Linear Algebra and its Applications 573 (2019) 144–165

Contents lists available at ScienceDirect

Linear Algebra and its Applications

www.elsevier.com/locate/laa

Structural conditions for projection-cost preservation via randomized matrix multiplication

Agniva Chowdhury [a,1], Jiasen Yang [a,1], Petros Drineas [b,*]

[a] *Department of Statistics, Purdue University, West Lafayette, IN, United States of America*
[b] *Department of Computer Science, Purdue University, West Lafayette, IN, United States of America*

ARTICLE INFO

Article history:
Received 17 August 2018
Accepted 9 March 2019
Available online 18 March 2019
Submitted by V. Mehrmann

MSC:
15A23
15A45
65F30
68W20
68W25

Keywords:
Projection-cost preservation
Randomized linear algebra
Matrix sketching
Leverage scores

ABSTRACT

Projection-cost preservation is a low-rank approximation guarantee which ensures that the cost of any rank-k projection can be preserved using a smaller sketch of the original data matrix. We present a general structural result outlining four sufficient conditions to achieve projection-cost preservation. These conditions can be satisfied using tools from the Randomized Linear Algebra literature.

* Corresponding author.
E-mail addresses: chowdhu5@purdue.edu (A. Chowdhury), jiaseny@purdue.edu (J. Yang), pdrineas@purdue.edu (P. Drineas).
[1] Both authors contributed equally to this work.

https://doi.org/10.1016/j.laa.2019.03.013
0024-3795/

1. Introduction

Projection-cost preservation is a low-rank approximation guarantee which ensures that the cost of any rank-k projection can be preserved using a smaller sketch of the original data matrix. It has recently emerged as a fundamental principle in the design and analysis of sketching-based algorithms for common matrix operations that are critical in data mining and machine learning.

Prior to introducing the formal definition for cost-preserving projections, we state the general *constrained low-rank approximation* problem. For any matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$ and a set $\mathbf{\Omega}$ of orthogonal projection matrices $\mathbf{P} \in \mathbb{R}^{d\times d}$ with $\text{rank}(\mathbf{P}) = k$, we seek the minimizer of the following optimization problem:

$$\mathbf{P}^* = \underset{\mathbf{P}\in\mathbf{\Omega}}{\text{argmin}} \|\mathbf{A} - \mathbf{AP}\|_F^2 . \tag{1}$$

Here, the term $\|\mathbf{A}-\mathbf{AP}\|_F^2$ is called the *cost* of projection $\mathbf{P}$; recall that $\|\mathbf{X}\|_F^2 = \sum_{i,j} \mathbf{X}_{ij}^2$. Two simple examples will help clarify the importance of the above formulation. First, let $\mathbf{\Omega}$ be the set of all rank-k orthogonal projection matrices: in this case, the problem of eqn. (1) is equivalent to finding the best rank-k approximation to $\mathbf{A}$, which can be computed via the Singular Value Decomposition (SVD) in polynomial time. Second, let $\mathbf{A}$ be a data matrix whose columns represent d points in $\mathbb{R}^n$ and let $\mathbf{\Omega}$ be the set of all orthogonal rank-k projection matrices of the form $\mathbf{P} = \mathbf{XX}^\mathsf{T}$. If we let $\mathbf{X} \in \mathbb{R}^{d\times k}$ denote the rescaled *cluster-membership matrix*, *i.e.*, $\mathbf{X}_{ij} = 1/\sqrt{s_j}$ if the i-th column of $\mathbf{A}$ belongs to the j-th cluster and zero otherwise (s_j is the size of the j-th cluster),[2] then the constrained low-rank approximation problem of eqn. (1) is equivalent to the well-known k-means clustering problem [10].

The above discussion shows that, depending on the set $\mathbf{\Omega}$, the optimization problem of eqn. (1) can be easy or very hard to solve exactly. Indeed, the low-rank approximation problem can be solved in polynomial time via the SVD, whereas the k-means clustering problem is NP-hard [3] and a polynomial time algorithm is unlikely. The projection-cost preservation formulation of the optimization problem in eqn. (1) replaces the full matrix $\mathbf{A}$ by a smaller sketch $\widetilde{\mathbf{A}} \in \mathbb{R}^{s\times d}$, with $s \ll n$, in order to reduce the solution cost. The following definition first appeared in [8,23]; see Section 1.1 for a discussion of prior work.[3]

Definition 1. Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ be the input matrix and let $\mathbf{W} \in \mathbb{R}^{s\times n}$ with $s \ll n$ be a sketching matrix. The matrix $\mathbf{WA}$ is a rank-k *projection-cost preserving sketch* of $\mathbf{A}$ with error $\varepsilon \in [0, 1]$ if it satisfies

[2] Note that every row of $\mathbf{X}$ has exactly one non-zero element and its columns are pairwise orthogonal and normal.

[3] The original definitions of [8,23] included a non-negative, fixed constant c as an additive term, which does not add any generality in our setting.

$$(1-\varepsilon)\|\mathbf{A}-\mathbf{AP}\|_F^2 \leq \|\mathbf{WA}-\mathbf{WAP}\|_F^2 \leq (1+\varepsilon)\|\mathbf{A}-\mathbf{AP}\|_F^2, \tag{2}$$

for all rank-k projection matrices $\mathbf{P} \in \mathbb{R}^{d\times d}$ $(1 \leq k < d)$.

In words, the so-called *projection-cost preserving sketch* $\mathbf{WA}$ can replace the original matrix $\mathbf{A}$ in the optimization problem of eqn. (1) with a small loss in accuracy (see Lemma 3 of [8]) and thus one can solve the *sketched* problem instead of the original problem. We will now slightly manipulate Definition 1 by rewriting the rank-k projection matrix $\mathbf{P} \in \mathbb{R}^{d\times d}$ as follows: let $\mathbf{X} \in \mathbb{R}^{d\times(d-k)}$ be a matrix whose columns form a basis for the subspace that is *orthogonal* to the subspace spanned by $\mathbf{P}$. Thus, $\mathbf{P} = \mathbf{I}_d - \mathbf{X}\mathbf{X}^\mathsf{T}$ and $\mathbf{X}^\mathsf{T}\mathbf{X} = \mathbf{I}_{d-k}$ ($\mathbf{I}$ is a square identity matrix of appropriate dimensions). We can now rewrite the exact and approximate cost of the projection $\mathbf{P}$ as follows:

$$\|\mathbf{A}-\mathbf{AP}\|_F^2 = \|\mathbf{AXX}^\mathsf{T}\|_F^2 = \|\mathbf{AX}\|_F^2, \text{ and}$$
$$\|\mathbf{WA}-\mathbf{WAP}\|_F^2 = \|\mathbf{WAXX}^\mathsf{T}\|_F^2 = \|\mathbf{WAX}\|_F^2.$$

The final equalities in both derivations follow from the unitary invariance of the Frobenius norm. We can now state our (equivalent) definition of cost-preserving projections.

Definition 2 *(Cost-preserving projections).* Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ be the input matrix and let $\mathbf{W} \in \mathbb{R}^{s\times n}$ with $s \ll n$ be a sketching matrix. Then, $\mathbf{WA}$ is a rank-k projection-cost preserving sketch of $\mathbf{A} \in \mathbb{R}^{n\times d}$ with error $\varepsilon \in [0,1]$ if it satisfies

$$\left|\|\mathbf{WAX}\|_F^2 - \|\mathbf{AX}\|_F^2\right| \leq \varepsilon \|\mathbf{AX}\|_F^2, \tag{3}$$

for all matrices $\mathbf{X} \in \mathbb{R}^{d\times(d-k)}$ such that $\mathbf{X}^\mathsf{T}\mathbf{X} = \mathbf{I}_{d-k}$ $(1 \leq k < d)$.

Building upon the above definition, our main contribution is a general, structural result (Theorem 2) presenting four sufficient conditions that a sketching matrix $\mathbf{W}$ should satisfy in order to guarantee that the sketched matrix $\mathbf{WA}$ is a cost-preserving projection. The proposed sufficient conditions all boil down to sketching-based matrix multiplication (see Section 4.1 for a review), a fundamental and well-studied primitive of Randomized Linear Algebra (RLA). Such structural results have been of paramount importance in the RLA community, as they typically decouple the linear-algebraic component of a problem from the randomized algorithms that are employed to satisfy the structural conditions. See [19,15,13] for similar structural results for a variety of linear algebraic problems for which randomized algorithms have been designed, including ℓ_2 regression, SVD approximation, the Column Subset Selection Problem, etc. In Section 4, we instantiate our main result (Theorem 2) to show how different constructions of the sketching matrix $\mathbf{W}$ satisfy the structural conditions of our theorem. We hope that this linear-algebraic exposition of the cost-preserving projection problem will help bring it to the forefront of the linear algebra community and stimulate further research on this fundamental problem.

1.1. Related work

The importance of cost-preserving projections was first recognized by [8,23], who coined the aforementioned term for the problem of Definition 2. Their work provided several ways to construct provably accurate projection-cost preserving sketches and also demonstrated their applicability to constrained low-rank approximation problems, such as k-means clustering. Their work recognized the importance of structural results for cost-preserving projections and actually presented a related structural theorem, which is considerably more involved and complicated than our Theorem 2. In a more recent paper [9], the authors showed that ridge leverage score sampling also satisfies the cost-preserving projection guarantee. Interestingly, the results of [8,23] and [9] are quite independent and different from each other in terms of proof strategies, at least to the best of our understanding. A major motivation of our work was the unification of these two seemingly different approaches for cost-preserving projections using the same structural result.

We do note that prior to [8,23], the idea of projection-cost preservation was also discussed in [16] (see their Definition 2) and it was also implicit in [10,5]. Cost-preserving projections have also been connected with the construction of *coresets*[4] in machine learning, computational geometry, and theoretical computer science. We refer the interested reader to [4,16–18] for detailed discussions. In particular, [18] addressed the existence and efficient construction of coresets using a definition that is almost identical to Definition 2.

2. Notation

Given a matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$, let $\mathbf{A}_{i*}$ denote the i-th row of $\mathbf{A}$ as a row vector and let $\mathbf{A}_{*i}$ denote the i-th column of $\mathbf{A}$ as a column vector. Let the (thin) SVD of $\mathbf{A}$ be $\mathbf{A} = \mathbf{U}_{n\times r}\mathbf{\Sigma}_{r\times r}\mathbf{V}^{\mathsf{T}}_{d\times r}$; subscripts denote matrix dimensions and r denotes the rank of the matrix $\mathbf{A}$. It is well-known that $\mathbf{U}^{\mathsf{T}}\mathbf{U} = \mathbf{V}^{\mathsf{T}}\mathbf{V} = \mathbf{I}_r$ and $\mathbf{\Sigma} = \text{diag}\{\sigma_1, \ldots, \sigma_r\}$ consists of the non-zero singular values of $\mathbf{A}$ sorted in non-increasing order, $\sigma_1 \geq \cdots \geq \sigma_r > 0$. Let $\mathbf{A}_m$ be the best rank-m approximation to $\mathbf{A}$ and let $\mathbf{A}_{m,\perp} = \mathbf{A} - \mathbf{A}_m$. We will use the slightly non-standard notation $\mathbf{\Sigma}_m \triangleq \text{diag}\,\{\sigma_1, \ldots, \sigma_m, 0, \ldots, 0\}$ to denote the $r \times r$ diagonal matrix whose top m entries are equal to the top m singular values of $\mathbf{A}$ and the bottom $r - m$ entries are set to zero. Similarly, $\mathbf{\Sigma}_{m,\perp} \triangleq \text{diag}\,\{0, \ldots, 0, \sigma_{m+1}, \ldots, \sigma_r\}$ is the $r \times r$ diagonal matrix whose top m entries are set to zero and the bottom $r - m$ entries are set to the bottom $r - m$ singular values of $\mathbf{A}$. Clearly, $\mathbf{\Sigma} = \mathbf{\Sigma}_m + \mathbf{\Sigma}_{m,\perp}$, $\mathbf{A}_m = \mathbf{U}\mathbf{\Sigma}_m\mathbf{V}^{\mathsf{T}}$, and $\mathbf{A}_{m,\perp} = \mathbf{U}\mathbf{\Sigma}_{m,\perp}\mathbf{V}^{\mathsf{T}}$.

We will make frequent use of matrix norms, norm inequalities, and matrix trace properties; we refer the reader to Chapter 2 of [14] for a quick introduction. We do note

[4] In words, coresets are small sets of points that approximate the shape and properties of a larger set of points.

the strong submultiplicativity property of the Frobenius norm, namely that for any two matrices $\mathbf{A}$ and $\mathbf{B}$ of suitable dimensions,

$$\|\mathbf{AB}\|_F \leq \min\{\|\mathbf{A}\|_2 \cdot \|\mathbf{B}\|_F\,,\ \|\mathbf{A}\|_F \cdot \|\mathbf{B}\|_2\}.$$

An important tool in our analysis will be *von Neumann's trace inequality*; recall that the trace of a square matrix $\mathbf{A}$, denoted $\mathrm{tr}(\mathbf{A})$, is the sum of its diagonal entries.

Proposition 1 *(Von Neumann's trace inequality [22]). For any matrices* $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$ *with singular values* $\sigma_1(\mathbf{A}) \geq \sigma_2(\mathbf{A}) \geq \cdots \geq \sigma_n(\mathbf{A})$ *and* $\sigma_1(\mathbf{B}) \geq \sigma_2(\mathbf{B}) \geq \cdots \geq \sigma_n(\mathbf{B})$ *respectively,*

$$|\,\mathrm{tr}(\mathbf{AB})| \leq \sum_{i=1}^{n} \sigma_i(\mathbf{A})\,\sigma_i(\mathbf{B})\,.$$

In Section 4, we will make frequent use of a fundamental result from probability theory, known as *Markov's inequality*. Let X be a random variable assuming non-negative values with expectation $\mathbb{E}[X]$. Then, for any $t > 0$,

$$\mathbb{P}\left(X \geq t \cdot \mathbb{E}[X]\right) \leq \frac{1}{t}\,.$$

We will also need the so-called *union bound*: given a set of random events $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_n$ holding with respective probabilities $p_1, p_2, \ldots, p_n$, the probability that at least one of these events holds (*i.e.*, the probability of the union of these events) is upper-bounded by $\sum_{i=1}^{n} p_i$.

3. Our structural result

We now describe our main structural result for cost-preserving projections (see Definition 2). Our structural result connects cost-preserving projections with sketching-based matrix multiplication, a well-studied primitive in the Randomized Linear Algebra community.

Prior to presenting our result we define the diagonal matrix $\widetilde{\mathbf{\Sigma}} \in \mathbb{R}^{r\times r}$ as

$$\widetilde{\mathbf{\Sigma}} = \mathrm{diag}\{d_1, d_2, \ldots, d_m, \ldots, d_q, 0, \ldots, 0\}, \tag{4}$$

with $d_1 \geq d_2 \geq \ldots \geq d_m \geq \ldots \geq d_q > 0$. As we will discuss in more detail below, q and m are positive integers between one and r that are selected by the user of our structural result in order to satisfy the four conditions of Theorem 2; the same is true for the values d_i, $i = 1, \ldots, q$. It is precisely this flexibility in the construction of the matrix $\widetilde{\mathbf{\Sigma}}$ that makes Theorem 2 able to accommodate different constructions of the sketching matrix $\mathbf{W}$.

Theorem 2. *Let* $\mathbf{A} \in \mathbb{R}^{n\times d}$ *be the input matrix and let* $\mathbf{X} \in \mathbb{R}^{d\times(d-k)}$ *be any matrix satisfying* $\mathbf{X}^\mathsf{T}\mathbf{X} = \mathbf{I}$*, with* $1 \leq k < d$*. Let the thin SVD of* $\mathbf{A}$ *be* $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}$*; recall the definition of* $\widetilde{\mathbf{\Sigma}}$ *from eqn. (4), and assume that the sketching matrix* $\mathbf{W} \in \mathbb{R}^{s\times n}$ *satisfies the following four conditions for some accuracy parameter* ε*:*

$$\left\|\widetilde{\mathbf{\Sigma}}\mathbf{U}^\mathsf{T}\mathbf{W}^\mathsf{T}\mathbf{W}\mathbf{U}\widetilde{\mathbf{\Sigma}} - \widetilde{\mathbf{\Sigma}}^2\right\|_2 \leq \varepsilon, \tag{5}$$

$$\left\|\widetilde{\mathbf{\Sigma}}\mathbf{U}^\mathsf{T}\mathbf{W}^\mathsf{T}\mathbf{W}\mathbf{A}_{m,\perp} - \widetilde{\mathbf{\Sigma}}\mathbf{U}^\mathsf{T}\mathbf{A}_{m,\perp}\right\|_F \leq \varepsilon \left\|\mathbf{A}\mathbf{X}\right\|_F, \tag{6}$$

$$\left\|\mathbf{A}_{m,\perp}^\mathsf{T}\mathbf{W}^\mathsf{T}\mathbf{W}\mathbf{A}_{m,\perp} - \mathbf{A}_{m,\perp}^\mathsf{T}\mathbf{A}_{m,\perp}\right\|_F \leq \frac{\varepsilon}{\sqrt{k}} \left\|\mathbf{A}\mathbf{X}\right\|_F^2, \text{ and} \tag{7}$$

$$\left|\left\|\mathbf{W}\mathbf{A}_{m,\perp}\right\|_F^2 - \left\|\mathbf{A}_{m,\perp}\right\|_F^2\right| \leq \varepsilon \left\|\mathbf{A}\mathbf{X}\right\|_F^2. \tag{8}$$

Then,

$$\left|\left\|\mathbf{W}\mathbf{A}\mathbf{X}\right\|_F^2 - \left\|\mathbf{A}\mathbf{X}\right\|_F^2\right| \leq \left(d_m^{-2} + 2\,d_m^{-1} + 2\right)\varepsilon \left\|\mathbf{A}\mathbf{X}\right\|_F^2. \tag{9}$$

Several comments are necessary to better understand the above structural result. First of all, the four conditions of Theorem 2 need to be satisfied for a user-specified matrix $\widetilde{\mathbf{\Sigma}}$. To be precise, the user of the structural result has the flexibility to choose q (the number of non-zero diagonal entries of $\widetilde{\mathbf{\Sigma}}$) as well as the values of its entries d_i, $i = 1, \ldots, q$, subject to the constraint that the entries are decreasing and strictly positive. Second, the user of the structural result has the flexibility to choose the parameter m (which ranges between one and q) that appears in the last three conditions of Theorem 2. In particular, m is used to define the optimal rank-m approximation $\mathbf{A}_m$ to the input matrix $\mathbf{A}$ (see Section 2 for notation) and the (perpendicular) matrix $\mathbf{A}_{m,\perp}$ which satisfies $\mathbf{A} = \mathbf{A}_m + \mathbf{A}_{m,\perp}$ and $\mathbf{A}_m^\mathsf{T}\mathbf{A}_{m,\perp} = \mathbf{0}$. We emphasize that the conditions of Theorem 2 only need to hold for *a single user-specified* choice of m without affecting the generality of the theorem's conclusion. Third, all four conditions of Theorem 2 boil down to sketching-based matrix multiplication, as described in Section 4.1. Fourth, the final error bound depends on the accuracy parameter ε as well as the (user-specified) value d_m. As we will see in two different constructions of the sketching matrix $\mathbf{W}$ in Section 4, d_m is a small constant and thus the term in parentheses in the right-hand side of eqn. (9) can be easily replaced by a constant. Fifth, we note that the third constraint is a bit tighter (by a factor of $1/\sqrt{k}$) compared to the other ones. This turns out not to be a problem for known constructions of $\mathbf{W}$ since the product $\mathbf{A}_{m,\perp}^\mathsf{T}\mathbf{A}_{m,\perp}$ is easy to approximate by sketching. Sixth, more general versions of our structural result are possible: for example, one could remove the assumption that the entries of the diagonal matrix $\widetilde{\mathbf{\Sigma}}$ are decreasing. A more general result could be derived for the general case $d_i \neq 0$ for all $i = 1, \ldots, q$. We are not aware of a construction of the sketching matrix $\mathbf{W}$ that would necessitate this more general setting and, therefore, we refrain from introducing additional complexity to Theorem 2.

Proof. Throughout the proof, we will make heavy use of the notation in Section 2. We also introduce an additional piece of notation, namely the diagonal matrix $\widetilde{\mathbf{\Sigma}}_\perp \in \mathbb{R}^{r\times r}$, which is defined as

$$\widetilde{\mathbf{\Sigma}}_\perp = \mathrm{diag}\{\underbrace{0,\ldots,0}_{q},\underbrace{1,\ldots,1}_{r-q}\}. \tag{10}$$

We note that the inverse of $\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp$ always exists, since it is a diagonal matrix with non-zero entries. For notational convenience, let $\mathbf{Z} \triangleq \mathbf{U}^\mathsf{T}(\mathbf{I}_n - \mathbf{W}^\mathsf{T}\mathbf{W})\mathbf{U}$ and $\mathbf{Y} \triangleq \mathbf{V}^\mathsf{T}\mathbf{X}$. Using properties of the matrix trace and the SVD of $\mathbf{A}$, we can rewrite the quantity that we seek to bound in Theorem 2 as

$$\begin{aligned}
\left|\|\mathbf{AX}\|_F^2 - \|\mathbf{WAX}\|_F^2\right| &= \left|\mathrm{tr}\left(\mathbf{X}^\mathsf{T}\mathbf{A}^\mathsf{T}\left(\mathbf{I}_n - \mathbf{W}^\mathsf{T}\mathbf{W}\right)\mathbf{AX}\right)\right| \\
&= \left|\mathrm{tr}\left(\mathbf{X}^\mathsf{T}\mathbf{V}\mathbf{\Sigma}\mathbf{U}^\mathsf{T}\left(\mathbf{I}_n - \mathbf{W}^\mathsf{T}\mathbf{W}\right)\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}\mathbf{X}\right)\right| \\
&= \left|\mathrm{tr}\left(\mathbf{Y}^\mathsf{T}\mathbf{\Sigma}\mathbf{Z}\mathbf{\Sigma}\mathbf{Y}\right)\right| \\
&= \left|\mathrm{tr}\left(\mathbf{Y}^\mathsf{T}(\mathbf{\Sigma}_m + \mathbf{\Sigma}_{m,\perp})\,\mathbf{Z}\,(\mathbf{\Sigma}_m + \mathbf{\Sigma}_{m,\perp})\mathbf{Y}\right)\right| \\
&\le \underbrace{\left|\mathrm{tr}\left(\mathbf{Y}^\mathsf{T}\mathbf{\Sigma}_m\mathbf{Z}\mathbf{\Sigma}_m\mathbf{Y}\right)\right|}_{\Delta_1} + \underbrace{\left|\mathrm{tr}\left(\mathbf{Y}^\mathsf{T}\mathbf{\Sigma}_{m,\perp}\mathbf{Z}\mathbf{\Sigma}_{m,\perp}\mathbf{Y}\right)\right|}_{\Delta_2} \\
&\quad + 2\underbrace{\left|\mathrm{tr}\left(\mathbf{Y}^\mathsf{T}\mathbf{\Sigma}_m\mathbf{Z}\mathbf{\Sigma}_{m,\perp}\mathbf{Y}\right)\right|}_{\Delta_3}.
\end{aligned} \tag{11}$$

In the above derivations, we also used the fact that $\mathbf{\Sigma} = \mathbf{\Sigma}_m + \mathbf{\Sigma}_{m,\perp}$ (see Section 2).

Bounding Δ_1. We start by bounding the first term in eqn. (11):

$$\begin{aligned}
\Delta_1 &= \left|\mathrm{tr}\left(\mathbf{Y}^\mathsf{T}\mathbf{\Sigma}_m\mathbf{Z}\mathbf{\Sigma}_m\mathbf{Y}\right)\right| \\
&= \left|\mathrm{tr}\left(\mathbf{Y}^\mathsf{T}\mathbf{\Sigma}_m\big(\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp\big)^{-1}\big(\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp\big)\mathbf{Z}\big(\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp\big)\big(\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp\big)^{-1}\mathbf{\Sigma}_m\mathbf{Y}\right)\right| \\
&= \left|\mathrm{tr}\left(\mathbf{Y}^\mathsf{T}\big(\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp\big)^{-1}\mathbf{\Sigma}_m\big(\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp\big)\mathbf{Z}\big(\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp\big)\mathbf{\Sigma}_m\big(\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp\big)^{-1}\mathbf{Y}\right)\right| \\
&= \left|\mathrm{tr}\left(\mathbf{Y}^\mathsf{T}\big(\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp\big)^{-1}\mathbf{\Sigma}_m\,\widetilde{\mathbf{\Sigma}}\,\mathbf{Z}\,\widetilde{\mathbf{\Sigma}}\,\mathbf{\Sigma}_m\big(\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp\big)^{-1}\mathbf{Y}\right)\right| \\
&= \left|\mathrm{tr}\left(\mathbf{Y}^\mathsf{T}\,\mathbf{\Sigma}_m\big(\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp\big)^{-1}\,\widetilde{\mathbf{\Sigma}}\,\mathbf{Z}\,\widetilde{\mathbf{\Sigma}}\,\big(\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp\big)^{-1}\mathbf{\Sigma}_m\,\mathbf{Y}\right)\right| \\
&= \left|\mathrm{tr}\left(\mathbf{Y}^\mathsf{T}\mathbf{\Sigma}_m\big(\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp\big)^{-1}\,\mathbf{E}_1\,\big(\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp\big)^{-1}\mathbf{\Sigma}_m\mathbf{Y}\right)\right| \\
&= \left|\mathrm{tr}\left(\mathbf{E}_1\left(\big(\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp\big)^{-1}\mathbf{\Sigma}_m\mathbf{Y}\right)\left(\big(\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp\big)^{-1}\mathbf{\Sigma}_m\mathbf{Y}\right)^\mathsf{T}\right)\right|.
\end{aligned} \tag{12}$$

In the above, we set $\mathbf{E}_1 = \widetilde{\mathbf{\Sigma}}\mathbf{Z}\widetilde{\mathbf{\Sigma}} = \widetilde{\mathbf{\Sigma}}\mathbf{U}^\mathsf{T}\,\mathbf{W}^\mathsf{T}\mathbf{W}\,\mathbf{U}\widetilde{\mathbf{\Sigma}} - \widetilde{\mathbf{\Sigma}}^2$. Further, we used the fact that $\mathbf{\Sigma}_m\big(\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp\big) = \mathbf{\Sigma}_m\,\widetilde{\mathbf{\Sigma}}$, which follows from $\mathbf{\Sigma}_m\,\widetilde{\mathbf{\Sigma}}_\perp = \mathbf{0}$ (recall that $m \le q$). The last equality follows from the invariance of matrix trace under cyclic permutations.

Next, we apply von Neumann's trace inequality and the condition of eqn. (5) on the right hand side of eqn. (12) to get

$$\begin{aligned}\Delta_1 &\leq \sum_{i=1}^{m} \sigma_i(\mathbf{E}_1) \cdot \sigma_i^2 \left(\left(\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp\right)^{-1} \mathbf{\Sigma}_m \mathbf{Y} \right) \\ &\leq \varepsilon \sum_{i=1}^{m} \sigma_i^2 \left(\left(\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp\right)^{-1} \mathbf{\Sigma}_m \mathbf{Y} \right) \leq \varepsilon \left\| \left(\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp\right)^{-1} \mathbf{\Sigma}_m \mathbf{Y} \right\|_F^2 .\end{aligned}$$

Notice that $\left(\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp\right)^{-1} \mathbf{\Sigma}_m = (\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp)_m^{-1} \mathbf{\Sigma}_m$, where $(\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp)_m^{-1} \in \mathbb{R}^{r \times r}$ is a diagonal matrix whose top m diagonal entries are equal to those of $\left(\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp\right)^{-1}$ and the remaining $r - m$ diagonal entries are set to zero. Then,

$$\begin{aligned}\left\| \left(\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp\right)^{-1} \mathbf{\Sigma}_m \mathbf{Y} \right\|_F^2 &= \left\| (\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp)_m^{-1} \mathbf{\Sigma}_m \mathbf{Y} \right\|_F^2 \\ &\leq \left\| (\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp)_m^{-1} \right\|_2^2 \cdot \|\mathbf{\Sigma}_m \mathbf{Y}\|_F^2 \leq d_m^{-2} \|\mathbf{A}_m \mathbf{X}\|_F^2 . \qquad (13)\end{aligned}$$

The last inequality follows from $\left\| (\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp)_m^{-1} \right\|_2 = d_m^{-1}$ and the fact that (see Section 2)

$$\|\mathbf{\Sigma}_m \mathbf{Y}\|_F^2 = \left\| \mathbf{U} \mathbf{\Sigma}_m \mathbf{V}^\mathsf{T} \mathbf{X} \right\|_F^2 = \|\mathbf{A}_m \mathbf{X}\|_F^2 .$$

Therefore, we have shown that

$$\Delta_1 \leq d_m^{-2} \varepsilon \|\mathbf{A}_m \mathbf{X}\|_F^2 \leq d_m^{-2} \varepsilon \|\mathbf{A}\mathbf{X}\|_F^2 , \qquad (14)$$

where the last inequality follows from $\|\mathbf{A}\mathbf{X}\|_F^2 = \|\mathbf{A}_m \mathbf{X}\|_F^2 + \|\mathbf{A}_{m,\perp} \mathbf{X}\|_F^2$ (by the matrix Pythagorean theorem).

Bounding Δ_2. We now manipulate the second term of eqn. (11). Let $\mathbf{X}_\perp \in \mathbb{R}^{d \times k}$ form a basis for the space that is orthogonal to the space spanned by the columns of $\mathbf{X}$ (thus, $\mathbf{X}\mathbf{X}^\mathsf{T} + \mathbf{X}_\perp \mathbf{X}_\perp^\mathsf{T} = \mathbf{I}_d$). Furthermore, recall that $\mathbf{Y} = \mathbf{V}^\mathsf{T} \mathbf{X}$, $\mathbf{Z} = \mathbf{U}^\mathsf{T} \left(\mathbf{I}_n - \mathbf{W}^\mathsf{T} \mathbf{W}\right) \mathbf{U}$, and $\mathbf{A}_{m,\perp} = \mathbf{U} \mathbf{\Sigma}_{m,\perp} \mathbf{V}^\mathsf{T}$. Using the above definitions and properties of the matrix trace, we get

$$\begin{aligned}\Delta_2 &= \left| \mathrm{tr}\left(\mathbf{Y}^\mathsf{T} \mathbf{\Sigma}_{m,\perp} \mathbf{Z} \mathbf{\Sigma}_{m,\perp} \mathbf{Y} \right) \right| = \left| \mathrm{tr}\left(\mathbf{X}^\mathsf{T} \mathbf{A}_{m,\perp}^\mathsf{T} \left(\mathbf{I}_n - \mathbf{W}^\mathsf{T} \mathbf{W}\right) \mathbf{A}_{m,\perp} \mathbf{X} \right) \right| \\ &= \left| \mathrm{tr}\left(\mathbf{A}_{m,\perp}^\mathsf{T} \left(\mathbf{I}_n - \mathbf{W}^\mathsf{T} \mathbf{W}\right) \mathbf{A}_{m,\perp} \mathbf{X}\mathbf{X}^\mathsf{T} \right) \right| \\ &= \left| \mathrm{tr}\left(\mathbf{A}_{m,\perp}^\mathsf{T} \left(\mathbf{I}_n - \mathbf{W}^\mathsf{T} \mathbf{W}\right) \mathbf{A}_{m,\perp} (\mathbf{I}_d - \mathbf{X}_\perp \mathbf{X}_\perp^\mathsf{T}) \right) \right| \\ &= \left| \mathrm{tr}\left(\mathbf{A}_{m,\perp}^\mathsf{T} \left(\mathbf{I}_n - \mathbf{W}^\mathsf{T} \mathbf{W}\right) \mathbf{A}_{m,\perp} - \mathbf{A}_{m,\perp}^\mathsf{T} \left(\mathbf{I}_n - \mathbf{W}^\mathsf{T} \mathbf{W}\right) \mathbf{A}_{m,\perp} \mathbf{X}_\perp \mathbf{X}_\perp^\mathsf{T} \right) \right| \\ &\leq \underbrace{\left| \mathrm{tr}\left(\mathbf{A}_{m,\perp}^\mathsf{T} \left(\mathbf{I}_n - \mathbf{W}^\mathsf{T} \mathbf{W}\right) \mathbf{A}_{m,\perp} \right) \right|}_{\Delta_{21}} + \underbrace{\left| \mathrm{tr}\left(\mathbf{A}_{m,\perp}^\mathsf{T} \left(\mathbf{I}_n - \mathbf{W}^\mathsf{T} \mathbf{W}\right) \mathbf{A}_{m,\perp} \mathbf{X}_\perp \mathbf{X}_\perp^\mathsf{T} \right) \right|}_{\Delta_{22}} . \qquad (15)\end{aligned}$$

We now bound Δ_{21} and Δ_{22} separately. Using the structural condition of eqn. (8), we obtain

$$\begin{aligned}\Delta_{21} &= \left|\mathrm{tr}\left(\mathbf{A}_{m,\perp}^{\mathsf{T}}\left(\mathbf{I}_n - \mathbf{W}^{\mathsf{T}}\mathbf{W}\right)\mathbf{A}_{m,\perp}\right)\right| = \left|\mathrm{tr}\left(\mathbf{A}_{m,\perp}^{\mathsf{T}}\mathbf{A}_{m,\perp} - \mathbf{A}_{m,\perp}^{\mathsf{T}}\mathbf{W}^{\mathsf{T}}\mathbf{W}\mathbf{A}_{m,\perp}\right)\right| \\ &= \left|\|\mathbf{A}_{m,\perp}\|_F^2 - \|\mathbf{W}\mathbf{A}_{m,\perp}\|_F^2\right| \leq \varepsilon\,\|\mathbf{A}\mathbf{X}\|_F^2\,. \end{aligned} \tag{16}$$

In order to bound Δ_{22}, we will apply *von Neumann*'s trace inequality, the Cauchy-Schwarz inequality, and the structural condition of eqn. (7). For notational convenience, let $\mathbf{E}_2 = \mathbf{A}_{m,\perp}^{\mathsf{T}}\left(\mathbf{I}_n - \mathbf{W}^{\mathsf{T}}\mathbf{W}\right)\mathbf{A}_{m,\perp}$ and proceed as follows:

$$\begin{aligned}\Delta_{22} &= \left|\mathrm{tr}\left(\mathbf{A}_{m,\perp}^{\mathsf{T}}\left(\mathbf{I}_n - \mathbf{W}^{\mathsf{T}}\mathbf{W}\right)\mathbf{A}_{m,\perp}\mathbf{X}_\perp\mathbf{X}_\perp^{\mathsf{T}}\right)\right| = \left|\mathrm{tr}\left(\mathbf{E}_2\mathbf{X}_\perp\mathbf{X}_\perp^{\mathsf{T}}\right)\right| \\ &\leq \sum_{i=1}^{k}\sigma_i(\mathbf{E}_2)\cdot\sigma_i(\mathbf{X}_\perp\mathbf{X}_\perp^{\mathsf{T}}) \leq \left[\sum_{i=1}^{k}\sigma_i^2(\mathbf{E}_2)\right]^{\frac{1}{2}}\cdot\left[\sum_{i=1}^{k}\sigma_i^2(\mathbf{X}_\perp\mathbf{X}_\perp^{\mathsf{T}})\right]^{\frac{1}{2}} \qquad (17) \\ &\leq \sqrt{k}\,\|\mathbf{E}_2\|_F \leq \varepsilon\,\|\mathbf{A}\mathbf{X}\|_F^2\,. \qquad (18)\end{aligned}$$

It is important to note that the matrix $\mathbf{X}_\perp$ has rank k and thus has at most k non-zero singular values (all equal to one), which explains the fact that the summation in eqn. (17) stops at k. The last two inequalities follow from $\sum_{i=1}^{k}\sigma_i^2(\mathbf{E}_2) \leq \|\mathbf{E}_2\|_F^2$ and the structural condition of eqn. (7). Combining eqns. (15), (16), and (18), we obtain

$$\Delta_2 \leq 2\,\varepsilon\,\|\mathbf{A}\mathbf{X}\|_F^2\,. \tag{19}$$

Bounding Δ_3. Finally, we consider the third term of eqn. (11). Again, using $\mathbf{Y} = \mathbf{V}^{\mathsf{T}}\mathbf{X}$, $\mathbf{Z} = \mathbf{U}^{\mathsf{T}}\left(\mathbf{I}_n - \mathbf{W}^{\mathsf{T}}\mathbf{W}\right)\mathbf{U}$, and $\mathbf{A}_{m,\perp} = \mathbf{U}\boldsymbol{\Sigma}_{m,\perp}\mathbf{V}^{\mathsf{T}}$, we get

$$\begin{aligned}\Delta_3 &= \left|\mathrm{tr}\left(\mathbf{Y}^{\mathsf{T}}\boldsymbol{\Sigma}_m\mathbf{Z}\boldsymbol{\Sigma}_{m,\perp}\mathbf{Y}\right)\right| \\ &= \left|\mathrm{tr}\left(\mathbf{Y}^{\mathsf{T}}\boldsymbol{\Sigma}_m\mathbf{U}^{\mathsf{T}}\left(\mathbf{I}_n - \mathbf{W}^{\mathsf{T}}\mathbf{W}\right)\mathbf{A}_{m,\perp}\mathbf{X}\right)\right| \\ &= \left|\mathrm{tr}\left(\mathbf{Y}^{\mathsf{T}}\boldsymbol{\Sigma}_m\left(\widetilde{\boldsymbol{\Sigma}} + \widetilde{\boldsymbol{\Sigma}}_\perp\right)^{-1}\left(\widetilde{\boldsymbol{\Sigma}} + \widetilde{\boldsymbol{\Sigma}}_\perp\right)\mathbf{U}^{\mathsf{T}}\left(\mathbf{I}_n - \mathbf{W}^{\mathsf{T}}\mathbf{W}\right)\mathbf{A}_{m,\perp}\mathbf{X}\right)\right| \\ &= \left|\mathrm{tr}\left(\mathbf{Y}^{\mathsf{T}}\left(\widetilde{\boldsymbol{\Sigma}} + \widetilde{\boldsymbol{\Sigma}}_\perp\right)^{-1}\boldsymbol{\Sigma}_m\left(\widetilde{\boldsymbol{\Sigma}} + \widetilde{\boldsymbol{\Sigma}}_\perp\right)\mathbf{U}^{\mathsf{T}}\left(\mathbf{I}_n - \mathbf{W}^{\mathsf{T}}\mathbf{W}\right)\mathbf{A}_{m,\perp}\mathbf{X}\right)\right| \\ &= \left|\mathrm{tr}\left(\mathbf{Y}^{\mathsf{T}}\left(\widetilde{\boldsymbol{\Sigma}} + \widetilde{\boldsymbol{\Sigma}}_\perp\right)^{-1}\boldsymbol{\Sigma}_m\widetilde{\boldsymbol{\Sigma}}\,\mathbf{U}^{\mathsf{T}}\left(\mathbf{I}_n - \mathbf{W}^{\mathsf{T}}\mathbf{W}\right)\mathbf{A}_{m,\perp}\mathbf{X}\right)\right| \\ &= \left|\mathrm{tr}\left(\mathbf{Y}^{\mathsf{T}}\boldsymbol{\Sigma}_m\left(\widetilde{\boldsymbol{\Sigma}} + \widetilde{\boldsymbol{\Sigma}}_\perp\right)^{-1}\widetilde{\boldsymbol{\Sigma}}\,\mathbf{U}^{\mathsf{T}}\left(\mathbf{I}_n - \mathbf{W}^{\mathsf{T}}\mathbf{W}\right)\mathbf{A}_{m,\perp}\mathbf{X}\right)\right|. \end{aligned} \tag{20}$$

In the above, we repeatedly used the fact that matrix multiplication of diagonal matrices is commutative and $\boldsymbol{\Sigma}_m\widetilde{\boldsymbol{\Sigma}}_\perp = \mathbf{0}$. Next, we apply *von Neumann*'s trace inequality and the Cauchy-Schwarz inequality to eqn. (20) to get

$$\begin{aligned}\Delta_3 &\le \sum_{i=1}^{m} \sigma_i\left(\mathbf{Y}^\mathsf{T}\mathbf{\Sigma}_m\big(\widetilde{\mathbf{\Sigma}}+\widetilde{\mathbf{\Sigma}}_\perp\big)^{-1}\right)\cdot\sigma_i\left(\widetilde{\mathbf{\Sigma}}\mathbf{U}^\mathsf{T}(\mathbf{I}_n-\mathbf{W}^\mathsf{T}\mathbf{W})\mathbf{A}_{m,\perp}\mathbf{X}\right)\\ &\le \left[\sum_{i=1}^{m}\sigma_i^2\left(\mathbf{Y}^\mathsf{T}\mathbf{\Sigma}_m\big(\widetilde{\mathbf{\Sigma}}+\widetilde{\mathbf{\Sigma}}_\perp\big)^{-1}\right)\right]^{\frac{1}{2}}\cdot\left[\sum_{i=1}^{m}\sigma_i^2\left(\widetilde{\mathbf{\Sigma}}\mathbf{U}^\mathsf{T}(\mathbf{I}_n-\mathbf{W}^\mathsf{T}\mathbf{W})\mathbf{A}_{m,\perp}\mathbf{X}\right)\right]^{\frac{1}{2}}\\ &= \|\mathbf{Y}^\mathsf{T}\mathbf{\Sigma}_m\big(\widetilde{\mathbf{\Sigma}}+\widetilde{\mathbf{\Sigma}}_\perp\big)^{-1}\|_F\cdot\|\widetilde{\mathbf{\Sigma}}\mathbf{U}^\mathsf{T}(\mathbf{I}_n-\mathbf{W}^\mathsf{T}\mathbf{W})\mathbf{A}_{m,\perp}\mathbf{X}\|_F\\ &\le \|\mathbf{Y}^\mathsf{T}\mathbf{\Sigma}_m\big(\widetilde{\mathbf{\Sigma}}+\widetilde{\mathbf{\Sigma}}_\perp\big)^{-1}\|_F\cdot\|\widetilde{\mathbf{\Sigma}}\mathbf{U}^\mathsf{T}(\mathbf{I}_n-\mathbf{W}^\mathsf{T}\mathbf{W})\mathbf{A}_{m,\perp}\|_F\,. \end{aligned} \tag{21}$$

The last inequality follows from strong submultiplicativity and the fact that $\|\mathbf{X}\|_2 = 1$. Note that in eqn. (13) we proved that $\|\mathbf{Y}^\mathsf{T}\mathbf{\Sigma}_m\big(\widetilde{\mathbf{\Sigma}}+\widetilde{\mathbf{\Sigma}}_\perp\big)^{-1}\|_F \le d_m^{-1}\|\mathbf{A}_m\mathbf{X}\|_F$, while the structural condition of eqn. (6) gives

$$\|\widetilde{\mathbf{\Sigma}}\mathbf{U}^\mathsf{T}(\mathbf{I}_n-\mathbf{W}^\mathsf{T}\mathbf{W})\mathbf{A}_{m,\perp}\|_F = \|\widetilde{\mathbf{\Sigma}}\mathbf{U}^\mathsf{T}\mathbf{A}_{m,\perp}-\widetilde{\mathbf{\Sigma}}\mathbf{U}^\mathsf{T}\,\mathbf{W}^\mathsf{T}\mathbf{W}\,\mathbf{A}_{m,\perp}\|_F \le \varepsilon\|\mathbf{A}\mathbf{X}\|_F\,.$$

Thus, eqn. (21) can be bounded as

$$\Delta_3 \le d_m^{-1}\,\|\mathbf{A}_m\mathbf{X}\|_F\cdot\varepsilon\|\mathbf{A}\mathbf{X}\|_F \le\; d_m^{-1}\,\varepsilon\,\|\mathbf{A}\mathbf{X}\|_F^2\,, \tag{22}$$

where the last inequality follows from $\|\mathbf{A}\mathbf{X}\|_F^2 = \|\mathbf{A}_m\mathbf{X}\|_F^2 + \|\mathbf{A}_{m,\perp}\mathbf{X}\|_F^2$ (by the matrix Pythagorean theorem).

Final bound. Combining eqns. (11), (14), (19), and (22) concludes the proof of eqn. (9). □

4. Satisfying the conditions of Theorem 2

In this section we show how to satisfy the conditions of Theorem 2 using various constructions for the sketching matrix $\mathbf{W}$. Of particular interest is that our structural result of Theorem 2 unifies the sketching matrix constructions of [8] and [9].

4.1. Review of randomized matrix multiplication

We present a brief review of randomized matrix multiplication and some relevant theoretical results from prior work that will be useful in this section. Consider a simple algorithm (Algorithm 1) to construct a *sampling-and-rescaling* matrix $\mathbf{W} \in \mathbb{R}^{s\times n}$. Using Algorithm 1 we can approximate the matrix product $\mathbf{AB}$ by $\mathbf{AW}^\mathsf{T}\mathbf{WB}$, where $\mathbf{AW}^\mathsf{T}$ is the sketch of $\mathbf{A}$ and $\mathbf{WB}$ is the sketch of $\mathbf{B}$.

Lemma 3. *Given matrices $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{n\times p}$, let $\mathbf{W} \in \mathbb{R}^{s\times n}$ be constructed using Algorithm 1. Then,*

$$\mathbb{E}\left[\|\mathbf{AW}^\mathsf{T}\mathbf{WB}-\mathbf{AB}\|_F^2\right] \le \sum_{i=1}^{n}\frac{\|\mathbf{A}_{*i}\|_2^2\cdot\|\mathbf{B}_{i*}\|_2^2}{sp_i}\,. \tag{23}$$

Algorithm 1 Construct sampling-and-rescaling matrix.

Input: Probabilities p_k, $k = 1, \ldots, n$; integer $s \ll n$;
Initialize: $\mathbf{W} \leftarrow \mathbf{0}_{s \times n}$;
for $i = 1$ **to** s **do**
  Pick $j_i \in \{1, \ldots, n\}$ with $\mathbb{P}(j_i = k) = p_k$;
  $\mathbf{W}_{ij_i} \leftarrow (s\, p_{j_i})^{-\frac{1}{2}}$;
end for
Output: Sampling-and-rescaling matrix $\mathbf{W}$;

Furthermore, if $m = p$,

$$\mathbb{E}\left[\left(\mathrm{tr}\left(\mathbf{A}\mathbf{W}^{\mathsf{T}}\mathbf{W}\mathbf{B} - \mathbf{A}\mathbf{B}\right)\right)^2\right] \leq \sum_{i=1}^{n} \frac{\left[(\mathbf{B}\mathbf{A})_{ii}\right]^2}{s p_i}\,. \tag{24}$$

Eqn. (23) was proven in [11] (see Lemma 3); the proof of eqn. (24) is a simple exercise using the setup of Lemma 3 in [11].

Lemma 4. *Given matrices $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times p}$ and some constant $\beta \in (0, 1]$, let $\mathbf{W} \in \mathbb{R}^{s\times n}$ be constructed using Algorithm 1 with*

$$p_i \geq \beta \frac{\|\mathbf{A}_{*i}\|_2^2}{\|\mathbf{A}\|_F^2}\,, \quad \forall i = 1, \ldots, n$$

such that $\sum_{i=1}^{n} p_i = 1$. Then,

$$\mathbb{E}\left[\|\mathbf{A}\mathbf{W}^{\mathsf{T}}\mathbf{W}\mathbf{B} - \mathbf{A}\mathbf{B}\|_F^2\right] \leq \frac{1}{\beta s}\|\mathbf{A}\|_F^2 \cdot \|\mathbf{B}\|_F^2\,. \tag{25}$$

Furthermore, if $m = p$,

$$\mathbb{E}\left[\left(\mathrm{tr}\left(\mathbf{A}\mathbf{W}^{\mathsf{T}}\mathbf{W}\mathbf{B} - \mathbf{A}\mathbf{B}\right)\right)^2\right] \leq \frac{1}{\beta s}\|\mathbf{A}\|_F^2 \cdot \|\mathbf{B}\|_F^2\,. \tag{26}$$

The proof of the above lemma is immediate from Lemma 3 (with an application of the Cauchy-Schwartz inequality which implies that $((\mathbf{B}\mathbf{A})_{ii})^2 \leq \|\mathbf{A}_{*i}\|_2^2 \cdot \|\mathbf{B}_{i*}\|_2^2$). Finally, the next lemma appeared in [6] as Theorem 3 and is a strengthening of Theorem 4.2 of [20] for the special case when $\|\mathbf{A}\|_2 \leq 1$. We also note that Lemma 5 is implicit in [9].

Lemma 5. *Given $\mathbf{A} \in \mathbb{R}^{m\times n}$ with $\|\mathbf{A}\|_2 \leq 1$, let $\mathbf{W} \in \mathbb{R}^{s\times n}$ be constructed using Algorithm 1 with $p_i \geq \beta \|\mathbf{A}_{i*}\|_2^2 / \|\mathbf{A}\|_F^2$ for all $i = 1, \ldots, n$ and $\beta \in (0, 1]$ such that $\sum_{i=1}^{n} p_i = 1$. Let δ be a failure probability and $\varepsilon > 0$ be an accuracy parameter. If the number of sampled columns s satisfies*

$$s \geq 2\left(1 + \frac{\varepsilon}{3}\right)\frac{\|\mathbf{A}\|_F^2}{\beta\,\varepsilon^2}\,\ln\left(\frac{4\,(1 + \|\mathbf{A}\|_F^2)}{\delta}\right),$$

then, with probability at least $1-\delta$,

$$\left\|\mathbf{A}\mathbf{W}^{\mathsf{T}}\mathbf{W}\mathbf{A}-\mathbf{A}\mathbf{A}^{\mathsf{T}}\right\|_2 \leq \varepsilon.$$

4.2. Leverage score–based sampling

Our first approach constructs a *sampling-and-rescaling* matrix $\mathbf{W}$ using Algorithm 1 with sampling probabilities p_i, $i=1,\ldots,n$:

$$p_i \triangleq \frac{1}{2}\frac{\|(\mathbf{U}_k)_{i*}\|_2^2}{k}+\frac{1}{2}\frac{\|(\mathbf{A}_{k,\perp})_{i*}\|_2^2}{\|(\mathbf{A}_{k,\perp})\|_F^2}. \tag{27}$$

Clearly $\sum_{i=1}^{n} p_i = 1$. Recall that $(\mathbf{U}_k)_{i*}$ denotes the i-th row of the matrix of the top k left singular vectors of $\mathbf{A}$, while $(\mathbf{A}_{k,\perp})_{i*}$ denotes the i-th row of the matrix $\mathbf{A}_{k,\perp}=\mathbf{A}-\mathbf{A}_k$ (here $\mathbf{A}_k$ is the best rank-k approximation to $\mathbf{A}$). It is well-known that the quantities $\|(\mathbf{U}_k)_{i*}\|_2^2$ for $i=1,\ldots,n$ correspond to the so-called leverage scores of the best rank-k approximation to $\mathbf{A}$ (see [12,21] for detailed discussions of the leverage scores and their properties). The sampling probabilities p_i of eqn. (27) are a linear combination of the aforementioned leverage scores and a quantity that depends on the row norms of the residual matrix $\mathbf{A}_{k,\perp}$. It is worth noting that, to the best of our knowledge, using only the leverage scores as the sampling probabilities to construct the sampling-and-rescaling matrix $\mathbf{W}$ would not suffice to satisfy all conditions of Theorem 2.

We are now ready to apply Theorem 2 in order to analyze the performance of the matrix $\mathbf{W}$ that is constructed using the above procedure. First, recall that, as users of Theorem 2, we have full control in the construction of the matrix $\widetilde{\mathbf{\Sigma}}$. Towards that end, let $\widetilde{\mathbf{\Sigma}}$ be constructed as

$$\widetilde{\mathbf{\Sigma}} = \mathrm{diag}\{\underbrace{1,\ldots,1}_{k},\underbrace{0,\ldots,0}_{r-k}\}, \tag{28}$$

where both m and q (two parameters associated with the matrix $\widetilde{\mathbf{\Sigma}}$) are set to be equal to k. Trivially, $d_m=1$ by the construction of $\widetilde{\mathbf{\Sigma}}$. Therefore the constant at the right-hand side of eqn. (9) is equal to five.

Satisfying the condition of eqn. (5) Using our definition for $\widetilde{\mathbf{\Sigma}}$, we rewrite the left hand side of the structural eqn. (5) as follows:

$$\begin{aligned}\left\|\widetilde{\mathbf{\Sigma}}\mathbf{U}^{\mathsf{T}}\mathbf{W}^{\mathsf{T}}\mathbf{W}\mathbf{U}\widetilde{\mathbf{\Sigma}}-\widetilde{\mathbf{\Sigma}}^2\right\|_2 &= \left\|\begin{pmatrix}\mathbf{I}_k & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{U}^{\mathsf{T}}\mathbf{W}^{\mathsf{T}}\mathbf{W}\mathbf{U}\begin{pmatrix}\mathbf{I}_k & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}-\begin{pmatrix}\mathbf{I}_k & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\right\|_2\\ = \left\|\begin{pmatrix}\mathbf{U}_k^{\mathsf{T}}\mathbf{W}^{\mathsf{T}}\mathbf{W}\mathbf{U}_k-\mathbf{I}_k & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\right\|_2 &= \left\|\mathbf{U}_k^{\mathsf{T}}\mathbf{W}^{\mathsf{T}}\mathbf{W}\mathbf{U}_k-\mathbf{I}_k\right\|_2.\end{aligned} \tag{29}$$

Notice that the sampling probabilities p_i satisfy $p_i \geq \frac{1}{2}\frac{\|(\mathbf{U}_k)_{i*}\|_2^2}{k}$ for $i = 1, \ldots, n$. Thus, combining eqn. (29) and Lemma 5, we get

$$\mathbb{P}\left(\|\widetilde{\mathbf{\Sigma}}\mathbf{U}^\mathsf{T}\mathbf{W}^\mathsf{T}\mathbf{W}\mathbf{U}\widetilde{\mathbf{\Sigma}} - \widetilde{\mathbf{\Sigma}}^2\|_2 \geq \varepsilon\right) \leq \frac{\delta}{4}. \tag{30}$$

For the above bound to hold, we need to set the number of sampled rows of $\mathbf{A}$,

$$s \geq \left(1 + \frac{\varepsilon}{3}\right)\frac{4\,k\,\ln\left(16\,(1+k)/\delta\right)}{\varepsilon^2}.$$

Satisfying the condition of eqn. (6) We start by proving a simple inequality that will be useful in subsequent derivations (recall the definition of $\mathbf{X}_\perp$ from Section 3):

$$\|\mathbf{A}_{k,\perp}\|_F = \|\mathbf{A} - \mathbf{A}_k\|_F^2 \leq \|\mathbf{A} - \mathbf{A}\mathbf{X}_\perp\mathbf{X}_\perp^\mathsf{T}\|_F = \|\mathbf{A}\mathbf{X}\mathbf{X}^\mathsf{T}\|_F = \|\mathbf{A}\mathbf{X}\|_F. \tag{31}$$

The inequality in the above derivation is due to the fact that $\mathbf{A}_k$ is the *best* rank-k approximation to $\mathbf{A}$. We now use the definition of $\widetilde{\mathbf{\Sigma}}$ to rewrite the condition as follows:

$$\begin{aligned}
&\left\|\widetilde{\mathbf{\Sigma}}\mathbf{U}^\mathsf{T}\mathbf{W}^\mathsf{T}\mathbf{W}\mathbf{A}_{k,\perp} - \widetilde{\mathbf{\Sigma}}\mathbf{U}^\mathsf{T}\mathbf{A}_{k,\perp}\right\|_F \\
&= \left\|\begin{pmatrix}\mathbf{I}_k & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{U}_k^\mathsf{T}\\ \mathbf{U}_{k,\perp}^\mathsf{T}\end{pmatrix}\mathbf{W}^\mathsf{T}\mathbf{W}\mathbf{A}_{k,\perp} - \begin{pmatrix}\mathbf{I}_k & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{U}_k^\mathsf{T}\\ \mathbf{U}_{k,\perp}^\mathsf{T}\end{pmatrix}\mathbf{A}_{k,\perp}\right\|_F \\
&= \left\|\begin{pmatrix}\mathbf{U}_k^\mathsf{T}\\ \mathbf{0}\end{pmatrix}\mathbf{W}^\mathsf{T}\mathbf{W}\mathbf{A}_{k,\perp} - \begin{pmatrix}\mathbf{U}_k^\mathsf{T}\\ \mathbf{0}\end{pmatrix}\mathbf{A}_{k,\perp}\right\|_F = \left\|\mathbf{U}_k^\mathsf{T}\mathbf{W}^\mathsf{T}\mathbf{W}\mathbf{A}_{k,\perp} - \underbrace{\mathbf{U}_k^\mathsf{T}\mathbf{A}_{k,\perp}}_{\mathbf{0}}\right\|_F.
\end{aligned} \tag{32}$$

We emphasize that $\mathbf{U}_k^\mathsf{T}\mathbf{A}_{k,\perp} = \mathbf{0}$. Using the fact that $p_i \geq \frac{1}{2}\frac{\|(\mathbf{U}_k)_{i*}\|_2^2}{k}$ and applying Lemma 4, we obtain

$$\mathbb{E}\left[\|\mathbf{U}_k^\mathsf{T}\mathbf{W}^\mathsf{T}\mathbf{W}\mathbf{A}_{k,\perp}\|_F^2\right] \leq \frac{2}{s}\|\mathbf{U}_k\|_F^2 \cdot \|\mathbf{A}_{k,\perp}\|_F^2 = \frac{2k}{s}\|\mathbf{A}_{k,\perp}\|_F^2 \leq \frac{2k}{s}\|\mathbf{A}\mathbf{X}\|_F^2,$$

where we used the fact $\|\mathbf{U}_k\|_F^2 = k$, and the last inequality follows from eqn. (31). Applying Markov's inequality, we get

$$\mathbb{P}\left(\|\mathbf{A}_{k,\perp}^\mathsf{T}\mathbf{W}^\mathsf{T}\mathbf{W}\mathbf{U}_k\|_F \geq \varepsilon\,\|\mathbf{A}\mathbf{X}\|_F\right) \leq \frac{\delta}{4}. \tag{33}$$

The above bound holds if the number of sampled rows $s \geq 8k/\delta\varepsilon^2$.

Satisfying the conditions of eqns. (7) *and* (8) We note that $p_i \geq \frac{1}{2}\frac{\|(\mathbf{A}_{k,\perp})_{i*}\|_2^2}{\|(\mathbf{A}_{k,\perp})\|_F^2}$ for $i = 1, \ldots, n$. Applying Lemma 4 and using eqn. (31), we obtain

$$\mathbb{E}\left[\|\mathbf{A}_{k,\perp}^{\mathsf{T}}\mathbf{W}^{\mathsf{T}}\mathbf{W}\mathbf{A}_{k,\perp}-\mathbf{A}_{k,\perp}^{\mathsf{T}}\mathbf{A}_{k,\perp}\|_F^2\right] \leq \frac{2}{s}\|\mathbf{A}_{k,\perp}\|_F^4 \leq \frac{2}{s}\|\mathbf{A}\mathbf{X}\|_F^4,$$

$$\mathbb{E}\left[\left(\|\mathbf{W}\mathbf{A}_{k,\perp}\|_F^2-\|\mathbf{A}_{k,\perp}\|_F^2\right)^2\right] \leq \frac{2}{s}\|\mathbf{A}_{k,\perp}\|_F^4 \leq \frac{2}{s}\|\mathbf{A}\mathbf{X}\|_F^4.$$

In the above derivations, we used the fact that

$$\operatorname{tr}\left(\mathbf{A}_{k,\perp}^{\mathsf{T}}\mathbf{W}^{\mathsf{T}}\mathbf{W}\mathbf{A}_{k,\perp}-\mathbf{A}_{k,\perp}^{\mathsf{T}}\mathbf{A}_{k,\perp}\right)=\|\mathbf{W}\mathbf{A}_{k,\perp}\|_F^2-\|\mathbf{A}_{k,\perp}\|_F^2.$$

Next, by Markov's inequality, we get

$$\mathbb{P}\left(\left|\|\mathbf{W}\mathbf{A}_{k,\perp}\|_F^2-\|\mathbf{A}_{k,\perp}\|_F^2\right| \geq \frac{\varepsilon}{\sqrt{k}}\|\mathbf{A}\mathbf{X}\|_F^2\right) \leq \frac{\delta}{4}, \tag{34}$$

$$\mathbb{P}\left(\|\mathbf{A}_{k,\perp}^{\mathsf{T}}\mathbf{W}^{\mathsf{T}}\mathbf{W}\mathbf{A}_{k,\perp}-\mathbf{A}_{k,\perp}^{\mathsf{T}}\mathbf{A}_{k,\perp}\|_F \geq \frac{\varepsilon}{\sqrt{k}}\|\mathbf{A}\mathbf{X}\|_F^2\right) \leq \frac{\delta}{4}. \tag{35}$$

For the above bounds to hold, we need to set the number of sampled rows $s \geq 8k/\delta\varepsilon^2$. It is worth noting that the bound of eqn. (34) is *stronger* (by a factor of $1/\sqrt{k}$) compared to what is needed in Theorem 2. This improved bound comes for free given the value of s used in the construction of $\mathbf{W}$ and does not affect the tightness of the overall bound.

Finally, applying the union bound to eqns. (30), (33), (34), and (35), we conclude that if the number of sampled rows s satisfies

$$s \geq \max\left\{\left(1+\frac{\varepsilon}{3}\right)\frac{4\,k\,\ln\left(16\,(1+k)/\delta\right)}{\varepsilon^2}, \frac{8k}{\delta\,\varepsilon^2}\right\},$$

then all four structural conditions of Theorem 2 hold with probability at least $1-\delta$. Therefore, the number of sampled rows s is, asymptotically (assuming that δ is constant), $s = \mathcal{O}\left(k\ln k/\varepsilon^2\right)$.

We conclude this section by noting that a similar proof strategy (using the same construction for the matrix $\widetilde{\mathbf{\Sigma}}$ of eqn. (28)) could also be used to prove that all five constructions of sketching matrices described in Lemma 11 of [8] return cost-preserving projections.

4.3. Ridge leverage score sampling

Our second approach constructs a *sampling-and-rescaling* matrix $\mathbf{W}$ using Algorithm 1 with sampling probabilities p_i that are proportional to the so-called *ridge leverage scores* [2,9] of the rows of the matrix $\mathbf{A}$. To properly define the ridge leverage scores of the rows of $\mathbf{A}$, we first define the $r \times r$ diagonal matrix $\mathbf{\Sigma}_\lambda$ as follows:

$$\mathbf{\Sigma}_\lambda = \operatorname{diag}\left\{\frac{\sigma_1}{\sqrt{\sigma_1^2+\lambda}}, \ldots, \frac{\sigma_r}{\sqrt{\sigma_r^2+\lambda}}\right\}. \tag{36}$$

Recall that r is the rank of the matrix $\mathbf{A}$. The i-th *row ridge leverage score*, denoted by τ_i^λ, of $\mathbf{A}$ with respect to the ridge parameter $\lambda > 0$ is given by

$$\tau_i^\lambda \triangleq \left(\mathbf{A}(\mathbf{A}^\mathsf{T}\mathbf{A} + \lambda \mathbf{I}_d)^{-1}\mathbf{A}^\mathsf{T}\right)_{ii} = \|(\mathbf{U}\mathbf{\Sigma}_\lambda)_{i*}\|_2^2, \tag{37}$$

for $i = 1, \ldots, n$. Recall that $(\mathbf{U}\mathbf{\Sigma}_\lambda)_{i*}$ denotes the i-th row of the matrix of *all* the left singular vectors of $\mathbf{A}$, rescaled by the diagonal entries of $\mathbf{\Sigma}_\lambda$. The last equality in eqn. (37) follows by using the SVD of $\mathbf{A}$ and the definition of the matrix $\mathbf{\Sigma}_\lambda$. Let d_λ denote the sum of the ridge leverage scores, *i.e.*,

$$d_\lambda = \sum_{i=1}^n \tau_i^\lambda = \sum_{i=1}^n \|(\mathbf{U}\mathbf{\Sigma}_\lambda)_{i*}\|_2^2 = \|\mathbf{U}\mathbf{\Sigma}_\lambda\|_F^2 = \|\mathbf{\Sigma}_\lambda\|_F^2 . \tag{38}$$

The last equality follows by the unitary invariance of the Frobenius norm. We can now define the sampling probabilities p_i, $i = 1, \ldots, n$, as

$$p_i \triangleq \frac{\tau_i^\lambda}{\sum_{i=1}^n \tau_i^\lambda} = \frac{\tau_i^\lambda}{d_\lambda} = \frac{\|(\mathbf{U}\mathbf{\Sigma}_\lambda)_{i*}\|_2^2}{\|\mathbf{\Sigma}_\lambda\|_F^2} . \tag{39}$$

Clearly, $\sum_{i=1}^n p_i = 1$. In the remainder of this section, we will analyze the special case where

$$\lambda = \frac{\|\mathbf{A}_{k,\perp}\|_F^2}{k} . \tag{40}$$

This is the case analyzed in [9] and the simplest known setting for λ that returns provably accurate approximations for cost-preserving projections via ridge leverage score sampling.

We now proceed to apply Theorem 2 in order to analyze the performance of the matrix $\mathbf{W}$ that is constructed using the ridge leverage scores as sampling probabilities. Recall that, as users of Theorem 2, we have full control of the construction of the matrix $\widetilde{\mathbf{\Sigma}}$. Towards that end, let

$$\widetilde{\mathbf{\Sigma}} = \mathbf{\Sigma}_\lambda. \tag{41}$$

For the parameters associated with $\widetilde{\mathbf{\Sigma}}$ in eqn. (4), we will set q to r, the rank of the matrix $\mathbf{A}$; and set m to be the index of smallest non-zero singular value of $\mathbf{A}$ such that

$$\sigma_m^2 \geq \lambda \geq \sigma_{m+1}^2. \tag{42}$$

Several observations follow from the above definitions. First of all, the diagonal entries of $\widetilde{\mathbf{\Sigma}}$ (denoted as d_i in eqn. (4)) are set to $d_i = \sigma_i/\sqrt{\sigma_i^2+\lambda}$. We can upper-bound d_m^{-1} as

$$d_m^{-1} = \sqrt{1 + \frac{\lambda}{\sigma_m^2}} \leq \sqrt{2},$$

where the last inequality follows from our choice for m in eqn. (42). This implies that the constant in the right-hand side of eqn. (9) is at most $4 + 2\sqrt{2}$. Second, using our choices for m (eqn. (42)) and λ (eqn. (40)), we get

$$\|\mathbf{A}_{k,\perp}\|_F^2 = k\lambda \geq k\sigma_{m+1}^2 \geq \sigma_{m+1}^2 + \sigma_{m+2}^2 + \ldots \sigma_{m+k}^2 \,.$$

Adding $\|\mathbf{A}_{k,\perp}\|_F^2$ on both sides of the above inequality yields

$$\begin{aligned} 2\,\|\mathbf{A}_{k,\perp}\|_F^2 &\geq (\sigma_{m+1}^2 + \sigma_{m+2}^2 + \ldots \sigma_{m+k}^2) + (\sigma_{k+1}^2 + \sigma_{k+2}^2 + \ldots \sigma_r^2) \\ &\geq \sigma_{m+1}^2 + \sigma_{m+2}^2 + \ldots \sigma_r^2 = \|\mathbf{A}_{m,\perp}\|_F^2 \,, \end{aligned} \tag{43}$$

where the last inequality holds because $m + k \geq k + 1$. Combining eqns. (31) and (43) results in the following inequality, which will be quite useful in this section:

$$\|\mathbf{A}_{m,\perp}\|_F^2 \leq 2\,\|\mathbf{A}_{k,\perp}\|_F^2 \leq 2\,\|\mathbf{A}\mathbf{X}\|_F^2 \,. \tag{44}$$

Third, we can upper-bound the sum of the ridge leverage scores (denoted as d_λ in eqn. (38)) as follows:

$$\begin{aligned} d_\lambda = \|\mathbf{\Sigma}_\lambda\|_F^2 &= \sum_{i=1}^{r} \frac{\sigma_i^2}{\sigma_i^2 + \lambda} = \sum_{i=1}^{k} \frac{\sigma_i^2}{\sigma_i^2 + \lambda} + \sum_{i=k+1}^{r} \frac{\sigma_i^2}{\sigma_i^2 + \lambda} \\ &\leq k + \sum_{i=k+1}^{r} \frac{\sigma_i^2}{\lambda} = k + \frac{\|\mathbf{A}_{k,\perp}\|_F^2}{\lambda} = k + k = 2k. \end{aligned} \tag{45}$$

In the above, we upper-bounded the top k diagonal entries of the matrix $\mathbf{\Sigma}_\lambda$ (squared) by one and the bottom $r - k$ entries by σ_i^2/λ. To conclude, we used our specific choice for λ from eqn. (40). Again, this upper-bound for d_λ will be useful later in this section.

Satisfying the condition of eqn. (5) Applying Lemma 5 and setting s, the number of sampled rows, to be at least

$$s \geq \left(1 + \frac{\varepsilon}{3}\right) \frac{4\,k\,\ln\left(16\,(1+2k)/\delta\right)}{\varepsilon^2} \,,$$

we obtain

$$\mathbb{P}\left(\|\mathbf{\Sigma}_\lambda \mathbf{U}^\mathsf{T}\mathbf{W}^\mathsf{T}\mathbf{W}\mathbf{U}\mathbf{\Sigma}_\lambda - \mathbf{\Sigma}_\lambda^2\|_2 \geq \varepsilon\right) \leq \frac{\delta}{4} \,. \tag{46}$$

In the above we used $d_\lambda \leq 2k$ from eqn. (45).

Satisfying the condition of eqn. (6) Applying Lemma 4, we get

$$\mathbb{E}\left[\left\|\mathbf{\Sigma}_\lambda \mathbf{U}^\mathsf{T}\mathbf{W}^\mathsf{T}\mathbf{W}\mathbf{A}_{m,\perp} - \mathbf{\Sigma}_\lambda \mathbf{U}^\mathsf{T}\mathbf{A}_{m,\perp}\right\|_F^2\right] \leq \frac{1}{s}\left\|\mathbf{\Sigma}_\lambda \mathbf{U}^\mathsf{T}\right\|_F^2 \cdot \left\|\mathbf{A}_{m,\perp}\right\|_F^2$$
$$= \frac{d_\lambda}{s}\left\|\mathbf{A}_{m,\perp}\right\|_F^2 \leq \frac{4k}{s}\left\|\mathbf{A}\mathbf{X}\right\|_F^2 . \quad (47)$$

The last inequality follows since $d_\lambda = \|\mathbf{\Sigma}_\lambda \mathbf{U}^\mathsf{T}\|_F^2$ is upper-bounded by $2k$ from eqn. (45); we also used eqn. (44). Next, applying Markov's inequality, if the number of sampled rows $s \geq {}^{16\,k}/_{\delta\varepsilon^2}$, then

$$\mathbb{P}\left(\left\|\mathbf{\Sigma}_\lambda \mathbf{U}^\mathsf{T}\mathbf{W}^\mathsf{T}\mathbf{W}\mathbf{A}_{m,\perp} - \mathbf{\Sigma}_\lambda \mathbf{U}^\mathsf{T}\mathbf{A}_{m,\perp}\right\|_F > \varepsilon \left\|\mathbf{A}\mathbf{X}\right\|_F\right) \leq \frac{\delta}{4} . \quad (48)$$

Satisfying the conditions of eqns. (7) *and* (8) We start with eqn. (8). Using standard properties of the trace, we get

$$\begin{aligned}\left\|\mathbf{W}\mathbf{A}_{m,\perp}\right\|_F^2 &= \left\|\mathbf{W}\mathbf{U}\mathbf{\Sigma}_{m,\perp}\right\|_F^2 = \operatorname{tr}\left(\mathbf{\Sigma}_{m,\perp}\mathbf{U}^\mathsf{T}\mathbf{W}^\mathsf{T}\mathbf{W}\mathbf{U}\mathbf{\Sigma}_{m,\perp}\right) \\ &= \operatorname{tr}\left(\mathbf{\Sigma}_{m,\perp}\mathbf{\Sigma}_\lambda^{-1}\mathbf{\Sigma}_\lambda \mathbf{U}^\mathsf{T}\mathbf{W}^\mathsf{T}\mathbf{W}\mathbf{U}\mathbf{\Sigma}_{m,\perp}\right) \\ &= \operatorname{tr}\left(\mathbf{\Sigma}_\lambda \mathbf{U}^\mathsf{T}\mathbf{W}^\mathsf{T}\mathbf{W}\mathbf{U}\mathbf{\Sigma}_{m,\perp}^2\mathbf{\Sigma}_\lambda^{-1}\right) . \quad (49)\end{aligned}$$

Similarly,

$$\left\|\mathbf{A}_{m,\perp}\right\|_F^2 = \operatorname{tr}\left(\mathbf{\Sigma}_\lambda \mathbf{U}^\mathsf{T}\mathbf{U}\mathbf{\Sigma}_{m,\perp}^2\mathbf{\Sigma}_\lambda^{-1}\right) . \quad (50)$$

Combining Lemma 4 with eqns. (49) and (50), we have

$$\begin{aligned}&\mathbb{E}\left[\left(\left\|\mathbf{W}\mathbf{A}_{m,\perp}\right\|_F^2 - \left\|\mathbf{A}_{m,\perp}\right\|_F^2\right)^2\right] \\ &= \mathbb{E}\left[\left(\operatorname{tr}(\mathbf{\Sigma}_\lambda \mathbf{U}^\mathsf{T}\mathbf{W}^\mathsf{T}\mathbf{W}\mathbf{U}\mathbf{\Sigma}_{m,\perp}^2\mathbf{\Sigma}_\lambda^{-1} - \mathbf{\Sigma}_\lambda \mathbf{U}^\mathsf{T}\mathbf{U}\mathbf{\Sigma}_{m,\perp}^2\mathbf{\Sigma}_\lambda^{-1})\right)^2\right] \\ &\leq \frac{1}{s}\left\|\mathbf{\Sigma}_\lambda \mathbf{U}^\mathsf{T}\right\|_F^2 \cdot \left\|\mathbf{U}\mathbf{\Sigma}_{m,\perp}^2\mathbf{\Sigma}_\lambda^{-1}\right\|_F^2 = \frac{d_\lambda}{s}\left\|\mathbf{\Sigma}_{m,\perp}^2\mathbf{\Sigma}_\lambda^{-1}\right\|_F^2 \\ &\leq \frac{2k}{s}\left\|\mathbf{\Sigma}_{m,\perp}\right\|_F^2 \cdot \left\|\mathbf{\Sigma}_{m,\perp}\mathbf{\Sigma}_\lambda^{-1}\right\|_2^2 , \quad (51)\end{aligned}$$

where the last inequality follows from strong submultiplicativity and eqn. (45). Note that $\mathbf{\Sigma}_{m,\perp}\mathbf{\Sigma}_\lambda^{-1}$ is a diagonal matrix whose i-th diagonal entry is equal to

$$(\mathbf{\Sigma}_{m,\perp}\mathbf{\Sigma}_\lambda^{-1})_{ii} = \begin{cases} 0, & i \leq m ; \\ \sqrt{\sigma_i^2 + \lambda}, & i \geq m+1 . \end{cases}$$

It now follows that

$$\|\mathbf{\Sigma}_{m,\perp}\mathbf{\Sigma}_{\lambda}^{-1}\|_2 = \sqrt{\sigma_{m+1}^2 + \lambda} \leq \sqrt{2\lambda} = \sqrt{\frac{2}{k}}\|\mathbf{A}_{k,\perp}\|_F\,,$$

where the inequality follows from eqn. (42) and the last equality follows from eqn. (40). In addition, eqn. (44) yields

$$\|\mathbf{\Sigma}_{m,\perp}\|_F^2 = \|\mathbf{A}_{m,\perp}\|_F^2 \leq 2\|\mathbf{A}\mathbf{X}\|_F^2\,,$$

and thus we get,

$$\mathbb{E}\left[\left(\|\mathbf{W}\mathbf{A}_{m,\perp}\|_F^2 - \|\mathbf{A}_{m,\perp}\|_F^2\right)^2\right] \leq \frac{2k}{s}\cdot 2\|\mathbf{A}\mathbf{X}\|_F^2 \cdot \frac{2\|\mathbf{A}_{k,\perp}\|_F^2}{k} \leq \frac{8}{s}\,\|\mathbf{A}\mathbf{X}\|_F^4\,, \quad (52)$$

where the last inequality follows from eqn. (31). Using similar algebraic manipulations and Lemma 4, we obtain

$$\mathbb{E}\left(\left\|\mathbf{A}_{m,\perp}^{\mathsf{T}}\mathbf{W}^{\mathsf{T}}\mathbf{W}\mathbf{A}_{m,\perp} - \mathbf{A}_{m,\perp}^{\mathsf{T}}\mathbf{A}_{m,\perp}\right\|_F^2\right) \leq \frac{8}{s}\,\|\mathbf{A}\mathbf{X}\|_F^4\,. \quad (53)$$

Next, applying Markov's inequality and setting the number of sampled rows $s \geq {}^{32k}/_{\delta\varepsilon^2}$, we get

$$\mathbb{P}\left(\left|\|\mathbf{W}\mathbf{A}_{m,\perp}\|_F^2 - \|\mathbf{A}_{m,\perp}\|_F^2\right| \geq \frac{\varepsilon}{\sqrt{k}}\,\|\mathbf{A}\mathbf{X}\|_F^2\right) \leq \frac{\delta}{4}\,, \quad (54)$$

$$\mathbb{P}\left(\|\mathbf{A}_{m,\perp}^{\mathsf{T}}\mathbf{W}^{\mathsf{T}}\mathbf{W}\mathbf{A}_{m,\perp} - \mathbf{A}_{m,\perp}^{\mathsf{T}}\mathbf{A}_{m,\perp}\|_F \geq \frac{\varepsilon}{\sqrt{k}}\,\|\mathbf{A}\mathbf{X}\|_F^2\right) \leq \frac{\delta}{4}\,. \quad (55)$$

It is worth noting that the bound of eqn. (54) is *stronger* (by a factor of ${}^1/\sqrt{k}$) compared to what is needed in Theorem 2. This improved bound comes for free given the value of s used in the construction of $\mathbf{W}$ and does not affect the tightness of the overall bound.

Finally, applying the union bound to eqns. (46), (48), (54), and (55), we observe that if the number of sampled rows

$$s \geq \max\left\{\left(1+\frac{\varepsilon}{3}\right)\frac{4\,k\,\ln\left({}^{16\,(1+2k)}/_{\delta}\right)}{\varepsilon^2}\,,\,\frac{32k}{\delta\,\varepsilon^2}\right\},$$

then all four structural conditions of Theorem 2 hold with probability at least $1-\delta$. Therefore, the number of sampled rows s is, asymptotically (assuming that δ is constant), $s = \mathcal{O}\left({}^{k\ln k}/_{\varepsilon^2}\right)$.

5. Other constructions for the sketching matrix $\mathbf{W}$

In this section, we briefly point to various constructions for the sketching matrix $\mathbf{W} \in \mathbb{R}^{s\times n}$. We are particularly interested in the case where the rows of $\mathbf{W}$ are pairwise orthogonal and normal, *i.e.*, $\mathbf{W}\mathbf{W}^{\mathsf{T}} = \mathbf{I}_s$, and thus $\mathbf{W}^{\mathsf{T}}\mathbf{W}$ is an orthogonal projector.

Several random projection–based constructions for the sketching matrix $\mathbf{W}$ were summarized in [8], including the (relatively) sparse matrix of [1], the (very) sparse embedding matrix of [7], and the *oblivious sparse norm-approximating projections* (OSNAP) of [24]. In all three cases, the rows of $\mathbf{W}$ are not exactly orthonormal, but rather close to being orthonormal, *i.e.*, $\mathbf{W}\mathbf{W}^\mathsf{T}$ is approximately equal to the identity matrix. At the same time, [8] also considered cases where the rows of $\mathbf{W}$ are exactly orthonormal. One such case is a construction of $\mathbf{W}$ using a randomized SVD of $\mathbf{A}$ (see also [25,19]). In this case, the rows of $\mathbf{W}$ are approximations to the top s left singular vectors of $\mathbf{A}$ (see Theorem 8 of [8] for details). Another case is the so-called *non-oblivious random projection* of [8]. In this case, the rows of $\mathbf{W}$ form an orthonormal basis for the range of $\mathbf{AR}$, where $\mathbf{R}$ is a Johnson–Lindenstrauss random projection matrix (see Theorem 16 of [8] for further details).

Our next result is a special case of Theorem 2 showing that if the sketching matrix $\mathbf{W}$ has orthonormal rows then only two of our structural conditions (eqns. (5) and (8)) suffice to ensure the projection-cost preservation guarantee.

Lemma 6. *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ be the input matrix and let $\mathbf{X} \in \mathbb{R}^{d\times(d-k)}$ be any matrix satisfying $\mathbf{X}^\mathsf{T}\mathbf{X} = \mathbf{I}$, with $1 \le k < d$. Let the thin SVD of $\mathbf{A}$ be $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}$. Recall the definition of $\widetilde{\mathbf{\Sigma}}$ from eqn. (4), and assume that the sketching matrix $\mathbf{W} \in \mathbb{R}^{s\times n}$ satisfies $\mathbf{W}\mathbf{W}^\mathsf{T} = \mathbf{I}_s$, as well as the following two conditions (for some accuracy parameter ε):*

$$\left\|\widetilde{\mathbf{\Sigma}}\mathbf{U}^\mathsf{T}\mathbf{W}^\mathsf{T}\mathbf{W}\mathbf{U}\widetilde{\mathbf{\Sigma}} - \widetilde{\mathbf{\Sigma}}^2\right\|_2 \le \varepsilon, \quad \textit{and} \tag{56}$$

$$\left|\|\mathbf{W}\mathbf{A}_{m,\perp}\|_F^2 - \|\mathbf{A}_{m,\perp}\|_F^2\right| \le \varepsilon\,\|\mathbf{A}\mathbf{X}\|_F^2\,. \tag{57}$$

Then,

$$\left|\|\mathbf{W}\mathbf{A}\mathbf{X}\|_F^2 - \|\mathbf{A}\mathbf{X}\|_F^2\right| \le \left(d_m^{-2} + 2\,d_m^{-1} + 1\right)\varepsilon\,\|\mathbf{A}\mathbf{X}\|_F^2\,. \tag{58}$$

Proof. Again, for notational convenience, let $\mathbf{Z} \triangleq \mathbf{U}^\mathsf{T}(\mathbf{I}_n - \mathbf{W}^\mathsf{T}\mathbf{W})\mathbf{U}$ and $\mathbf{Y} \triangleq \mathbf{V}^\mathsf{T}\mathbf{X}$. If $\mathbf{W}$ has orthonormal rows, then $\mathbf{W}^\mathsf{T}\mathbf{W}$ is an orthogonal projector, *i.e.*, it is idempotent and symmetric positive semi-definite (SPSD), with all its singular values equal to zero and one. Hence, $\mathbf{Z}$ is also SPSD and its square root $\mathbf{Z}^{1/2}$ is well-defined. Using the definitions of $\mathbf{Z}$ and $\mathbf{Y}$, eqns. (56) and (57) become:

$$\left\|\mathbf{Z}^{1/2}\widetilde{\mathbf{\Sigma}}\right\|_2 \le \sqrt{\varepsilon}, \quad \text{and} \tag{59}$$

$$\left\|\mathbf{Z}^{1/2}\mathbf{\Sigma}_{m,\perp}\right\|_F \le \sqrt{\varepsilon}\left\|\mathbf{\Sigma}\mathbf{Y}\right\|_F. \tag{60}$$

We rewrite the left hand side of eqn. (58) as

$$\begin{aligned}\left|\|\mathbf{A}\mathbf{X}\|_F^2 - \mathbf{W}\mathbf{A}\mathbf{X}\|_F^2\right| &= \left|\mathrm{tr}\left(\mathbf{X}^\mathsf{T}\mathbf{A}^\mathsf{T}\left(\mathbf{I}_n - \mathbf{W}^\mathsf{T}\mathbf{W}\right)\mathbf{A}\mathbf{X}\right)\right| \\ &= \left|\mathrm{tr}\left(\mathbf{X}^\mathsf{T}\mathbf{V}\mathbf{\Sigma}\mathbf{U}^\mathsf{T}\left(\mathbf{I}_n - \mathbf{W}^\mathsf{T}\mathbf{W}\right)\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}\mathbf{X}\right)\right|\end{aligned}$$

$$= \left|\mathrm{tr}\left(\mathbf{Y}^{\mathsf{T}}\mathbf{\Sigma}\mathbf{Z}\mathbf{\Sigma}\mathbf{Y}\right)\right| = \left\|\mathbf{Z}^{1/2}\mathbf{\Sigma}\mathbf{Y}\right\|_F^2.$$

Thus,

$$\begin{aligned}\sqrt{\left|\|\mathbf{A}\mathbf{X}\|_F^2 - \|\mathbf{W}\mathbf{A}\mathbf{X}\|_F^2\right|} &= \left\|\mathbf{Z}^{1/2}\mathbf{\Sigma}\mathbf{Y}\right\|_F = \left\|\mathbf{Z}^{1/2}(\mathbf{\Sigma}_m + \mathbf{\Sigma}_{m,\perp})\mathbf{Y}\right\|_F \\ &\leq \left\|\mathbf{Z}^{1/2}\mathbf{\Sigma}_m\mathbf{Y}\right\|_F + \left\|\mathbf{Z}^{1/2}\mathbf{\Sigma}_{m,\perp}\mathbf{Y}\right\|_F. \end{aligned} \tag{61}$$

Using norm invariance properties and the definition of $\mathbf{Y}$, we get

$$\|\mathbf{A}\mathbf{X}\|_F^2 = \|\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathsf{T}}\mathbf{X}\|_F^2 = \|\mathbf{\Sigma}\mathbf{Y}\|_F^2. \tag{62}$$

Thus, it suffices to prove that

$$\sqrt{\left|\|\mathbf{A}\mathbf{X}\|_F^2 - \|\mathbf{W}\mathbf{A}\mathbf{X}\|_F^2\right|} \leq \sqrt{\varepsilon}\,(1 + d_m^{-1})\,\|\mathbf{\Sigma}\mathbf{Y}\|_F. \tag{63}$$

Similarly to the proof of Theorem 2, using the commutativity property of diagonal matrices and the fact that $\left(\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp\right)\mathbf{\Sigma}_m = \widetilde{\mathbf{\Sigma}}\,\mathbf{\Sigma}_m$, we have

$$\begin{aligned}\mathbf{\Sigma}_m &= (\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp)^{-1}(\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp)\mathbf{\Sigma}_m = (\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp)^{-1}\widetilde{\mathbf{\Sigma}}\mathbf{\Sigma}_m \\ &= \widetilde{\mathbf{\Sigma}}(\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp)^{-1}\mathbf{\Sigma}_m = \widetilde{\mathbf{\Sigma}}(\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp)_m^{-1}\mathbf{\Sigma}_m. \end{aligned} \tag{64}$$

Notice that $\left(\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp\right)^{-1}\mathbf{\Sigma}_m = (\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp)_m^{-1}\mathbf{\Sigma}_m$, where $(\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp)_m^{-1} \in \mathbb{R}^{r\times r}$ is a diagonal matrix whose top m diagonal entries are equal to those of $\left(\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp\right)^{-1}$ and the remaining $r - m$ diagonal entries are set to zero. In order to bound the first term in eqn. (61), we use eqn. (64) as follows:

$$\begin{aligned}\left\|\mathbf{Z}^{1/2}\mathbf{\Sigma}_m\mathbf{Y}\right\|_F &= \left\|\mathbf{Z}^{1/2}\widetilde{\mathbf{\Sigma}}(\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp)_m^{-1}\mathbf{\Sigma}_m\mathbf{Y}\right\|_F \leq \left\|\mathbf{Z}^{1/2}\widetilde{\mathbf{\Sigma}}\right\|_2 \cdot \left\|(\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp)_m^{-1}\right\|_2 \cdot \|\mathbf{\Sigma}_m\mathbf{Y}\|_F \\ &\leq \sqrt{\varepsilon}\,d_m^{-1}\,\|\mathbf{\Sigma}_m\mathbf{Y}\|_F = \sqrt{\varepsilon}\,d_m^{-1}\,\|\mathbf{A}_m\mathbf{X}\|_F \leq \sqrt{\varepsilon}\,d_m^{-1}\,\|\mathbf{A}\mathbf{X}\| \\ &= \sqrt{\varepsilon}\,d_m^{-1}\,\|\mathbf{\Sigma}\mathbf{Y}\|_F. \end{aligned} \tag{65}$$

The first inequality is due to strong submultiplicativity; the second inequality follows from eqn. (59) and the fact that $\left\|(\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{\Sigma}}_\perp)_m^{-1}\right\|_2 = d_m^{-1}$; and the last inequality uses $\|\mathbf{A}\mathbf{X}\|_F^2 = \|\mathbf{A}_m\mathbf{X}\|_F^2 + \|\mathbf{A}_{m,\perp}\mathbf{X}\|_F^2$ (by the matrix Pythagorean theorem). Next, using strong submultiplicativity and eqn. (60), we bound the second term of eqn. (61):

$$\left\|\mathbf{Z}^{1/2}\mathbf{\Sigma}_{m,\perp}\mathbf{Y}\right\|_F \leq \left\|\mathbf{Z}^{1/2}\mathbf{\Sigma}_{m,\perp}\right\|_F \|\mathbf{Y}\|_2 \leq \sqrt{\varepsilon}\,\|\mathbf{\Sigma}\mathbf{Y}\|_F, \tag{66}$$

where we used $\|\mathbf{Y}\|_2 = \|\mathbf{V}^{\mathsf{T}}\mathbf{X}\|_2 \leq 1$. Finally, we combine eqns. (61), (65), and (66) to get

$$\sqrt{|\|\mathbf{AX}\|_F^2 - \|\mathbf{WAX}\|_F^2|} \leq \sqrt{\varepsilon}\,(1 + d_m^{-1})\,\|\mathbf{\Sigma Y}\|_F\,,$$

which concludes the proof. □

6. Conclusion

Building upon the definition of cost-preserving projections, we have presented a simple structural result connecting the construction of projection-cost preserving sketches to sketching-based matrix multiplication. Our work unifies and generalizes prior known constructions for projection-cost preserving sketches based on (variants of) the standard leverage scores, ridge leverage scores, as well as other constructions.

An interesting open problem would be to understand whether similar structural results for cost-preserving projections can be derived for other Schatten p-norms, *e.g.*, for the Schatten infinity norm, which corresponds to the well-known matrix two-norm. Preliminary work in this direction includes Lemma 26 in [8]; to the best of our knowledge, other Schatten p-norms have not been studied in prior work. Additionally, it would be interesting to study alternative sets of structural conditions that guarantee projection-cost preservation, with the end goal of fully characterizing the problem by presenting both necessary and sufficient conditions for various Schatten p-norms.

Acknowledgements

We would like to thank Ilse Ipsen for many useful comments and, in particular, for deriving the proof of Lemma 6 in Section 5. AC and PD were supported by NSF IIS-1661760, NSF IIS-1661756, NSF CCF-1814041, and NSF DMS-1760353. JY was supported by NSF IIS-1149789 and NSF IIS-1618690.

References

[1] Dimitris Achlioptas, Database-friendly random projections: Johnson-Lindenstrauss with binary coins, J. Comput. System Sci. 66 (4) (2003) 671–687.
[2] Ahmed Alaoui, Michael W. Mahoney, Fast randomized kernel ridge regression with statistical guarantees, in: Advances in Neural Information Processing Systems 28, 2015, pp. 775–783.
[3] Daniel Aloise, Amit Deshpande, Pierre Hansen, Preyas Popat, NP-hardness of euclidean sum-of-squares clustering, Mach. Learn. 75 (2) (2009) 245–248.
[4] Olivier Bachem, Mario Lucic, Andreas Krause, Practical coreset constructions for machine learning, arXiv:1703.06476, 2017.
[5] C. Boutsidis, A. Zouzias, M.W. Mahoney, P. Drineas, Randomized dimensionality reduction for k-means clustering, IEEE Trans. Inform. Theory 61 (2) (2015) 1045–1062.
[6] Agniva Chowdhury, Jiasen Yang, Petros Drineas, An iterative, sketching-based framework for ridge regression, in: Proceedings of the 35th International Conference on Machine Learning, 2018, pp. 988–997.
[7] Kenneth L. Clarkson, David P. Woodruff, Low-rank approximation and regression in input sparsity time, J. ACM 63 (6) (2017) 54.
[8] Michael B. Cohen, Sam Elder, Cameron Musco, Christopher Musco, Madalina Persu, Dimensionality reduction for k-means clustering and low rank approximation, in: Proceedings of the 47th Annual ACM Symposium on Theory of Computing, 2015, pp. 163–172.

[9] Michael B. Cohen, Cameron Musco, Christopher Musco, Input sparsity time low-rank approximation via ridge leverage score sampling, in: Proceedings of the 28th Annual ACM-SIAM Symposium on Discrete Algorithms, 2017, pp. 1758–1777.
[10] P. Drineas, A. Frieze, R. Kannan, S. Vempala, V. Vinay, Clustering large graphs via the singular value decomposition, Mach. Learn. 56 (1–3) (2004) 9–33.
[11] Petros Drineas, Ravi Kannan, Michael W. Mahoney, Fast Monte Carlo algorithms for matrices I: Approximating matrix multiplication, SIAM J. Comput. 36 (1) (2006) 132–157.
[12] Petros Drineas, Michael W. Mahoney, RandNLA: Randomized numerical linear algebra, Commun. ACM 59 (6) (2016) 80–90.
[13] Petros Drineas, Michael W. Mahoney, Structural properties underlying high-quality randomized numerical linear algebra algorithms, in: Handbook of Big Data, Chapman & Hall/CRC Press, 2016, pp. 137–154.
[14] Petros Drineas, Michael W. Mahoney, Lectures on randomized numerical linear algebra, in: The Mathematics of Data, in: IAS/Park City Mathematics Series, vol. 25, 2018, pp. 1–45.
[15] Petros Drineas, Michael W. Mahoney, S. Muthukrishnan, Tamás Sarlós, Faster least squares approximation, Numer. Math. 117 (2011) 219–249.
[16] Dan Feldman, Melanie Schmidt, Christian Sohler, Turning big data into tiny data: constant-size coresets for k-means, PCA and projective clustering, in: Proceedings of the 24th Annual ACM-SIAM Symposium on Discrete Algorithms, 2013, pp. 1434–1453.
[17] Dan Feldman, Tamir Tassa, More constraints, smaller coresets: constrained matrix approximation of sparse big data, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015, pp. 249–258.
[18] Dan Feldman, Mikhail Volkov, Daniela Rus, Dimensionality reduction of massive sparse datasets using coresets, in: Advances in Neural Information Processing Systems 29, 2016.
[19] Nathan Halko, Per-Gunnar Martinsson, Joel A. Tropp, Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions, SIAM Rev. 53 (2) (2011) 217–288.
[20] John T. Holodnak, Ilse C.F. Ipsen, Randomized approximation of the Gram matrix: exact computation and probabilistic bounds, SIAM J. Matrix Anal. Appl. 36 (1) (2015) 110–137.
[21] Michael W. Mahoney, Petros Drineas, CUR matrix decompositions for improved data analysis, Proc. Natl. Acad. Sci. USA 106 (3) (2009).
[22] Leon Mirsky, A trace inequality of John von Neumann, Monatsh. Math. 79 (4) (1975) 303–306.
[23] Cameron N. Musco, Dimensionality Reduction for k-Means Clustering, M.Sc. Thesis, MIT, 2015.
[24] Jelani Nelson, Huy L. Nguyên, Osnap: faster numerical linear algebra algorithms via sparser subspace embeddings, in: Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science, IEEE, 2013, pp. 117–126.
[25] Tamás Sarlós, Improved approximation algorithms for large matrices via random projections, in: Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science, IEEE, 2006, pp. 143–152.