} (1/p_i - 1)\mathbf{u}_i\mathbf{u}_i^T & \text{if } \mathbf{a}_i \text{ is sampled in } \tilde{\mathbf{A}}, \\ -\mathbf{u}_i\mathbf{u}_i^T & \text{otherwise.} \end{cases}$$

In the case that $\|\mathbf{Y}_{i-1}\|_2 < \varepsilon$, by construction, $\|\mathbf{Y}_j\|_2 < \varepsilon$ for all $j < i-1$. So we have

$$\mathbf{Y}_{i-1} = (\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1/2}(\tilde{\mathbf{A}}_{i-1}^T\tilde{\mathbf{A}}_{i-1} - \mathbf{A}_{i-1}^T\mathbf{A}_{i-1})(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1/2}.$$

Since $\|\mathbf{Y}_{i-1}\|_2 \leq \varepsilon$, we can see that

$$(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1/2}(\tilde{\mathbf{A}}_{i-1}^T\tilde{\mathbf{A}}_{i-1})(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1/2} \preceq (\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1/2}(\mathbf{A}_{i-1}^T\mathbf{A}_{i-1})(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1/2} + \varepsilon\mathbf{I}.$$

Multiplying on both right and left by $(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{1/2}$ gives

$$\tilde{\mathbf{A}}_{i-1}^T\tilde{\mathbf{A}}_{i-1} \preceq \mathbf{A}_{i-1}^T\mathbf{A}_{i-1} + \varepsilon(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I}).$$

Hence, we have

$$\begin{aligned} \tilde{l}_i &= \min((1+\varepsilon)\mathbf{a}_i^T(\tilde{\mathbf{A}}_{i-1}^T\tilde{\mathbf{A}}_{i-1} + \lambda\mathbf{I})^{-1}\mathbf{a}_i, 1) \\ &\geq \min((1+\varepsilon)\mathbf{a}_i^T(\mathbf{A}_{i-1}^T\mathbf{A}_{i-1} + \lambda\mathbf{I} + \varepsilon(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I}))^{-1}\mathbf{a}_i, 1) \\ &\geq \min((1+\varepsilon)\mathbf{a}_i^T((1+\varepsilon)(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I}))^{-1}\mathbf{a}_i, 1) \\ &= \mathbf{a}_i^T(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{a}_i \qquad (3.1) \\ &= \mathbf{u}_i^T\mathbf{u}_i\,. \end{aligned}$$

Thus, $p_i = \min(c\tilde{l}_i, 1) \geq \min(c\mathbf{u}_i^T\mathbf{u}_i, 1)$. If $p_i = 1$, then $\mathbf{X}_i = \mathbf{0}$. Otherwise, we have $p_i \geq c\mathbf{u}_i^T\mathbf{u}_i$ and

$$\|\mathbf{X}_i\|_2 \leq \max\{1, 1/p_i - 1\} \cdot \|\mathbf{u}_i\mathbf{u}_i^T\|_2 \leq \mathbf{u}_i^T\mathbf{u}_i/p_i \leq 1/c. \tag{3.2}$$

Further

$$\begin{aligned}
\mathbb{E}\left[\mathbf{X}_i^2 \mid \mathbf{Y}_{i-1}, \ldots, \mathbf{Y}_0\right] &\preceq p_i \cdot (1/p_i - 1)^2(\mathbf{u}_i\mathbf{u}_i^T)^2 + (1-p_i)\cdot(\mathbf{u}_i\mathbf{u}_i^T)^2 \\
&= (\mathbf{u}_i\mathbf{u}_i^T)^2 \cdot (1-p_i)/p_i \\
&\preceq \mathbf{u}_i\mathbf{u}_i^T \cdot \left(\mathbf{u}_i^T\mathbf{u}_i/p_i\right) \\
&\preceq \mathbf{u}_i\mathbf{u}_i^T/c. && \text{(by equation (3.2))}
\end{aligned}$$

And so, for the predictable quadratic variation process of the martingale $\{\mathbf{Y}_i\}$

$$\mathbf{W}_i := \sum_{k=1}^{i} \mathbb{E}\left[\mathbf{X}_k^2 \mid \mathbf{Y}_{k-1}, \ldots, \mathbf{Y}_0\right],$$

we have

$$\|\mathbf{W}_i\|_2 \leq \left\|\sum_{k=1}^{i} \mathbf{u}_k\mathbf{u}_k^T/c\right\|_2 \leq 1/c.$$

Therefore by, Theorem 3.2, we have

$$\begin{aligned}
P\left[\|\mathbf{Y}_n\|_2 \geq \varepsilon\right] &\leq d \cdot \exp\left(\frac{-\varepsilon^2/2}{1/c + \varepsilon/(3c)}\right) \\
&\leq d \cdot \exp(-c\varepsilon^2/4) \\
&= 1/d.
\end{aligned}$$

This implies that with high probability

$$\|(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1/2}(\tilde{\mathbf{A}}^T\tilde{\mathbf{A}} + \lambda\mathbf{I})(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1/2} - \mathbf{I}\|_2 \leq \varepsilon$$

and so

$$(1-\varepsilon)(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I}) \preceq \tilde{\mathbf{A}}^T\tilde{\mathbf{A}} + \lambda\mathbf{I} \preceq (1+\varepsilon)(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I}).$$

Subtracting $\lambda\mathbf{I} = (\delta/\varepsilon)\mathbf{I}$ from all sides, we get

$$(1-\varepsilon)\mathbf{A}^T\mathbf{A} - \delta\mathbf{I} \preceq \tilde{\mathbf{A}}^T\tilde{\mathbf{A}} \preceq (1+\varepsilon)\mathbf{A}^T\mathbf{A} + \delta\mathbf{I}.$$

Finally, note that, since we set $\mathbf{X}_i = \mathbf{0}$ if $\|\mathbf{Y}_{i-1}\|_2 \geq \varepsilon$, $\|\mathbf{Y}_n\|_2 < \varepsilon$ implies $\|\mathbf{Y}_i\|_2 < \varepsilon$ for all $i < n$. We thus have the desired bound on $\tilde{l}_i$ by equation (3.1). □

If we set c in ONLINE-SAMPLE to be proportional to $\log n$ rather than $\log d$, we would be able to take a union bound over all the rows and guarantee that with high probability all the approximate online leverage scores $\tilde{l}_i$ are close to true online leverage scores l_i. Thus Theorem 2.2 would imply that ONLINE-SAMPLE only selects $\mathcal{O}(d \log n \log(\|\mathbf{A}\|_2^2/\lambda)/\varepsilon^2)$ rows with high probability.

In order to remove the dependency on n, we have to sacrifice achieving close approximations to l_i at every step. Instead, we show that the *sum* of the computed approximate online leverage scores is still small with high probability, using a custom Chernoff bound.

Lemma 3.4. *After running* ONLINE-SAMPLE*, it holds with high probability that*

$$\sum_{i=1}^{n} \tilde{l}_i = \mathcal{O}(d \log(\|\mathbf{A}\|_2^2/\lambda)).$$

Proof. Define

$$\delta_i := \log\det(\tilde{\mathbf{A}}_i^T \tilde{\mathbf{A}}_i + \lambda \mathbf{I}) - \log\det(\tilde{\mathbf{A}}_{i-1}^T \tilde{\mathbf{A}}_{i-1} + \lambda \mathbf{I}).$$

The proof closely follows the idea from the proof of Theorem 2.2. We will aim to show that large values of $\tilde{l}_i$ correlate with large values of δ_i. Then, the sum of δ_i can be bounded by the logarithm of the ratio of the determinants of $\tilde{\mathbf{A}}^T\tilde{\mathbf{A}} + \lambda\mathbf{I}$ and $\lambda\mathbf{I}$, giving us a bound on the sum of $\tilde{l}_i$. First, we will show that $\mathbb{E}\left[\exp(\tilde{l}_i/8 - \delta_i) \mid \tilde{\mathbf{A}}_{i-1}, \ldots, \tilde{\mathbf{A}}_0\right]$ is always at most 1. Note that if row i is sampled by ONLINE-SAMPLE, $\tilde{\mathbf{A}}_i^T\tilde{\mathbf{A}}_i + \lambda\mathbf{I} = \tilde{\mathbf{A}}_{i-1}^T\tilde{\mathbf{A}}_{i-1} + \lambda\mathbf{I} + \mathbf{a}_i\mathbf{a}_i^T/p_i$ By the matrix determinant lemma (Lemma 3.1), we thus have $\mathrm{e}^{\delta_i} = 1 + \mathbf{a}_i^T(\tilde{\mathbf{A}}_{i-1}^T\tilde{\mathbf{A}}_{i-1} + \lambda\mathbf{I})^{-1}\mathbf{a}_i/p$ in this case. Otherwise, if row i is not sampled, $\delta_i = 0$. Thus,

$$\begin{aligned}
\mathbb{E}\left[\exp(\tilde{l}_i/8 - \delta_i) \mid \tilde{\mathbf{A}}_{i-1}, \ldots, \tilde{\mathbf{A}}_0\right] &= p_i \cdot \mathrm{e}^{\tilde{l}_i/8}(1 + \mathbf{a}_i^T(\tilde{\mathbf{A}}_{i-1}^T\tilde{\mathbf{A}}_{i-1} + \lambda\mathbf{I})^{-1}\mathbf{a}_i/p_i)^{-1} + (1-p_i)\mathrm{e}^{\tilde{l}_i/8} \\
&\leq p_i \cdot (1 + \tilde{l}_i/4)(1 + \mathbf{a}_i^T(\tilde{\mathbf{A}}_{i-1}^T\tilde{\mathbf{A}}_{i-1} + \lambda\mathbf{I})^{-1}\mathbf{a}_i/p_i)^{-1} + (1-p_i)(1+\tilde{l}_i/4). \quad (3.3)
\end{aligned}$$

If $c\tilde{l}_i < 1$, we have $p_i = c\tilde{l}_i$ and $\tilde{l}_i = (1+\varepsilon)\mathbf{a}_i^T(\tilde{\mathbf{A}}_{i-1}^T\tilde{\mathbf{A}}_{i-1} + \lambda\mathbf{I})^{-1}\mathbf{a}_i$, and so,

$$\begin{aligned}
\mathbb{E}\left[\exp(\tilde{l}_i/8 - \delta_i) \mid \tilde{\mathbf{A}}_{i-1}, \ldots, \tilde{\mathbf{A}}_0\right] &\leq c\tilde{l}_i \cdot (1 + \tilde{l}_i/4)(1 + 1/((1+\varepsilon)c))^{-1} + (1 - c\tilde{l}_i)(1 + \tilde{l}_i/4) \\
&= (1 + \tilde{l}_i/4)(c\tilde{l}_i(1 + 1/((1+\varepsilon)c))^{-1} + 1 - c\tilde{l}_i) \\
&= (1 + \tilde{l}_i/4)(1 - \tilde{l}_i/4) \\
&\leq 1.
\end{aligned}$$

Otherwise, we have $p_i = 1$ and so, by (3.3),

$$\begin{aligned}
\mathbb{E}\left[\exp(\tilde{l}_i/8 - \delta_i) \mid \tilde{\mathbf{A}}_{i-1}, \ldots, \tilde{\mathbf{A}}_0\right] &\leq (1 + \tilde{l}_i/4)(1 + \mathbf{a}_i^T(\tilde{\mathbf{A}}_{i-1}^T\tilde{\mathbf{A}}_{i-1} + \lambda\mathbf{I})^{-1}\mathbf{a}_i)^{-1} \\
&\leq (1 + \tilde{l}_i/4)(1 + \tilde{l}_i)^{-1} \\
&\leq 1.
\end{aligned}$$

We will now analyze the expected product of $\exp(\tilde{l}_i/8 - \delta_i)$ over the first k steps, $\mathbb{E}\left[\exp\left(\sum_{i=1}^k \tilde{l}_i/8 - \delta_i\right)\right]$. Since conditioned on the first k steps, $\exp(\tilde{l}_k/8 - \delta_k)$ is independent of $\exp(\tilde{l}_i/8 - \delta_i)$ for all $i < k$, for $k \geq 1$ we have

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\sum_{i=1}^{k} \tilde{l}_i/8 - \delta_i\right)\right] &= \mathbb{E}_{\text{first } k-1 \text{ steps}}\left[\exp\left(\sum_{i=1}^{k-1} \tilde{l}_i/8 - \delta_i\right) \mathbb{E}\left[\exp(\tilde{l}_k/8 - \delta_k) \mid \tilde{\mathbf{A}}_{k-1}, \ldots, \tilde{\mathbf{A}}_0\right]\right] \\
&\leq \mathbb{E}\left[\exp\left(\sum_{i=1}^{k-1} \tilde{l}_i/8 - \delta_i\right)\right],
\end{aligned}$$

and so by induction on k

$$\mathbb{E}\left[\exp\left(\sum_{i=1}^{n} \tilde{l}_i/8 - \delta_i\right)\right] \leq 1.$$

Hence by Markov's inequality

$$P\left[\sum_{i=1}^{n} \tilde{l}_i > 8d + 8\sum_{i=1}^{n} \delta_i\right] \leq \mathrm{e}^{-d}.$$

By Lemma 3.3, with high probability we have $\tilde{\mathbf{A}}^T\tilde{\mathbf{A}} + \lambda\mathbf{I} \preceq (1+\varepsilon)(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})$. We also have with high probability

$$\det(\tilde{\mathbf{A}}^T\tilde{\mathbf{A}} + \lambda\mathbf{I}) \leq (1+\varepsilon)^d(\|\mathbf{A}\|_2^2 + \lambda)^d,$$
$$\log\det(\tilde{\mathbf{A}}^T\tilde{\mathbf{A}} + \lambda\mathbf{I}) \leq d(1+\log(\|\mathbf{A}\|_2^2 + \lambda)).$$

Hence, with high probability it holds that

$$\begin{aligned}\sum_{i=1}^{n} \delta_i &= \log\det(\tilde{\mathbf{A}}^T\tilde{\mathbf{A}} + \lambda\mathbf{I}) - d\log(\lambda) \\ &\leq d(1+\log(\|\mathbf{A}\|_2^2 + \lambda) - \log(\lambda)) \\ &= d(1+\log(1+\|\mathbf{A}\|_2^2/\lambda)).\end{aligned}$$

And so, with high probability,

$$\begin{aligned}\sum_{i=1}^{n} \tilde{l}_i &\leq 8d + 8\sum_{i=1}^{n} \delta_i \\ &\leq 16d + 8d\log(1+\|\mathbf{A}\|_2^2/\lambda) \\ &= \mathcal{O}(d\log(\|\mathbf{A}\|_2^2/\lambda)).\end{aligned}$$

□

Proof of Theorem 2.3. The statement follows immediately from Lemmas 3.3 and 3.4. □

Observe that by Theorem 2.3, ONLINE-SAMPLE stores $O(d\log d\log(\varepsilon\|\mathbf{A}\|_2^2/\delta)/\varepsilon^2)$ rows in memory. We now consider a simple modification of the algorithm, SLIM-SAMPLE (Figure 2), that removes the $1/\varepsilon^2$ factor from the working memory usage with no additional cost.

Lemma 3.5. *Let $\tilde{\mathbf{A}}$ be the matrix returned by* SLIM-SAMPLE$(\mathbf{A}, \varepsilon, \delta)$. *Then, with high probability,*

$$(1-\varepsilon)\mathbf{A}^T\mathbf{A} - \delta\mathbf{I} \preceq \tilde{\mathbf{A}}^T\tilde{\mathbf{A}} \preceq (1+\varepsilon)\mathbf{A}^T\mathbf{A} + \delta\mathbf{I},$$

and the number of rows in $\tilde{\mathbf{A}}$ is $\mathcal{O}(d\log d\log(\varepsilon\|\mathbf{A}\|_2^2/\delta)/\varepsilon^2)$.

Moreover, with high probability, the memory requirement of SLIM-SAMPLE *is dominated by storing $\mathcal{O}(d\log d\log(\varepsilon\|\mathbf{A}\|_2^2/\delta))$ rows of $\mathbf{A}$.*

Proof. As the samples are independent, the statement follows from Theorem 2.1 and Lemmas 3.3 and 3.4. □

$\tilde{\mathbf{A}} = \text{SLIM-SAMPLE}(\mathbf{A}, \varepsilon, \delta)$, where $\mathbf{A}$ is an $n \times d$ matrix with rows $\mathbf{a}_1, \ldots, \mathbf{a}_n$, $\varepsilon \in (0,1)$, $\delta > 0$.

1. Set $\lambda := \delta/\varepsilon$, $c := 8 \log d/\varepsilon^2$.
2. Let $\tilde{\mathbf{A}}_0$ be a $0 \times d$ matrix.
3. Let $\tilde{l}_1, \ldots, \tilde{l}_n$ be the approximate online leverage scores computed by an independent instance of $\text{ONLINE-SAMPLE}(\mathbf{A}, 1/2, \delta/(2\varepsilon))$.
4. For $i = 1, \ldots, n$:
   (a) Let $p_i := \min(c\tilde{l}_i, 1)$.
   (b) Set $\tilde{\mathbf{A}}_i := \begin{cases} \begin{bmatrix} \tilde{\mathbf{A}}_{i-1} \\ \mathbf{a}_i/\sqrt{p_i} \end{bmatrix} & \text{with probability } p_i, \\ \tilde{\mathbf{A}}_{i-1} & \text{otherwise.} \end{cases}$
5. Return $\tilde{\mathbf{A}} := \tilde{\mathbf{A}}_n$.

Figure 2: The low-memory online sampling algorithm

4 Asymptotically optimal algorithm

In addition to sampling by online leverage scores, we introduce a row sampling algorithm, ONLINE-BSS (Figure 3), which improves the row count of ONLINE-SAMPLE by a $\log d$ factor, to

$$\mathcal{O}(d \log(\varepsilon \|\mathbf{A}\|_2^2/\delta)/\varepsilon^2).$$

This improved bound matches the lower bound for online sampling given in Theorem 5.1. This approach uses a variant of the deterministic "BSS" method, introduced by Batson, Spielman, and Srivastava in [2]. It is well known that this method yields spectral approximations with a $\log d$ factor fewer rows than leverage scores sampling in the offline setting, and we show that this improvement extends to online approximation.

Unlike the original BSS algorithm of [2], our algorithm is randomized. It is similar to, and inspired by, the randomized version of BSS from [19], especially "Algorithm 1" from that paper. In both algorithms, like in online leverage score sampling, when a new row is processed, a probability p_i is assigned to it, and it is kept with probability p_i and rejected otherwise. The key difference between the algorithms is in the definition of p_i. Like ONLINE-SAMPLE, at each step, ONLINE-BSS maintains a row sample $\tilde{\mathbf{A}}_i$ which approximates the matrix $\mathbf{A}_i$ that has been seen so far. However, p_i cannot be computed solely based on $\tilde{\mathbf{A}}_{i-1}$—it is necessary to "remember" the entire input. Thus, ONLINE-BSS is *not memory efficient*, using $O(d^2)$ space. One may improve the memory dependence by simply running ONLINE-BSS on the output stream of rows produced by ONLINE-SAMPLE. This reduces the storage cost to the size of that output spectral approximation. Of course, this does not mean that ONLINE-BSS leads to a space savings over ONLINE-SAMPLE. However the number of rows in its output stream will be less than that of ONLINE-SAMPLE, by a $\log d$ factor.

We also remark that ONLINE-SAMPLE gives bounds on both the size of the output spectral approximation and its accuracy with high probability. In contrast, ONLINE-BSS gives an *expected* bound on the output size, while it *never* fails to output a correct spectral approximation. These guarantees are similar to those given in [19]. Below, we give present the performance guarantees of ONLINE-BSS and its analysis.

Theorem 4.1. *Let* $\tilde{\mathbf{A}}$ *be the matrix output by* ONLINE-BSS$(\mathbf{A}, \varepsilon, \delta)$ *(Figure 3).*

1. *We always have* $(1-\varepsilon)\mathbf{A}^T\mathbf{A} - \delta\mathbf{I} \prec \tilde{\mathbf{A}}^T\tilde{\mathbf{A}} \prec (1+\varepsilon)\mathbf{A}^T\mathbf{A} + \delta\mathbf{I}$.
2. *The expected number of rows in* $\tilde{\mathbf{A}}$ *is* $\mathcal{O}(d\log(\varepsilon\|\mathbf{A}\|_2^2/\delta)/\varepsilon^2)$.

$\tilde{\mathbf{A}} =$ ONLINE-BSS$(\mathbf{A}, \varepsilon, \delta)$, where $\mathbf{A}$ is an $n \times d$ matrix with rows $\mathbf{a}_1, \ldots, \mathbf{a}_n$, $\varepsilon \in (0,1)$, $\delta > 0$.

1. Set $c_U = \frac{2}{\varepsilon} + 1$ and $c_L = \frac{2}{\varepsilon} - 1$.
2. Let $\tilde{\mathbf{A}}_0$ be a $0 \times d$ matrix, $\mathbf{B}_0^U = \delta\mathbf{I}$, $\mathbf{B}_0^L = -\delta\mathbf{I}$.
3. For $i = 1, \ldots, n$:
   (a) Let $\mathbf{X}_{i-1}^U = (\mathbf{B}_{i-1}^U - \tilde{\mathbf{A}}_{i-1}^T\tilde{\mathbf{A}}_{i-1})$, $\mathbf{X}_{i-1}^L = (\tilde{\mathbf{A}}_{i-1}^T\tilde{\mathbf{A}}_{i-1} - \mathbf{B}_{i-1}^L)$.
   (b) Let $p_i := \min(c_U\mathbf{a}_i^T(\mathbf{X}_{i-1}^U)^{-1}\mathbf{a}_i + c_L\mathbf{a}_i^T(\mathbf{X}_{i-1}^L)^{-1}\mathbf{a}_i, 1)$.
   (c) Set $\tilde{\mathbf{A}}_i := \begin{cases} \begin{bmatrix} \tilde{\mathbf{A}}_{i-1} \\ \mathbf{a}_i/\sqrt{p_i} \end{bmatrix} & \text{with probability } p_i, \\ \tilde{\mathbf{A}}_{i-1} & \text{otherwise.} \end{cases}$
   (d) Set $\mathbf{B}_i^U = \mathbf{B}_{i-1}^U + (1+\varepsilon)\mathbf{a}_i\mathbf{a}_i^T$, $\mathbf{B}_i^L = \mathbf{B}_{i-1}^L + (1-\varepsilon)\mathbf{a}_i\mathbf{a}_i^T$.
4. Return $\tilde{\mathbf{A}} := \tilde{\mathbf{A}}_n$.

Figure 3: The Online BSS Algorithm

Proof of Theorem 4.1 Part 1. As in [2], a key of idea of ONLINE-BSS is to maintain two matrices, $\mathbf{B}_i^U$ and $\mathbf{B}_i^L$, acting as upper and lower "barriers." We will prove that the current approximation $\tilde{\mathbf{A}}_i$ always falls between them:

$$\mathbf{B}_i^L \prec \tilde{\mathbf{A}}_i^T\tilde{\mathbf{A}}_i \prec \mathbf{B}_i^U. \tag{4.1}$$

Equivalently, $\mathbf{X}_i^U$ and $\mathbf{X}_i^L$ will always remain positive definite. Since, at the completion of the algorithm, $\mathbf{B}_n^U = (1+\varepsilon)\mathbf{A}^T\mathbf{A} + \delta\mathbf{I}$ and $\mathbf{B}_n^L = (1-\varepsilon)\mathbf{A}^T\mathbf{A} - \delta\mathbf{I}$ this ensures that the final approximation *always* satisfies the approximation bound in claim (1) of the theorem. p_i is chosen at step 3(b) to ensure this invariant—if either $\mathbf{X}_i^U$ or $\mathbf{X}_i^L$ are too small (we are too close to one of the barriers) then at least one of $\mathbf{a}_i^T(\mathbf{X}_{i-1}^U)^{-1}\mathbf{a}_i$ or $\mathbf{a}_i^T(\mathbf{X}_{i-1}^L)^{-1}\mathbf{a}_i$ will be large and so p_i will be large.

We can prove this invariant if (4.1) holds, by induction on i. The base case follows from the initialization of $\tilde{\mathbf{A}}_0$ with $\tilde{\mathbf{A}}_0^T\tilde{\mathbf{A}}_0 = \mathbf{0}$, $\mathbf{B}_0^U = \delta\mathbf{I}$, and $\mathbf{B}_0^L = -\delta\mathbf{I}$ since clearly $-\delta\mathbf{I} \prec \mathbf{0} \prec \delta\mathbf{I}$. For each successive step, we consider two possibilities.

Case 1: $p_i = 1$. When $p_i = 1$, $\tilde{\mathbf{A}}_i^T\tilde{\mathbf{A}}_i = \tilde{\mathbf{A}}_{i-1}^T\tilde{\mathbf{A}}_{i-1} + \mathbf{a}_i\mathbf{a}_i^T$. Since we set $\mathbf{B}_i^U = \mathbf{B}_{i-1}^U + (1+\varepsilon)\mathbf{a}_i\mathbf{a}_i^T$ and $\mathbf{B}_i^L = \mathbf{B}_{i-1}^L + (1-\varepsilon)\mathbf{a}_i\mathbf{a}_i^T$, we can see that $\mathbf{X}_i^U = \mathbf{X}_{i-1}^U + \varepsilon\mathbf{a}_i\mathbf{a}_i^T$ and $\mathbf{X}_i^L = \mathbf{X}_{i-1}^L + \varepsilon\mathbf{a}_i\mathbf{a}_i^T$. Since by the induction assumption, $\mathbf{X}_{i-1}^U$ and $\mathbf{X}_{i-1}^L$ are both positive definite, so are $\mathbf{X}_i^U$ and $\mathbf{X}_i^L$, giving the claim.

Case 2: $p_i < 1$. In this case, with probability p_i, $\tilde{\mathbf{A}}_i^T\tilde{\mathbf{A}}_i = \tilde{\mathbf{A}}_{i-1}^T\tilde{\mathbf{A}}_{i-1} + \mathbf{a}_i\mathbf{a}_i^T/p$. With probability $1-p_i$, $\tilde{\mathbf{A}}_i^T\tilde{\mathbf{A}}_i = \tilde{\mathbf{A}}_{i-1}^T\tilde{\mathbf{A}}_{i-1}$. Thus, in any case, $\tilde{\mathbf{A}}_i^T\tilde{\mathbf{A}}_i \prec \tilde{\mathbf{A}}_{i-1}^T\tilde{\mathbf{A}}_{i-1} + \mathbf{a}_i\mathbf{a}_i^T/p$. In turn, $\mathbf{X}_i^U \succeq \mathbf{X}_{i-1}^U - \mathbf{a}_i\mathbf{a}_i^T/p_i$. Observe that $c_U = 2/\varepsilon + 1 > 1$ and thus $p_i > \mathbf{a}_i^T(\mathbf{X}_{i-1}^U)^{-1}\mathbf{a}_i$. This gives $\mathbf{X}_{i-1}^U \succ \mathbf{a}_i\mathbf{a}_i^T/p_i$ and so since $\mathbf{X}_i^U \succeq \mathbf{X}_{i-1}^U - \mathbf{a}_i\mathbf{a}_i^T/p_i$, it must be postive definite.

Analogously, since $\mathbf{B}_i^L = \mathbf{B}_{i-1}^L + (1-\varepsilon)\mathbf{a}_i\mathbf{a}_i^T$, we have $\mathbf{X}_i^L \succeq \mathbf{X}_{i-1}^L - (1-\varepsilon)\mathbf{a}_i\mathbf{a}_i^T \succeq \mathbf{X}_{i-1}^L - \mathbf{a}_i\mathbf{a}_i^T$. Since $c_L = 2/\varepsilon - 1 > 1$ for $\varepsilon \in (0,1)$, and since $p_i < 1$ (by the fact that we are in Case 2), we have $\mathbf{a}_i^T(\mathbf{X}_{i-1}^L)^{-1}\mathbf{a}_i < 1$. This in turn gives $\mathbf{X}_{i-1}^L \succ \mathbf{a}_i\mathbf{a}_i^T$ and thus since $\mathbf{X}_i^L \succeq \mathbf{X}_{i-1}^L - \mathbf{a}_i\mathbf{a}_i^T$, it must be positive definite, giving the claim in this case.

Thus, we have shown (4.1) for all i. In particular, $\mathbf{B}_n^U \prec \tilde{\mathbf{A}}^T\tilde{\mathbf{A}} \prec \mathbf{B}_n^L$. We can see by construction that

$$\mathbf{B}_n^U = (1+\varepsilon)\mathbf{A}^T\mathbf{A} + \delta\mathbf{I} \text{ and } \mathbf{B}_n^L = (1-\varepsilon)\mathbf{A}^T\mathbf{A} - \delta\mathbf{I}.$$

Thus, we have $(1-\varepsilon)\mathbf{A}^T\mathbf{A} - \delta\mathbf{I} \prec \tilde{\mathbf{A}}^T\tilde{\mathbf{A}} \prec (1+\varepsilon)\mathbf{A}^T\mathbf{A} + \delta\mathbf{I}$, which gives the first claim of the theorem. □

In our proof of the second claim, bounding the expected number of rows sampled, we will need the following technical lemma, which is derived from the Sherman–Morrison formula [24].

Lemma 4.2. *Given a positive definite matrix $\mathbf{X}$, two vectors $\mathbf{u}$ and $\mathbf{v}$, two scalar multipliers a and b, and a probability p, define the random variable $\widehat{\mathbf{X}}$ to be $\mathbf{X} - a\mathbf{u}\mathbf{u}^T$ with probability p and $\mathbf{X} - b\mathbf{u}\mathbf{u}^T$ otherwise. Then if $\mathbf{u}^T\mathbf{X}^{-1}\mathbf{u} = 1$,*

$$\mathbb{E}\left[\mathbf{v}^T\widehat{\mathbf{X}}^{-1}\mathbf{v} - \mathbf{v}^T\mathbf{X}^{-1}\mathbf{v}\right] = (\mathbf{v}^T\mathbf{X}^{-1}\mathbf{u})^2 \cdot \frac{pa + (1-p)b - ab}{(1-a)(1-b)}.$$

Proof. We apply the Sherman–Morrison formula to each of the two possibilities ($\widehat{\mathbf{X}} = \mathbf{X}\mathbf{X} - a\mathbf{u}\mathbf{u}^T$ and $\widehat{\mathbf{X}} = \mathbf{X} + b\mathbf{u}\mathbf{u}^T$ respectively). These give respective $\widehat{\mathbf{X}}^{-1}$ values of

$$\widehat{\mathbf{X}}^{-1} = \mathbf{X}^{-1} + a \cdot \frac{\mathbf{X}^{-1}\mathbf{u}\mathbf{u}^T\mathbf{X}^{-1}}{1 - a\mathbf{u}^T\mathbf{X}^{-1}\mathbf{u}} = \mathbf{X}^{-1} + \frac{a}{1-a} \cdot \mathbf{X}^{-1}\mathbf{u}\mathbf{u}^T\mathbf{X}^{-1}$$

and

$$\widehat{\mathbf{X}}^{-1} = \mathbf{X}^{-1} + b \cdot \frac{\mathbf{X}^{-1}\mathbf{u}\mathbf{u}^T\mathbf{X}^{-1}}{1 - b\mathbf{u}^T\mathbf{X}^{-1}\mathbf{u}} = \mathbf{X}^{-1} + \frac{b}{1-b} \cdot \mathbf{X}^{-1}\mathbf{u}\mathbf{u}^T\mathbf{X}^{-1}.$$

The values of $\mathbf{v}^T\widehat{\mathbf{X}}^{-1}\mathbf{v} - \mathbf{v}^T\mathbf{X}^{-1}\mathbf{v}$ are then respectively

$$\frac{a}{1-a} \cdot \mathbf{v}^T\mathbf{X}^{-1}\mathbf{u}\mathbf{u}^T\mathbf{X}^{-1}\mathbf{v} = (\mathbf{v}^T\mathbf{X}^{-1}\mathbf{u})^2 \cdot \frac{a}{1-a}$$

and

$$\frac{b}{1-b} \cdot \mathbf{v}^T \mathbf{X}^{-1} \mathbf{u} \mathbf{u}^T \mathbf{X}^{-1} \mathbf{v} = (\mathbf{v}^T \mathbf{X}^{-1} \mathbf{u})^2 \cdot \frac{b}{1-b}.$$

Combining these gives the stated result. □

Proof of Theorem 4.1 Part 2. We will show that the probability that row $\mathbf{a}_i$ is included in $\tilde{\mathbf{A}}$ is at most $8/\varepsilon^2 \cdot l_i$, where l_i is the online $2\delta/\varepsilon$-ridge leverage score of $\mathbf{a}_i$, i. e.,

$$l_i = \min(\mathbf{a}_i^T \left(\mathbf{A}_{i-1}^T \mathbf{A}_{i-1} + 2\delta/\varepsilon \cdot \mathbf{I}\right)^{-1} \mathbf{a}_i \, , 1).$$

Since $\sum_{i=1}^n l_i = \mathcal{O}(d \log(\varepsilon \|\mathbf{A}\|_2^2/\delta))$ by Theorem 2.2, this implies that $\tilde{\mathbf{A}}$ has $\mathcal{O}(d \log(\varepsilon \|\mathbf{A}\|_2^2/\delta)/\varepsilon^2)$ rows in expectation, completing the second claim of the theorem.

First, we introduce some notation to help in the analysis. Let q_i be the probability that row $\mathbf{a}_i$ is sampled in the algorithm. Note that q_i is fixed and we seek to prove that $q_i \leq 8/\varepsilon^2 \cdot l_i$. The probability p_i that $\mathbf{a}_i$ is sampled at step i is a random variable. We have $q_i = \mathbb{E}\left[p_i\right]$. Thus it suffices to prove $\mathbb{E}\left[p_i\right] \leq 8/\varepsilon^2 \cdot l_i$. We define

$$\begin{aligned}
\mathbf{C}_{i,j}^U &= \delta \mathbf{I} + \frac{\varepsilon}{2} \mathbf{A}_i^T \mathbf{A}_i + \left(1 + \frac{\varepsilon}{2}\right) \mathbf{A}_j^T \mathbf{A}_j \\
\mathbf{C}_{i,j}^L &= -\delta \mathbf{I} - \frac{\varepsilon}{2} \mathbf{A}_i^T \mathbf{A}_i + \left(1 - \frac{\varepsilon}{2}\right) \mathbf{A}_j^T \mathbf{A}_j \, .
\end{aligned}$$

Note that $\mathbf{C}_{i,i}^U = \mathbf{B}_i^U$, $\mathbf{C}_{i,i}^L = \mathbf{B}_i^L$, and for $j \leq i$, $\mathbf{C}_{i,j}^U \succeq \mathbf{B}_j^U$ and $\mathbf{C}_{i,j}^L \preceq \mathbf{B}_j^L$. We can then define

$$\begin{aligned}
\mathbf{Y}_{i,j}^U &= \mathbf{C}_{i,j}^U - \tilde{\mathbf{A}}_j^T \tilde{\mathbf{A}}_j \\
\mathbf{Y}_{i,j}^L &= \tilde{\mathbf{A}}_j^T \tilde{\mathbf{A}}_j - \mathbf{C}_{i,j}^L \, .
\end{aligned}$$

We then have, similarly, $\mathbf{Y}_{i,i}^U = \mathbf{X}_i^U$, $\mathbf{Y}_{i,i}^L = \mathbf{X}_i^L$, and for $j \leq i$, $\mathbf{Y}_{i,j}^U \succeq \mathbf{X}_j^U$ and $\mathbf{Y}_{i,j}^L \succeq \mathbf{X}_j^L$.

Assume that $l_i < 1$. Otherwise, since $p_i \leq 1$ (it is a probability), we trivially have $\mathbb{E}\left[p_i\right] \leq 8/\varepsilon^2 \cdot l_i$ as desired. Now, note that for all $i > 0$,

$$\begin{aligned}
\mathbf{a}_i^T (\mathbf{Y}_{i,0}^U)^{-1} \mathbf{a}_i = \mathbf{a}_i^T (\mathbf{Y}_{i,0}^L)^{-1} \mathbf{a}_i& \\
= \mathbf{a}_i^T \left(\frac{\varepsilon}{2} \mathbf{A}_i^T \mathbf{A}_i + \delta \mathbf{I}\right)^{-1} \mathbf{a}_i& \\
= \frac{2}{\varepsilon} \mathbf{a}_i^T \left(\mathbf{A}_i^T \mathbf{A}_i + \frac{2\delta}{\varepsilon} \mathbf{I}\right)^{-1} \mathbf{a}_i \quad \leq \quad \frac{2}{\varepsilon} l_i \, .&
\end{aligned} \tag{4.2}$$

Next, we will show that for $j < i-1$,

$$\mathbb{E}\left[\mathbf{a}_i^T (\mathbf{Y}_{i-1,j+1}^U)^{-1} \mathbf{a}_i\right] \leq \mathbb{E}\left[\mathbf{a}_i^T (\mathbf{Y}_{i-1,j}^U)^{-1} \mathbf{a}_i\right] \tag{4.3}$$

and

$$\mathbb{E}\left[\mathbf{a}_i^T (\mathbf{Y}_{i-1,j+1}^L)^{-1} \mathbf{a}_i\right] \leq \mathbb{E}\left[\mathbf{a}_i^T (\mathbf{Y}_{i-1,j}^L)^{-1} \mathbf{a}_i\right] \, . \tag{4.4}$$

Combined with (4.2) and the fact that $\mathbf{Y}^U_{i,i} = \mathbf{X}^U_i$, $\mathbf{Y}^L_{i,i} = \mathbf{X}^L_i$, (4.3) and (4.4) give

$$\begin{aligned}
\mathbb{E}\left[p_i\right] &= c_U \mathbb{E}\left[\mathbf{a}_i^T (\mathbf{X}^U_{i-1})^{-1}\mathbf{a}_i\right] + c_L \mathbb{E}\left[\mathbf{a}_i^T (\mathbf{X}^L_{i-1})^{-1}\mathbf{a}_i\right] \\
&= c_U \mathbb{E}\left[\mathbf{a}_i^T (\mathbf{Y}^U_{i-1,i-1})^{-1}\mathbf{a}_i\right] + c_L \mathbb{E}\left[\mathbf{a}_i^T (\mathbf{Y}^L_{i-1,i-1})^{-1}\mathbf{a}_i\right] \\
&\le c_U \mathbb{E}\left[\mathbf{a}_i^T (\mathbf{Y}^U_{i-1,0})^{-1}\mathbf{a}_i\right] + c_L \mathbb{E}\left[\mathbf{a}_i^T (\mathbf{Y}^L_{i-1,0})^{-1}\mathbf{a}_i\right] \\
&= \frac{2}{\varepsilon} l_i \cdot (c_U + c_L) = \frac{8}{\varepsilon^2} l_i\,.
\end{aligned}$$

The last equality follows from the fact that in ONLINE-BSS we set $c_U = 2/\varepsilon + 1$ and $c_L = 2/\varepsilon - 1$. This completes the claim that for all i, the probability q_i that row $\mathbf{a}_i$ is sampled is bounded by $q_i = \mathbb{E}\left[p_i\right] \le 8/\varepsilon^2 \cdot l_i$, giving the second part of Theorem 4.1.

It remains to prove (4.3) and (4.4). To do this we will show a somewhat stronger statement: conditioned on any choices for the first j rows, the expected value of $\mathbf{a}_i^T (\mathbf{Y}^U_{i-1,j+1})^{-1}\mathbf{a}_i$ is no larger than that of $\mathbf{a}_i^T (\mathbf{Y}^U_{i-1,j})^{-1}\mathbf{a}_i$, and analogously for $(\mathbf{Y}^L_{i-1,j+1})^{-1}$. Similar to the proof of part 1, we consider two cases:

Case 1: $p_{j+1} = 1$. In that case, the positive semidefinite matrix $\mathbf{a}_{j+1}\mathbf{a}_{j+1}^T$ is added at step $j+1$ to give $\tilde{\mathbf{A}}_{j+1}^T\tilde{\mathbf{A}}_{j+1} = \tilde{\mathbf{A}}_j^T\tilde{\mathbf{A}}_j + \mathbf{a}_{j+1}\mathbf{a}_{j+1}^T$. This gives that

$$\begin{aligned}
\mathbf{Y}^U_{i-1,j+1} &= \mathbf{C}^U_{i,j+1} - \tilde{\mathbf{A}}_{j+1}^T\tilde{\mathbf{A}}_{j+1} \\
&= \mathbf{C}^U_{i,j} + \left(1 + \frac{\varepsilon}{2}\right)\mathbf{a}_{j+1}\mathbf{a}_{j+1}^T - (\tilde{\mathbf{A}}_j^T\tilde{\mathbf{A}}_j + \mathbf{a}_{j+1}\mathbf{a}_{j+1}^T) \\
&= \mathbf{Y}^U_{i-1,j} + \frac{\varepsilon}{2}\cdot \mathbf{a}_{j+1}\mathbf{a}_{j+1}^T.
\end{aligned}$$

Thus we have $\mathbf{Y}^U_{i-1,j+1} \succeq \mathbf{Y}^U_{i-1,j}$ and so $\mathbf{a}_i^T (\mathbf{Y}^U_{i-1,j+1})^{-1}\mathbf{a}_i \le \mathbf{a}_i^T (\mathbf{Y}^U_{i-1,j})^{-1}\mathbf{a}_i$, giving (4.3). An analogous argument holds for $\mathbf{Y}^L_{i-1,j+1}$, giving (4.4).

Case 2: $p_{j+1} < 1$**.** This case is more tricky. Importantly, by how p_{j+1} is set in step 3(b) of ONLINE-BSS and by the observation that $\mathbf{Y}^U_{i-1,j} \succeq \mathbf{X}^U_j$ and $\mathbf{Y}^L_{i-1,j} \succeq \mathbf{X}^L_j$ for $j \le i-1$ (recall that we must prove (4.3) and (4.4) under the assumption that $j \le i-1$), we have

$$p_{j+1} \ge c_U \cdot \mathbf{a}_{j+1}^T (\mathbf{X}^U_j)^{-1}\mathbf{a}_{j+1} \ge c_U \cdot \mathbf{a}_{j+1}^T (\mathbf{Y}^U_{i-1,j})^{-1}\mathbf{a}_{j+1} \tag{4.5}$$

and

$$p_{j+1} \ge c_L \cdot \mathbf{a}_{j+1}^T (\mathbf{X}^L_j)^{-1}\mathbf{a}_{j+1} \ge c_L \cdot \mathbf{a}_{j+1}^T (\mathbf{Y}^L_{i-1,j})^{-1}\mathbf{a}_{j+1}. \tag{4.6}$$

Now, we define $\mathbf{w}_{j+1} = \mathbf{a}_{j+1}/\sqrt{p_{j+1}}$ and additionally

$$\begin{aligned}
s^U_{j+1} &= \mathbf{w}_{j+1}^T (\mathbf{Y}^U_{i-1,j})^{-1}\mathbf{w}_{j+1}^T \\
s^L_{j+1} &= \mathbf{w}_{j+1}^T (\mathbf{Y}^L_{i-1,j})^{-1}\mathbf{w}_{j+1}^T \\
\mathbf{u}^U_{j+1} &= \frac{\mathbf{w}_{j+1}}{\sqrt{s^U_{j+1}}} \\
\mathbf{u}^L_{j+1} &= \frac{\mathbf{w}_{j+1}}{\sqrt{s^L_{j+1}}}\,.
\end{aligned}$$

We then deploy Lemma 4.2 to bound the expectations in (4.3) and (4.4).

Upper barrier bound (4.3)**:** Here we apply Lemma 4.2 with $\mathbf{X} = \mathbf{Y}^U_{i-1,j}$, $\mathbf{u} = \mathbf{u}^U_{j+1}$, $\mathbf{v} = \mathbf{a}_i^T$, $a = s^U_{j+1}(1-p_{j+1}(1+\varepsilon/2))$, $b = -s^U_{j+1}p_{j+1}(1+\varepsilon/2)$, $p = p_{j+1}$. Note that we have

$$\mathbf{u}^T\mathbf{X}^{-1}\mathbf{u} = 1/s^U_{j+1}\cdot\mathbf{w}^T_{j+1}(\mathbf{Y}^U_{i-1,j})^{-1}\mathbf{w}_{j+1} = 1$$

as required. Additionally, with probability p_{j+1},

$$\begin{aligned}
\widehat{\mathbf{X}} &= \mathbf{X} - a\mathbf{u}\mathbf{u}^T \\
&= \mathbf{Y}^U_{i-1,j} - s^U_{j+1}(1-p_{j+1}(1+\varepsilon/2))\cdot\mathbf{u}^U_{j+1}{\mathbf{u}^U_{j+1}}^T \\
&= \mathbf{Y}^U_{i-1,j} - (1-p_{j+1}(1+\varepsilon/2))\cdot\mathbf{w}_{j+1}{\mathbf{w}_{j+1}}^T \\
&= \mathbf{Y}^U_{i-1,j} - \frac{1}{p_{j+1}}\cdot\mathbf{a}_{j+1}{\mathbf{a}_{j+1}}^T + (1+\varepsilon/2)\mathbf{a}_{j+1}{\mathbf{a}_{j+1}}^T.
\end{aligned}$$

Similarly, with probability $1-p_{j+1}$,

$$\begin{aligned}
\widehat{\mathbf{X}} &= \mathbf{X} - b\mathbf{u}\mathbf{u}^T \\
&= \mathbf{Y}^U_{i-1,j} + s^U_{j+1}p_{j+1}(1+\varepsilon/2)\mathbf{u}^U_{j+1}{\mathbf{u}^U_{j+1}}^T \\
&= \mathbf{Y}^U_{i-1,j} + (1+\varepsilon/2)\mathbf{a}_{j+1}{\mathbf{a}_{j+1}}^T.
\end{aligned}$$

That is, $\widehat{\mathbf{X}} = \mathbf{Y}^U_{i-1,j+1}$. Thus, Lemma 4.2 gives

$$\begin{aligned}
\mathbb{E}\left[\mathbf{a}_i^T\widehat{\mathbf{X}}^{-1}\mathbf{a}_i - \mathbf{v}^T\mathbf{X}^{-1}\mathbf{v}\right] &= \mathbb{E}\left[\mathbf{a}_i^T{\mathbf{Y}^U_{i-1,j+1}}^{-1}\mathbf{a}_i\right] - \mathbb{E}\left[\mathbf{a}_i^T{\mathbf{Y}^U_{i-1,j}}^{-1}\mathbf{a}_i\right] \\
&= (\mathbf{v}^T\mathbf{X}^{-1}\mathbf{u})^2\cdot\frac{pa+(1-p)b-ab}{(1-a)(1-b)}.
\end{aligned}$$

To prove (4.3) it suffices to show that

$$\frac{pa+(1-p)b-ab}{(1-a)(1-b)}$$

is non-positive. Letting $r = \mathbf{a}^T_{j+1}(\mathbf{Y}^U_{i-1,j})^{-1}\mathbf{a}^T_{j+1}$ we can write $a = r\cdot(1/p_{j+1}-(1+\varepsilon/2))$ and $b = -r\cdot(1+\varepsilon/2) < 0$. By (4.5), $r \le p_{j+1}/c_U$ and thus $a \le r/p_{j+1} \le 1/c_U = 1/(2/\varepsilon+1) < 1$. Thus, the denominator $(1-a)(1-b)$ is positive, and so it remains to show that the numerator $pa+(1-p)b-ab$ is non-positive. We can write

$$\begin{aligned}
pa+(1-p)b-ab &= -r\cdot\frac{\varepsilon}{2} + r^2\cdot\left(\frac{1}{p_{j+1}} - \left(1+\frac{\varepsilon}{2}\right)\right)\cdot\left(1+\frac{\varepsilon}{2}\right) \\
&\le -r\cdot\frac{\varepsilon}{2} + r^2\cdot\left(\frac{1}{p_{j+1}}\right)\cdot\left(1+\frac{\varepsilon}{2}\right) \\
&\le -r\cdot\frac{\varepsilon}{2} + r\cdot\frac{1}{2/\varepsilon+1}\cdot\left(1+\frac{\varepsilon}{2}\right) \\
&= -r\cdot\frac{\varepsilon}{2} + r\cdot\frac{\varepsilon}{2} \le 0.
\end{aligned}$$

This completes the argument that $\mathbb{E}\left[\mathbf{a}_i^T \mathbf{Y}^U_{i-1,j+1}{}^{-1}\mathbf{a}_i\right] - \mathbb{E}\left[\mathbf{a}_i^T \mathbf{Y}^U_{i-1,j}{}^{-1}\mathbf{a}_i\right] \le 0$, giving (4.3).

Lower barrier bound (4.4)**:** For the lower barrier bound we give a similar argument. We use $\mathbf{X} = \mathbf{Y}^L_{i-1,j}$, $\mathbf{u} = \mathbf{u}^L_{j+1}$, $\mathbf{v} = \mathbf{a}_i^T$, $a = -s^L_{j+1}(1 - p_{j+1}(1-\varepsilon/2))$, $b = s^L_{j+1} p_{j+1}(1-\varepsilon/2)$, and $p = p_{j+1}$. We again have

$$\mathbf{u}^T\mathbf{X}^{-1}\mathbf{u} = 1/s^L_{j+1} \cdot \mathbf{w}^T_{j+1}(\mathbf{Y}^L_{i-1,j})^{-1}\mathbf{w}_{j+1} = 1,$$

as required. Additionally, with probability p_{j+1}, we have

$$\begin{aligned}
\widehat{\mathbf{X}} &= \mathbf{X} - a\mathbf{u}\mathbf{u}^T \\
&= \mathbf{Y}^L_{i-1,j} + s^L_{j+1}(1 - p_{j+1}(1-\varepsilon/2)) \cdot \mathbf{u}^L_{j+1}\mathbf{u}^L_{j+1}{}^T \\
&= \mathbf{Y}^L_{i-1,j} + \frac{1}{p_{j+1}} \cdot \mathbf{a}_{j+1}\mathbf{a}_{j+1}{}^T - (1-\varepsilon/2)\mathbf{a}_{j+1}\mathbf{a}_{j+1}{}^T.
\end{aligned}$$

Similarly, with probability $1 - p_{j+1}$,

$$\begin{aligned}
\widehat{\mathbf{X}} &= \mathbf{X} - b\mathbf{u}\mathbf{u}^T \\
&= \mathbf{Y}^L_{i-1,j} - s^L_{j+1} p_{j+1}(1-\varepsilon/2)\mathbf{u}^L_{j+1}\mathbf{u}^L_{j+1}{}^T \\
&= \mathbf{Y}^L_{i-1,j} - (1-\varepsilon/2)\mathbf{a}_{j+1}\mathbf{a}_{j+1}{}^T.
\end{aligned}$$

That is, $\widehat{\mathbf{X}} = \mathbf{Y}^L_{i-1,j+1}$. Thus, Lemma 4.2 gives

$$\begin{aligned}
\mathbb{E}\left[\mathbf{a}_i^T\widehat{\mathbf{X}}^{-1}\mathbf{a}_i - \mathbf{v}^T\mathbf{X}^{-1}\mathbf{v}\right] &= \mathbb{E}\left[\mathbf{a}_i^T(\mathbf{Y}^L_{i-1,j+1})^{-1}\mathbf{a}_i\right] - \mathbb{E}\left[\mathbf{a}_i^T(\mathbf{Y}^L_{i-1,j})^{-1}\mathbf{a}_i\right] \\
&= (\mathbf{v}^T\mathbf{X}^{-1}\mathbf{u})^2 \cdot \frac{pa + (1-p)b - ab}{(1-a)(1-b)}.
\end{aligned}$$

Again, to prove (4.4) it suffices to show that

$$\frac{pa + (1-p)b - ab}{(1-a)(1-b)}$$

is non-positive. Let $r = \mathbf{a}^T_{j+1}(\mathbf{Y}^L_{i-1,j})^{-1}\mathbf{a}_{j+1}$. We can write $a = -r(1/p_{j+1} - (1-\varepsilon/2)) < 0$ and $b = r(1-\varepsilon/2)$. Note that by (4.6), $r \le p_{j+1}/c_L = p_{j+1}/(2/\varepsilon - 1) < 1$, and thus $b < 1$. So the denominator $(1-a)(1-b)$ is positive. It thus remains to show that the numerator $pa + (1-p)b - ab$ is non-positive. We simplify this numerator as

$$\begin{aligned}
pa + (1-p)b - ab &= -r \cdot \frac{\varepsilon}{2} + r^2\left(\frac{1}{p_{j+1}} - \left(1 - \frac{\varepsilon}{2}\right)\right) \cdot \left(1 - \frac{\varepsilon}{2}\right) \\
&\le -r \cdot \frac{\varepsilon}{2} + r^2 \cdot \left(\frac{1}{p_{j+1}}\right) \cdot \left(1 - \frac{\varepsilon}{2}\right) \\
&\le -r \cdot \frac{\varepsilon}{2} + r \cdot \frac{1}{2/\varepsilon - 1} \cdot \left(1 - \frac{\varepsilon}{2}\right) \\
&= -r \cdot \frac{\varepsilon}{2} + r \cdot \frac{\varepsilon}{2} \le 0,
\end{aligned}$$

giving the required bound. This proves (4.4) and completes the theorem. □

5 Matching lower bound

Here we show that the row count obtained by Theorem 4.1 is in fact optimal. While it is possible to obtain a spectral approximation with $\mathcal{O}(d/\varepsilon^2)$ rows in the offline setting, online sampling always incurs a loss of $\Omega\left(\log(\varepsilon\|\mathbf{A}\|_2^2/\delta)\right)$ and must sample $\Omega\left(d\log(\varepsilon\|\mathbf{A}\|_2^2/\delta)/\varepsilon^2\right)$ rows.

Theorem 5.1. *Assume that $\varepsilon\|\mathbf{A}\|_2^2 \geq c_1\delta$ and $\varepsilon \geq c_2/\sqrt{d}$, for fixed constants c_1 and c_2. Then any algorithm that selects rows in an online manner and outputs a spectral approximation to $\mathbf{A}^T\mathbf{A}$ with $(1+\varepsilon)$ multiplicative error and δ additive error with probability at least $1/2$ must sample $\Omega\left(d\log(\varepsilon\|\mathbf{A}\|_2^2/\delta)/\varepsilon^2\right)$ rows of $\mathbf{A}$ in expectation.*

Note that the lower bounds we assume on $\varepsilon\|\mathbf{A}\|_2^2$ and ε are very minor. They just ensure that $\log(\varepsilon\|\mathbf{A}\|_2^2/\delta) \geq 1$ and that ε is not so small that we can essentially sample all rows.

Proof. We apply Yao's minimax principle, constructing, for any large enough M, a distribution on inputs $\mathbf{A}$ with $\|\mathbf{A}\|_2^2 \leq M$ for which any deterministic online row selection algorithm that succeeds with probability at least $1/2$ must output $\Omega\left(d\log(\varepsilon M/\delta)/\varepsilon^2\right)$ rows in expectation. The best randomized algorithm that works with probability $1/2$ on any input matrix with $\|\mathbf{A}\|_2^2 \leq M$ therefore must select at least $\Omega\left(d\log(\varepsilon M/\delta)/\varepsilon^2\right)$ rows in expectation on the worst case input, giving us the theorem.

Our distribution is as follows. We select an integer N uniformly at random from $[1, \log(M\varepsilon/\delta)]$. We then stream in the vertex-edge incidence matrices of N complete graphs on d vertices. We double the weight of each successive graph. Intuitively, spectrally approximating a complete graph requires selecting $\Omega(d/\varepsilon^2)$ edges [2] (as long as $\varepsilon \geq c_2/\sqrt{d}$ for some fixed constant c_2). Each time we stream in a new graph with double the weight, we force the algorithm to add $\Omega(d/\varepsilon^2)$ more edges to its output, eventually forcing it to return $\Omega(d/\varepsilon^2 \cdot N)$ edges, which is $\Omega(d\log(M\varepsilon/\delta)/\varepsilon^2)$ in expectation.

Specifically, let $\mathbf{K}_d$ be the $\binom{d}{2} \times d$ vertex-edge incidence matrix of the complete graph on d vertices. $\mathbf{K}_d^T\mathbf{K}_d$ is the Laplacian matrix of the complete graph on d vertices. We weight the first graph so that its Laplacian has all its nonzero eigenvalues equal to δ/ε. (That is, each edge has weight $\delta/(d\varepsilon)$). In this way, even if we select $N = \lfloor\log(M\varepsilon/\delta)\rfloor$ we have overall $\|\mathbf{A}\|_2^2 \leq \delta/\varepsilon + 2\delta/\varepsilon + \cdots + 2^{\lfloor\log(M\varepsilon/\delta)\rfloor-1}\delta/\varepsilon \leq M$.

Even if $N = 1$, all nonzero eigenvalues of $\mathbf{A}^T\mathbf{A}$ are at least δ/ε, so achieving $(1+\varepsilon)$ multiplicative error and $\delta\mathbf{I}$ additive error is equivalent to achieving $(1+2\varepsilon)$ multiplicative error. $\mathbf{A}^T\mathbf{A}$ is a graph Laplacian so has a null space. However, as all rows are orthogonal to the null space, achieving additive error $\delta\mathbf{I}$ is equivalent to achieving additive error $\delta\mathbf{I}_r$ where $\mathbf{I}_r$ is the identity projected to the span of $\mathbf{A}^T\mathbf{A}$. $\delta\mathbf{I}_r \preceq \varepsilon\mathbf{A}^T\mathbf{A}$ which is why we must achieve $(1+2\varepsilon)$ multiplicative error.

In order for a deterministic algorithm to be correct with probability $1/2$ on our distribution, it must be correct for at least $1/2$ of our $\lfloor\log(M\varepsilon/\delta)\rfloor$ possible choices of N.

Let i be the lowest choice of N for which the algorithm is correct. By the lower bound of [2], the algorithm must output $\Omega(d/\varepsilon^2)$ rows of $\mathbf{A}_i$ to achieve a $(1+2\varepsilon)$ multiplicative-error spectral approximation. Here $\mathbf{A}_i$ is the input consisting of the vertex-edge incidence matrices of i increasingly weighted complete graphs. Call the output on this input $\tilde{\mathbf{A}}_i$. Now let j be the second lowest choice of N on which the algorithm is correct. Since the algorithm was correct on $\mathbf{A}_i$ to within a multiplicative $(1+2\varepsilon)$, to be correct on $\mathbf{A}_j$, it must output a set of edges $\tilde{\mathbf{A}}_j$ such that

$$(\mathbf{A}_j^T\mathbf{A}_j - \mathbf{A}_i^T\mathbf{A}_i) - 4\varepsilon\mathbf{A}_j^T\mathbf{A}_j \preceq \tilde{\mathbf{A}}_j^T\tilde{\mathbf{A}}_j - \tilde{\mathbf{A}}_i^T\tilde{\mathbf{A}}_i \preceq (\mathbf{A}_j^T\mathbf{A}_j - \mathbf{A}_i^T\mathbf{A}_i) + 4\varepsilon\mathbf{A}_j^T\mathbf{A}_j\,.$$

Since we double each successive copy of the complete graph, $\mathbf{A}_j^T\mathbf{A}_j \preceq 2(\mathbf{A}_j^T\mathbf{A}_j - \mathbf{A}_i^T\mathbf{A}_i)$. So, $\tilde{\mathbf{A}}_j^T\tilde{\mathbf{A}}_j - \tilde{\mathbf{A}}_i^T\tilde{\mathbf{A}}_i$ must be a $1+8\varepsilon$ spectral approximation to the true difference $\mathbf{A}_j^T\mathbf{A}_j - \mathbf{A}_i^T\mathbf{A}_i$. Noting that this difference is itself just a weighting of the complete graph, by the lower bound in [2] the algorithm must select $\Omega(d/\varepsilon^2)$ additional edges between the i^{th} and j^{th} input graphs. Iterating this argument over all $\lfloor \log(M\varepsilon/\delta)\rfloor/2$ inputs on which the algorithm must be correct, it must select a total of $\Omega(d\log(M\varepsilon/\delta)/\varepsilon^2)$ edges in expectation over all inputs. □

6 Future work

The main open question arising from the original publication of this work [13] was if one could prove that the algorithm of [16] works despite dependencies arising due to the row pruning step. By operating in the online setting, our algorithm avoids row pruning, and hence is able to skirt these dependencies, as the probability that a row is sampled only depends on earlier rows in the stream. However, because the streaming setting offers the potential for sampling fewer rows than in the online case, obtaining a rigorous proof of [16] is very interesting. This open question was essentially resolved in [18], which presents an algorithm similar to the one presented in [16] for insertion-only streams that admits a correct proof.

While our work focuses on spectral approximation, variants on (ridge) leverage score sampling and the BSS algorithm are also used to solve low-rank approximation problems, including column subset selection [5, 12] and projection-cost-preserving sketching [10, 12]. Compared with spectral approximation, there is less work on streaming sampling for low-rank approximation, and understanding how online algorithms may be used in this setting would an interesting directino. Since initial publication, this question has been studied extensively [6, 8, 7], with online ridge leverage scores being employed for online low-rank approximation of kernel matrices and for low-rank approximation in sliding window streams.

Acknowledgements. The authors would like to thank Kenneth Clarkson, Jonathan Kelner, Gary Miller, Christopher Musco and Richard Peng for helpful discussions and comments. Cameron Musco and Jakub Pachocki both acknowledge the Gene Golub SIAM Summer School program on Randomization in Numerical Linear Algebra, where work on this project was initiated.

References

[1] AHMED ALAOUI AND MICHAEL W. MAHONEY: Fast randomized kernel ridge regression with statistical guarantees. In *Adv. Neural Info. Proc. Sys. 30 (NIPS'15)*, pp. 775–783. Curran Assoc., Inc., 2015. NIPS. [arXiv:1411.0306] 4, 5

[2] JOSHUA BATSON, DANIEL A. SPIELMAN, AND NIKHIL SRIVASTAVA: Twice-Ramanujan sparsifiers. *SIAM J. Comput.*, 41(6):1704–1721, 2012. Preliminary version in STOC'09. [doi:10.1137/090772873, arXiv:0808.0163] 4, 13, 14, 20, 21

[3] ANTOINE BORDES AND LÉON BOTTOU: The huller: a simple and efficient online SVM. In *Machine Learning: ECML'05*, pp. 505–512. Springer, 2005. [doi:10.1007/11564096_48] 3

[4] CHRISTOS BOUTSIDIS, DAN GARBER, ZOHAR KARNIN, AND EDO LIBERTY: Online principal components analysis. In *Proc. 26th Ann. ACM-SIAM Symp. on Discrete Algorithms (SODA'15)*, pp. 887–901. ACM Press, 2015. [doi:10.1137/1.9781611973730.61] 3

[5] CHRISTOS BOUTSIDIS AND DAVID P. WOODRUFF: Optimal CUR matrix decompositions. *SIAM J. Comput.*, 46(2):543–589, 2017. Preliminary version in STOC'14. [arXiv:1405.7910] 21

[6] VLADIMIR BRAVERMAN, PETROS DRINEAS, CAMERON MUSCO, CHRISTOPHER MUSCO, JALAJ UPADHYAY, DAVID P. WOODRUFF, AND SAMSON ZHOU: Numerical linear algebra in the online and sliding window models. In *Proc. 61st FOCS*, pp. 517–528. IEEE Comp. Soc., 2020. [doi:10.1109/FOCS46700.2020.00055, arXiv:1805.03765] 3, 21

[7] DANIELE CALANDRIELLO, ALESSANDRO LAZARIC, AND MICHAL VALKO: Efficient second-order online kernel learning with adaptive embedding. In *Adv. Neural Info. Proc. Sys. 30 (NIPS'17)*, pp. 6140–6150. Curran Assoc., Inc., 2017. NIPS. 3, 21

[8] DANIELE CALANDRIELLO, ALESSANDRO LAZARIC, AND MICHAL VALKO: Second-order kernel online convex optimization with adaptive sketching. In *Proc. 34th Int. Conf. on Machine Learning (ICML'17)*, pp. 645–653, 2017. MLR Press. [arXiv:1706.04892] 3, 21

[9] KENNETH L. CLARKSON AND DAVID P. WOODRUFF: Low rank approximation and regression in input sparsity time. *J. ACM*, 63(6):54:1–54:45, 2017. Preliminary version in in STOC'13. [doi:10.1145/3019134, arXiv:1207.6365] 2

[10] MICHAEL B. COHEN, SAM ELDER, CAMERON MUSCO, CHRISTOPHER MUSCO, AND MADALINA PERSU: Dimensionality reduction for k-means clustering and low rank approximation. In *Proc. 47th STOC*, pp. 163–172. ACM Press, 2015. [doi:10.1145/2746539.2746569, arXiv:1410.6801] 21

[11] MICHAEL B. COHEN, YIN TAT LEE, CAMERON MUSCO, CHRISTOPHER MUSCO, RICHARD PENG, AND AARON SIDFORD: Uniform sampling for matrix approximation. In *Proc. 6th Innovations in Theoret. Comp. Sci. conf. (ITCS'15)*, pp. 181–190. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2015. [doi:10.1145/2688073.2688113, arXiv:1408.5099] 2, 3, 5

[12] MICHAEL B. COHEN, CAMERON MUSCO, AND CHRISTOPHER MUSCO: Input sparsity time low-rank approximation via ridge leverage score sampling. In *Proc. 28th Ann. ACM-SIAM Symp. on Discrete Algorithms (SODA'17)*, pp. 1758–1777. ACM Press, 2017. [doi:10.1137/1.9781611974782.115, arXiv:1511.07263] 21

[13] MICHAEL B. COHEN, CAMERON MUSCO, AND JAKUB PACHOCKI: Online row sampling. In *Proc 19th Internat. Workshop on Approximation Algorithms for Combinat. Opt. Probl. (APPROX'16)*, pp. 7:1–7:18. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2016. [doi:10.4230/LIPIcs.APPROX-RANDOM.2016.7] 21

[14] KOBY CRAMMER, OFER DEKEL, JOSEPH KESHET, SHAI SHALEV-SHWARTZ, AND YORAM SINGER: Online passive-aggressive algorithms. *J. Mach. Learning Res.*, 7:551–585, 2006. JMLR. 3

[15] MICHAEL KAPRALOV, YIN TAT LEE, CAMERON MUSCO, CHRISTOPHER MUSCO, AND AARON SIDFORD: Single pass spectral sparsification in dynamic streams. *SIAM J. Comput.*, 46(1):456–477, 2017. Preliminary version in in FOCS'14. [doi:10.1137/141002281, arXiv:1407.1289] 2

[16] JONATHAN A. KELNER AND ALEX LEVIN: Spectral sparsification in the semi-streaming setting. *Theory Comput. Sys.*, 53(2):243–262, 2013. [doi:10.1007/s00224-012-9396-1] 2, 3, 4, 21

[17] IOANNIS KOUTIS, GARY L. MILLER, AND RICHARD PENG: Approaching optimality for solving SDD linear systems. *SIAM J. Comput.*, 43(1):337–354, 2014. Preliminary version in in FOCS'10. [doi:10.1137/110845914, arXiv:1003.2958] 5

[18] RASMUS KYNG, JAKUB PACHOCKI, RICHARD PENG, AND SUSHANT SACHDEVA: A framework for analyzing resparsification algorithms. In *Proc. 28th Ann. ACM-SIAM Symp. on Discrete Algorithms (SODA'17)*, pp. 2032–2043. ACM Press, 2017. [doi:10.1137/1.9781611974782.132, arXiv:1611.06940] 3, 21

[19] YIN TAT LEE AND HE SUN: Constructing linear-sized spectral sparsification in almost-linear time. *SIAM J. Comput.*, 47(6):2315–2336, 2018. Preliminary version in in FOCS'15. [doi:10.1137/16M1061850, arXiv:1508.03261] 13, 14

[20] MU LI, GARY L. MILLER, AND RICHARD PENG: Iterative row sampling. In *Proc. 54th FOCS*, pp. 127–136. IEEE Comp. Soc., 2013. [doi:10.1109/FOCS.2013.22, arXiv:1211.2713] 2

[21] EDO LIBERTY, RAM SRIHARSHA, AND MAXIM SVIRIDENKO: An algorithm for online k-means clustering. In *Proc. 18th Workshop on Algorithm Engineering and Experiments (ALENEX'16)*, pp. 81–89, 2016. [doi:10.1137/1.9781611974317.7, arXiv:1412.5721] 3

[22] XIANGRUI MENG AND MICHAEL W. MAHONEY: Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *Proc. 45th STOC*, pp. 91–100. ACM Press, 2013. [doi:10.1145/2488608.2488621, arXiv:1210.3135] 2

[23] JELANI NELSON AND HUY LÊ NGUYỄN: OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *Proc. 54th FOCS*, pp. 117–126. IEEE Comp. Soc., 2013. [doi:10.1109/FOCS.2013.21, arXiv:1211.1002] 2

[24] JACK SHERMAN AND WINIFRED J. MORRISON: Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *Ann. Math. Stat.*, 21(1):124–127, 1950. JSTOR. 15

[25] DANIEL A. SPIELMAN AND NIKHIL SRIVASTAVA: Graph sparsification by effective resistances. *SIAM J. Comput.*, 40(6):1913–1926, 2011. Preliminary version in in STOC'08. [doi:10.1137/080734029, arXiv:0803.0929] 7

[26] DANIEL A. SPIELMAN AND SHANG-HUA TENG: Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *Proc. 36th STOC*, pp. 81–90. ACM Press, 2004. [doi:10.1145/1007352.1007372, arXiv:cs/0310051] 2

[27] DANIEL A. SPIELMAN AND SHANG-HUA TENG: Nearly linear time algorithms for preconditioning and solving symmetric, diagonally dominant linear systems. *SIAM J. Matrix Anal. Appl.*, 35(3):835–885, 2014. [doi:10.1137/090771430] 7

[28] JOEL A. TROPP: Freedman's inequality for matrix martingales. *Electr. Comm. Probability*, 16(25):262–270, 2011. [doi:10.1214/ECP.v16-1624, arXiv:1101.3039] 8

AUTHORS

Michael B. Cohen†
Massachusetts Institute of Technology
Cambridge, MA, USA

Cameron Musco
Assistant Professor
College of Information and Computer Sciences
University of Massachusetts Amherst
Amherst, MA, USA
cmusco@cs.umass.edu
`https://people.cs.umass.edu/~cmusco/`

Jakub Pachocki
Research Lead
Open AI
San Francisco, CA, USA
jakub@openai.com

ABOUT THE AUTHORS

MICHAEL B. COHEN (1992–2017) received his Ph. D. from the Massachusetts Institute of Technology posthumously in 2018. He was advised by Jonathan Kelner. He was supported by an NSF Graduate Student Fellowship and was awarded the Machtey Award for the best student paper at FOCS 2016. He made important contributions to algorithmic spectral graph theory, randomized numerical linear algebra, and convex optimization.

Michael tragically passed away in September 2017 due to complications from undiagnosed Type I diabetes. He is dearly missed and remembered for his unbounded excitement, his energy, and his love of both knowledge and people.

CAMERON MUSCO is an Assistant Professor at the University of Massachusetts Amherst. He received his Ph. D. from Massachusetts Institute of Technology in 2018, where he was advised by Nancy Lynch and supported by an NSF Graduate Student Fellowship. After MIT, Cameron was a postdoctoral researcher at Microsoft Research New England. He studies algorithms, often randomized ones, working at the intersection of theoretical computer science, numerical linear algebra, optimization, and machine learning

JAKUB PACHOCKI is a Research Lead at OpenAI. He received his Ph. D. from Carnegie Mellon University in 2016, where he was advised by Gary Miller. After that he was a postdoctoral fellow under Jelani Nelson at Harvard University.