Reduced-Rank Regression with Operator Norm Error

Praneeth Kacham PKACHAM@CS.CMU.EDU and **David P. Woodruff** DWOODRUF@CS.CMU.EDU
Computer Science Department
Carnegie Mellon University

Abstract

A common data analysis task is the reduced-rank regression problem:

$$\min_{\text{rank-}k\ X} \|AX - B\|,$$

where $A \in \mathbb{R}^{n \times c}$ and $B \in \mathbb{R}^{n \times d}$ are given large matrices and $\|\cdot\|$ is some norm. Here the unknown matrix $X \in \mathbb{R}^{c \times d}$ is constrained to be of rank k as it results in a significant parameter reduction of the solution when c and d are large. In the case of Frobenius norm error, there is a standard closed form solution to this problem and a fast algorithm to find a $(1 + \varepsilon)$-approximate solution. However, for the important case of operator norm error, no closed form solution is known and the fastest known algorithms take singular value decomposition time.

We give the first randomized algorithms for this problem running in time

$$(\mathsf{nnz}(A) + \mathsf{nnz}(B) + c^2) \cdot k/\varepsilon^{1.5} + (n + d)k^2/\varepsilon + c^\omega,$$

up to a polylogarithmic factor involving condition numbers, matrix dimensions, and dependence on $1/\varepsilon$. Here $\mathsf{nnz}(M)$ denotes the number of nonzero entries of a matrix M, and ω is the exponent of matrix multiplication. As both (1) spectral low rank approximation ($A = B$) and (2) linear system solving ($n = c$ and $d = 1$) are special cases, our time cannot be improved by more than a $1/\varepsilon$ factor (up to polylogarithmic factors) without a major breakthrough in linear algebra. Interestingly, known techniques for low rank approximation, such as alternating minimization or sketch-and-solve, provably fail for this problem. Instead, our algorithm uses an existential characterization of a solution, together with Krylov methods, low degree polynomial approximation, and sketching-based preconditioning.

1. Introduction

Given an $n \times c$ matrix A, an $n \times d$ matrix B, and an integer parameter k, the reduced-rank regression problem asks to solve for a rank at most k matrix $X \in \mathbb{R}^{c \times d}$ for which $\|AX - B\|$ is minimized in some norm. A standard motivation is that by constraining X to have rank at most k, the solution X can be represented using only $(c + d)k$ parameters rather than $c \cdot d$ parameters. Another important motivation is that the rank constraint provides regularization on the solution, which often leads to better generalization. Yet another motivation is that the solution X can be explained by at most k latent factors, and one can try to interpret the latent factors, plot them [3], and so on. This is commonly done in ecology, where reduced-rank regression is known as redundancy analysis [15], and is a type of ordination method [14]. For a survey, we refer the reader to the textbook by Velu and Reinsel [29] devoted to reduced-rank regression.

The $\min_{\text{rank-}k\ X} \|AX - B\|$ problem is only known to have a closed form solution when the error measure is the Frobenius norm. In this case, the solution is given by $X = A^+[AA^+B]_k$ (see, e.g., [7]). Here for a matrix M, $[M]_k$ denotes the best rank k approximation for M in Frobenius norm

and M^+ denotes the Moore-Penrose pseudo-inverse. This has a natural geometric interpretation - project each of the columns of B onto the column span of A and find the best rank-k approximation to the projected matrix. By the Pythagorean theorem, one can show there is no loss in this approach, as the optimal cost decomposes into the sum of squared distances of columns of B to the column span of A followed by the best rank-k approximation to the projected matrix inside of the column span of A.

In a number of applications, the Frobenius norm is not the right measure. For example, in cancer genetics more robust versions are desired, and versions based on the sum of Euclidean lengths instead of the sum of squared Euclidean lengths are sometimes used [25]. Still, in other applications, the operator norm error solution may give a solution of much better quality. Indeed, if B has a heavy tail of singular values, as is common for data analysis and learning applications, then it has no good rank-k approximation, much less one in the column span of A, and consequently, outputting an X' with $\|AX' - B\|_{\mathsf{F}}^2 \le (1+\varepsilon)\|AX_{\mathsf{F}} - B\|_{\mathsf{F}}^2$, where X_{F} is the optimal Frobenius norm solution, may be meaningless as one could just set $X' = 0$. Indeed, this is sometimes a motivation (see, e.g., [18]) for the low rank approximation problem with operator norm error, which is a special case of our problem when $A = B$, and a number of works [9; 12; 13; 28] suggest considering operator norm error in certain contexts.

It is tempting to think that the Frobenius norm solution holds also for other unitarily invariant norms, such as the operator norm. However, one can show this is not the case. Indeed, let X_{F} be the solution to $\min_{\text{rank-}k\ X}\|AX - B\|_{\mathsf{F}}$. It was shown by Boutsidis [2] that this is a $\sqrt{2}$-approximation, namely, that $\|AX_{\mathsf{F}} - B\|_2 \le \sqrt{2} \cdot \text{Opt}$ where $\text{Opt} = \min_{\text{rank-}k\ X}\|AX - B\|_2$. Unfortunately, the $\sqrt{2}$ factor is tight and there are instances where the Frobenius norm solution really does give at best a $\sqrt{2}$-approximation. Suppose, for example[1]

$$A = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \text{ and } B = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1+\gamma \end{bmatrix}.$$

For the problem $\min_{\text{rank-1}\ X}\|AX - B\|_{\mathsf{F}}$, the optimum solution is

$$X_{\mathsf{F}} = \begin{bmatrix} 0 & 0 \\ 0 & 1+\gamma \end{bmatrix}, \text{ with } AX_{\mathsf{F}} - B = -\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 0 \end{bmatrix}$$

and thus, $\|AX_{\mathsf{F}} - B\|_2 = \sqrt{2}$. On the other hand, for

$$X = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \ AX - B = -\begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1+\gamma \end{bmatrix},$$

and so $\|AX - B\|_2 = (1+\gamma)$. As $\gamma \to 0$, the approximation factor becomes arbitrarily close to $\sqrt{2}$.

We note that the reduced-rank regression problem in operator norm is non-convex in X due to the rank constraint, and it is not even clear this problem can be solved in polynomial time. Of the few techniques that are known for rank-constrained optimization, they do not apply here. One

---

1. We thank Ankur Moitra for pointing out this example to us.

common method is alternating minimization, writing the problem above as $\min_{U,V} \|AUV - B\|_2$, where $U \in \mathbb{R}^{n \times k}$ and $V \in \mathbb{R}^{k \times d}$. The idea is to fix U, then solve for V, then fix V and solve for U, and repeat. When U is fixed, then $V = (AU)^+B$ is the optimum, and when V is fixed, the solution turns out to be $U = A^+BV^+$, though this is not as obvious, see (1.3) in [16], taking $p \to \infty$, for a proof. It turns out if one initializes with the Frobenius norm solution U, V, then each of these operations does not change U or V, and so by the example above, alternating minimization gives at best a $\sqrt{2}$-approximation. Other techniques include sketching to a small problem, and solving the small problem in the sketch space; sketches are well-known not to apply to operator norm low rank approximation problems, motivating the first open question in [31].

This issue of polynomial time solvability was raised in the control theory literature by Sou and Rantzer [27], where a $(1 + \varepsilon)$-approximation was obtained, but the the time required to find the solution was at least the time to perform a singular value decomposition (SVD) on matrices A and B, which is prohibitive for large n, c, and d. This is a common setting of parameters and indeed, one of the motivations for constraining X to have rank at most k in the first place. This motivates the question:

> "*Are there fast algorithms for reduced-rank regression with operator norm error?*"

1.1. Main Result

We answer the question above by designing a new randomized algorithm running in time

$$O\left(\left(\frac{\mathsf{nnz}(B) \cdot k}{\varepsilon} + \frac{\mathsf{nnz}(A) \cdot k}{\varepsilon^{1.5}} + \frac{c^2 k}{\varepsilon^{1.5}} + \frac{(n+d)k^2}{\varepsilon}\right) \cdot \mathrm{polylog}(\kappa(B), n, d, k, 1/\varepsilon) + c^\omega\right).$$

Here, $\kappa(B)$ denotes $\sigma_1(B)/\sigma_{k+1}(B)$. This significantly improves over Sou and Rantzer's polynomial time result, which takes $\Omega(nd^2 + nc^2)$ time.

We note that spectral low rank approximation is a special case in which $A = B$, and the best known upper bound is $O(\mathsf{nnz}(A) \cdot k/\sqrt{\varepsilon})$ for this problem, up to logarithmic factors [18]. A major open question in randomized numerical linear algebra is to improve this bound (see, e.g., Open Question 1 of [31]), or show that it is not possible. We note that for $k = 1$, in the matrix-vector query model, $\Omega(1/\sqrt{\varepsilon})$ queries is known to be required if a slightly stronger guarantee than spectral low rank approximation is desired, even for adaptive algorithms [4; 26]. Another important point is that when $n = c$ and $d = 1$, this is just the time to solve an arbitrary linear system, for which the best known time is c^ω. Improving either spectral low rank approximation or linear system solving is a major open question, and barring that, our algorithm is optimal up to a $1/\varepsilon$ factor and polylogarithmic factors involving matrix dimensions and condition numbers.

1.2. Our Techniques

Throughout the paper, let $\mathrm{Opt} := \inf_{\text{rank-}k \ X} \|AX - B\|_2$, β be such that $(1 + \varepsilon)\mathrm{Opt} \leq \beta \leq (1 + 2\varepsilon)\mathrm{Opt}$, and let $\Delta := B^\mathsf{T}(I - AA^+)B$. The work of Sou and Rantzer [27] shows that $X_\beta = A^+[AA^+B(\beta^2 I - \Delta)^{-1/2}]_k(\beta^2 I - \Delta)^{1/2}$ satisfies $\|AX - B\|_2 < \beta$. For completeness, we give a short proof of this fact in this paper. It is not a priori clear how to extract a solution from this expression, while multiplying out all of the matrices, computing an inverse square root, and taking an SVD would take a prohibitive amount of time. This is essentially the algorithm of [27].

We instead show that not only the best rank k approximation of the matrix $AA^+B(\beta^2 I - \Delta)^{-1/2}$, but even a $1 + \varepsilon$ approximation in spectral norm yields an overall solution of cost at most $\beta(1 + O(\varepsilon))$. To obtain such a $1 + \varepsilon$ approximation, we next try to apply the iterative method of [18] which computes the Krylov matrix $K = [C \cdot G, (CC^{\mathsf{T}}) \cdot C \cdot G, (CC^{\mathsf{T}})^2 \cdot C \cdot G, \ldots, (CC^{\mathsf{T}})^{(q-1)/2} \cdot C \cdot G]$ where G is a Gaussian matrix with k columns, $q = O(\log(d/\varepsilon)\sqrt{1/\varepsilon})$ is an odd integer, and $C = AA^+B(\beta^2 I - \Delta)^{-1/2}$. The first problem with this approach is that we have to compute the matrix vector product CG and to do this, in each iteration we need to (1) multiply by the square root of an inverse (multiplication by $(\beta^2 I - \Delta)^{-1/2}$), and then (2) project onto the column span of A (multiplication by AA^+).

Computing exact matrix-vector products with the matrices AA^+ and $(\beta^2 I - \Delta)^{-1/2}$, is slow when c, d are large, and finding the matrices AA^+ and $(\beta^2 I - \Delta)^{-1/2}$ takes at least $\Omega(nc^2 + \mathsf{nnz}(B) \cdot c + d^\omega)$ time. To avoid such a running time, we show that the Block Krylov Iteration algorithm of Musco and Musco [18] works even with approximate matrix-vector products i.e., we only need algorithms to compute vectors $C \circ v$ and $C^{\mathsf{T}} \circ v'$ for arbitrary vectors v, v' such that $\|C \circ v - Cv\|_2$ and $\|C^{\mathsf{T}} \circ v' - C^{\mathsf{T}} v'\|_2$ are small. Here and throughout the paper, we use the notation $M \circ v$ to denote an approximation to the matrix-vector product Mv.

An important idea of Musco and Musco [18] is that the Krylov matrix K spans a rank k matrix $p(C)G = \sum_{\text{odd } i \leq q} p_i (CC^{\mathsf{T}})^{(i-1)/2} G$, where p is a polynomial, such that projecting the columns of the matrix C onto the column span of $p(C)G$ gives a good rank k approximation. To prove that the algorithm works even with approximate matrix-vector products, we first show that the approximations computed to matrices $(CC^{\mathsf{T}})^{(i-1)/2}CG$ for $i = 1, \ldots, q$ are good enough to imply that the approximate Krylov matrix K' spans a matrix Apx that is close to the matrix $p(C)G$ in Frobenius norm. To then conclude that the column space of Apx is also a good subspace to project the matrix C onto, we need to show that $(Apx)(Apx)^+ \approx (p(C)G)(p(C)G)^+$. We prove a simple lemma that shows if $\|p(C)G - Apx\|_{\mathsf{F}}$ is small, and $p(C)G$ has a good condition number, and so then $\|(p(C)G)(p(C)G)^+ - (Apx)(Apx)^+\|_2$ is small. Crucially, as G is a Gaussian matrix that has, with good probability a good condition number, we only have to bound $\sigma_1(p(C))/\sigma_k(p(C))$ to obtain a bound on the condition number of $p(C)G$. Using several properties of Chebyshev polynomials used to define the polynomial $p(x)$, we show that $\sigma_1(p(C))/\sigma_k(p(C))$ can be bounded in terms of $\kappa = \sigma_1(C)/\sigma_{k+1}(C)$, which finally shows that the k-dimensional column span of Apx is also a good subspace to project the columns of C.

As the parameters of the polynomial $p(x)$ are unknown, we cannot actually compute the matrix Apx and then project C onto the column span. But using the fact that K' spans Apx, we can conclude, similarly to the arguments of [18], that the best rank k Frobenius norm approximation of C in the span of K' is a good rank k approximation to C. Using the oracle to compute approximate matrix-vector products with the matrix C, we recover a $1 + \varepsilon$ approximation to the best rank k Frobenius norm approximation of C inside the span of K', which we then show is a $1 + \varepsilon$ approximation to a spectral norm low rank approximation of matrix C. Our analysis that the Block Krylov Iteration algorithm works with approximate matrix-vector products could help justify why the Block Krylov Iteration algorithm works well when using finite precision arithmetic rather than exact arithmetic. Our results address the comments of [19] about the stability of block Lanczos based methods for problems such as low rank approximation. Though several analyses of the noisy power method have been done previously [1; 10; 11], where each intermediate computation is corrupted by Gaussian noise, we are not aware of an analysis that works for worst case corruption. Also, previous work

bounds the amount of Gaussian noise that can be added in terms of a gap between σ_k and σ_{k+1}, which can be 0, and would not work for our analysis.

We return to the task at hand, i.e., of computing a low rank approximation of $AA^+B(\beta^2 I - \Delta)^{-1/2}$. We show that we can replace the matrix $(\beta^2 I - \Delta)^{-1/2}$ with the matrix $(1/\beta)r(\Delta/\beta^2)$, where $r(x)$ is a polynomial of degree $\widetilde{O}(1/\sqrt{\varepsilon})$, using polynomial approximation techniques based on Chebyshev polynomials (see, e.g., [24] and the references therein). Here we crucially use the fact that $(1 + 2\varepsilon)\mathsf{Opt} \geq \beta \geq (1 + \varepsilon)\mathsf{Opt} \geq (1 + \varepsilon)\|(I - AA^+)B\|_2$ to lower bound the minimum singular value of the matrix $(\beta^2 I - \Delta)$, thereby obtaining an upper bound on the number of terms required to approximate $(I - (\Delta/\beta^2))^{-1/2}$ with a Taylor series. Then we replace each monomial in the Taylor series with a low degree polynomial approximation to construct a polynomial $r(x)$. The replacement of $(\beta^2 I - \Delta)^{-1/2}$ with the matrix $r(\Delta/\beta^2)$ is done as we can give very fast algorithms to approximately multiply a vector with the matrix $r(\Delta/\beta^2)$, as discussed below.

Let $\mathcal{M}' = AA^+B \cdot r(\Delta/\beta^2)$. Recall $\Delta = B^\mathsf{T}(I - AA^+)B$. To approximate the matrix-vector product Δu for an arbitrary vector u, we need only approximate $B^\mathsf{T}AA^+Bu$, since $B^\mathsf{T}Bu$ can be computed exactly in $\mathsf{nnz}(B)$ time. For computing an approximation to $AA^+(Bu)$, we use fast sketching-based preconditioning methods for linear regression, which show given an arbitrary vector b and accuracy parameter $\varepsilon_{\mathrm{reg}}$ how to find an x for which $\|Ax - AA^+b\|_2 \leq \varepsilon_{\mathrm{reg}}\|(I - AA^+)b\|_2$ in time $O((\mathsf{nnz}(A) + c^2)\log(1/\varepsilon_{\mathrm{reg}}) + c^\omega)$, where $\omega \approx 2.376$ is the exponent of matrix multiplication [6; 17; 21]. We note that we only need to pay the c^ω time once to compute a preconditioner, after which each regression problem takes $O((\mathsf{nnz}(A) + c^2)\log(1/\varepsilon_{\mathrm{reg}}))$ time. This algorithm to approximately compute Δu for an arbitrary vector u is extended to approximate $r(\Delta/\beta^2) \cdot v$ for an arbitrary v. After approximating the product $r(\Delta/\beta^2) \cdot v$ with a vector y, we approximate the vector AA^+By again using the sketching-based preconditioning methods for linear regression.

Similarly we also give an algorithm to approximate $\mathcal{M}'^\mathsf{T}v'$ for an arbitrary vector v'. Thus, as discussed above, we can obtain using a Block Krylov algorithm, a matrix Z with orthonormal columns for which $\|ZZ^\mathsf{T}\mathcal{M}' - \mathcal{M}'\|_2 \leq (1 + \varepsilon)\sigma_{k+1}(\mathcal{M}')$ and then conclude that

$$\|AA^+Z(AA^+Z)^+B - B\|_2 \leq (1 + O(\varepsilon))\beta = (1 + O(\varepsilon))\mathsf{Opt}$$

and that the rank k matrix $X = A^+Z(AA^+Z)^+B$ is a $1 + O(\varepsilon)$ approximation for the problem $\min_{\mathrm{rank}\text{-}k}\|AX - B\|_2$.

The time complexity of our algorithm depends logarithmically on $\kappa(B) = \sigma_1(B)/\sigma_{k+1}(B)$ and $\kappa(AA^+B) = \sigma_1(AA^+B)/\sigma_{k+1}(AA^+B)$. We show that if $\widetilde{B} = B + \alpha GF^\mathsf{T}$ where G is an $n \times (k + 1)$ random Gaussian matrix and F^T has $k + 1$ orthonormal rows, then for a suitable value of α, the condition number $\kappa(AA^+\widetilde{B}) \leq (Cn/\varepsilon)\kappa(B)$ for a constant C. We also show that a $1 + \varepsilon$ approximation for reduced rank regression computed using the matrix $\widetilde{B}$ is a $1 + O(\varepsilon)$ approximation for reduced rank regression on matrix B, thus removing the dependence on $\kappa(AA^+B)$. Note that matrix-vector products with $\widetilde{B}$ can be computed in $\mathsf{nnz}(B) + (n + d)k$ time.

Our final dependence on ε in the running time is $1/\varepsilon^{3/2}$, ignoring polylogarithmic factors, where a factor of $1/\sqrt{\varepsilon}$ is from the number of iterations in the Block Krylov Iteration algorithm of Musco and Musco [18], a factor of $1/\sqrt{\varepsilon}$ is from the degree of the polynomial $r(x)$, which is used as a proxy for the matrix $(\beta^2 I - \Delta)^{-1/2}$ with a matrix $r(\Delta/\beta^2)$, and a factor of $1/\sqrt{\varepsilon}$ is due to the running time of high-precision regression methods based on the accuracy with which the approximate matrix products need to be computed.

2. Notation and Preliminaries

For a matrix M, $\mathsf{nnz}(M)$ denotes the number of nonzero entries in M. We refer to the Singular Value Decomposition (SVD) with only nonzero singular values as the "thin" SVD. Given an arbitrary matrix M, $\mathsf{colspan}(M)$ denotes the subspace spanned by the columns of M, and the matrix M^+ denotes the Moore-Penrose pseudo-inverse of matrix M. Given a subspace V, the matrix $\mathbb{P}_V$ denotes the projection onto the subspace V. Therefore $\mathbb{P}_V u = \arg\min_{v \in V} \|u - v\|_2$ for all vectors u. Given a matrix M, we use $\mathbb{P}_M$ to denote $\mathbb{P}_{\mathsf{colspan}(M)}$.

For a matrix M, the Frobenius norm $(\sum_{i,j} M_{i,j}^2)^{1/2}$ is denoted by $\|M\|_{\mathsf{F}}$ and the operator norm (or spectral norm) $\sup_x \|Mx\|_2/\|x\|_2$ is denoted by $\|M\|_2$. For a square matrix M, $\mathsf{tr}(M)$ denotes the sum of diagonal entries. For matrices M and M' of the same dimensions, $\langle M, M' \rangle$ denotes $\mathsf{tr}(M^\mathsf{T} M') = \sum_{i,j} M_{i,j} M'_{i,j}$. We use the following standard facts repeatedly throughout the paper: for any matrix M, (1) $\|M\|_2 \leq \|M\|_{\mathsf{F}}$, (2) $\|M\|_{\mathsf{F}} \leq \sqrt{\mathsf{rank}(M)}\|M\|_2$ and (3) $\mathbb{P}_M = MM^+$. For any matrices A, B and C, (i) $\mathsf{tr}(ABC) = \mathsf{tr}(BCA)$, (ii) $\|ABC\|_{\mathsf{F}} \leq \|A\|_2\|B\|_{\mathsf{F}}\|C\|_2$ and (iii) $\langle A, B \rangle \leq \|A\|_{\mathsf{F}}\|B\|_{\mathsf{F}}$.

For a symmetric matrix M, define $\mathsf{psd}(M)$ to be the closest positive semi-definite matrix to M in Frobenius norm. It can be shown that if $M = \sum_i \lambda_i v_i v_i^\mathsf{T}$, then $\mathsf{psd}(M) = \sum_{i:\lambda_i \geq 0} \lambda_i v_i v_i^\mathsf{T}$.

Weyl's Inequality. For matrices A and B, Weyl's inequality gives that $\sigma_{i+j-1}(A + B) \leq \sigma_i(A) + \sigma_j(B)$ for all i and j. In particular, if $\|A - B\|_2 \leq \varepsilon$, $|\sigma_i(A) - \sigma_i(B)| \leq \varepsilon$ for all i.

Polynomials and Matrices. Let $p(x) = \sum_{i=0}^{d} p_i x^i$ be a degree d polynomial. We define $\|p\|_1 := \sum_i |p_i|$ to be the sum of absolute values of the coefficients of the polynomial $p(x)$. Given $A \in \mathbb{R}^{n \times d}$, let $A = U\Sigma V^\mathsf{T}$ be the singular value decomposition of A with $\Sigma \in \mathbb{R}^{n \times d}$. Define $p(A) := Up(\Sigma)V^\mathsf{T}$ where $p(\Sigma)$ is the matrix with main diagonal entries $p(\sigma_1), \ldots, p(\sigma_d)$. It is easy to check that the singular values of $p(A)$ are equal to $|p(\sigma_1)|, \ldots, |p(\sigma_d)|$.

Singular Value Excess. Let $A \in \mathbb{R}^{n \times d}$ with $n \geq d$ be an arbitrary matrix. Let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d \geq 0$ be the singular values of matrix A. The Singular Value Excess of matrix A, denoted by $\mathsf{sve}(A)$, is defined as the number of singular values of matrix A that are greater than or equal to 1 i.e.,

$$\mathsf{sve}(A) = |\{i \in [d] \,|\, \sigma_i \geq 1\}|.$$

As eigenvalues of matrix $I - A^\mathsf{T} A$ are $1 - \sigma_1^2 \leq \cdots \leq 1 - \sigma_d^2$, $\mathsf{sve}(A)$ is equal to the number of non-positive eigenvalues of the matrix $I - A^\mathsf{T} A$. For any symmetric matrix M, let $k^-(M)$ denote the number of non-positive eigenvalues of the matrix M. For any matrix A, $\mathsf{sve}(A) = k^-(I - A^\mathsf{T} A)$.

Sketching Based Preconditioning for High-Precision Regression. Given a matrix $A \in \mathbb{R}^{n \times c}$ and a vector $b \in \mathbb{R}^n$, we use fast sketching based preconditioning methods given by the following theorem to obtain a $(1 + \varepsilon)$ approximation to the problem $\min_x \|Ax - b\|_2$. See [31] and references therein for more background.

Theorem 1 (High Precision Regression/Approximate Projections) *Given a matrix $A \in \mathbb{R}^{n \times c}$ and a vector $b \in \mathbb{R}^n$, we can compute a vector x in time $O((\mathsf{nnz}(A) + c^2)\log(1/\varepsilon) + c^\omega)$ that satisfies $\|Ax - b\|_2^2 \leq (1 + \varepsilon)\|AA^+ b - b\|_2^2$. By the Pythagorean theorem, the vector x obtained satisfies $\|AA^+ b - Ax\|_2^2 \leq \varepsilon\|AA^+ b - b\|_2^2$.*

We have to pay c^ω only once to compute a preconditioner. Thereafter, every regression problem can be solved in time $O((\mathsf{nnz}(A) + c^2)\log(1/\varepsilon))$. Throughout the paper, we use HIGHPRECISION-REGRESSION(A, b, ε) to denote the algorithm implied by Theorem 1. We extend the notation to

compute approximate projections of each of the columns of matrix B, instead of just a single vector b, onto the column space of A.

Low Rank Approximation(LRA). Let $A \in \mathbb{R}^{n \times c}$ and $A = U\Sigma V^\mathsf{T}$ be its "thin" Singular Value Decomposition, where $U^\mathsf{T}U = I$, $V^\mathsf{T}V = I$ and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_{\mathrm{rank}(A)})$ with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\mathrm{rank}(A)} > 0$. For any $k \leq \mathrm{rank}(A)$, we define $[A]_k := \sum_{i=1}^{k} \sigma_i U_{*i}(V^\mathsf{T})_{i*}$, where U_{*i} denotes the i-th column of matrix U and V_{i*}^T denotes the i-th row of matrix V^T. The matrix $[A]_k$ optimally solves the problems $\min_{\mathrm{rank}\text{-}k\ X}\|A - X\|_\mathsf{F}$ and $\min_{\mathrm{rank}\text{-}k\ X}\|A - X\|_2$. As computing $[A]_k$ exactly is expensive, we use the Block Krylov Iteration algorithm of [18] to obtain a matrix $Z \in \mathbb{R}^{n \times k}$ for which $ZZ^\mathsf{T}A$ is a good solution to the Frobenius norm and spectral norm low rank approximation problems.

Theorem 2 (Musco and Musco [18]) *Given a matrix $A \in \mathbb{R}^{n \times d}$ such that the products $Av \in \mathbb{R}^n$ and $A^\mathsf{T}v' \in \mathbb{R}^d$ can be computed in time T for any vectors $v \in \mathbb{R}^d$ and $v' \in \mathbb{R}^n$, the Block Krylov Iteration algorithm runs in time*

$$O\left(T\frac{k \log d}{\varepsilon^{1/2}} + \frac{nk^2 \log^2(d)}{\varepsilon} + \frac{k^3 \log^3(d)}{\varepsilon^{3/2}}\right)$$

and returns a matrix $Z \in \mathbb{R}^{n \times k}$ with orthonormal columns for which

$$\|A - ZZ^\mathsf{T}A\|_2 \leq (1 + \varepsilon)\|A - [A]_k\|_2 \text{ and } \|A - ZZ^\mathsf{T}A\|_\mathsf{F} \leq (1 + \varepsilon)\|A - [A]_k\|_\mathsf{F}.$$

Frobenius Norm Reduced-Rank Regression. As discussed in the introduction, there is a closed form solution to the reduced-rank Frobenius norm regression problem.

Lemma 3 (Lemma 4.1 of [31], Lemma 2 of [18]) *Given matrices $A \in \mathbb{R}^{n \times c}$, $B \in \mathbb{R}^{n \times d}$, and a rank parameter $k \leq c$, let matrix Q denote an orthonormal basis for the column span of A. Then $\min_{\mathrm{rank}\text{-}k\ X}\|AX - B\|_\mathsf{F} = \|Q[Q^\mathsf{T}B]_k - B\|_\mathsf{F} = \|[AA^+B]_k - B\|_\mathsf{F}$. If $\bar{U}\bar{\Sigma}^2\bar{U}^\mathsf{T}$ is the SVD of $Q^\mathsf{T}AA^\mathsf{T}Q$, and $\bar{U}_k$ denotes the first k columns of $\bar{U}$, then $[Q^\mathsf{T}B]_k = \bar{U}_k\bar{U}_k^\mathsf{T}Q^\mathsf{T}B$, and therefore*

$$\min_{\mathrm{rank}\text{-}k\ X}\|AX - B\|_\mathsf{F} = \|Q[Q^\mathsf{T}B]_k - B\|_\mathsf{F} = \|(Q\bar{U}_k)(Q\bar{U}_k)^\mathsf{T}B - B\|_\mathsf{F}.$$

Chebyshev Polynomials. The Chebyshev polynomials are defined as

$$T_0(x) = 1, \ T_1(x) = x \text{ and } T_i(x) = 2xT_{i-1}(x) - T_{i-2}(x)$$

for all $i \geq 2$. Thus $T_i(x)$ is a polynomial of degree i. It can be shown that if i is odd, then $T_i(x)$ has only odd degree monomials. Chebyshev polynomial T_i has the property that $\|T_i\|_1 \leq (1 + \sqrt{2})^i$ for all i. See [18] for more properties of Chebyshev polynomials.

3. Previous work

Let $A \in \mathbb{R}^{n \times c}$ be a matrix and $U\Sigma V^\mathsf{T}$ be the "thin" SVD of A, where U is an orthonormal basis for the column space of A. Note that the projection matrix onto the column space of A is given by $AA^+ = UU^\mathsf{T}$. The first algorithm to solve $\min_{\mathrm{rank}\text{-}k\ X}\|AX - B\|_2$ was by Sou and Rantzer [27]. They consider the following problem:

$$\text{minimize rank}(X)$$

$$\text{such that } \|AX - B\|_2 < 1. \tag{1}$$

As multiplying a matrix with a projection matrix does not increase the operator norm, we have that $\|AX - B\|_2 \geq \|(I - AA^+)(AX - B)\|_2 = \|(I - AA^+)B\|_2$. Thus the problem is feasible only when $\|(I - AA^+)B\|_2 = \|(I - UU^\mathsf{T})B\|_2 < 1$. The following theorem characterizes the solution for (1).

Theorem 4 (Sou and Rantzer [27]) *Given matrices $A \in \mathbb{R}^{n \times c}$ and a matrix $B \in \mathbb{R}^{n \times d}$, if there is a matrix Y such that $\|AY - B\|_2 < 1$, then the optimum value of (1) is $\mathrm{sve}(B)$ where $\mathrm{sve}(B)$ denotes the number of singular values of B that are greater than or equal to 1.*

For an arbitrary $s > 0$, consider the problem (1) with matrices A/s and B/s. The problem is feasible if and only if $\|(I - UU^\mathsf{T})(B/s)\|_2 < 1$, i.e., if and only if $\|(I - UU^\mathsf{T})B\|_2 < s$. Suppose s is such that $s > \|(I - UU^\mathsf{T})B\|_2$. Then Theorem 4 implies that there is a rank k matrix X such that $\|(A/s)X - (B/s)\|_2 < 1$ if and only if $k \geq \mathrm{sve}(B/s)$, i.e., $\sigma_{k+1}(B/s) < 1$. This argument shows that for any $s > \max(\sigma_{k+1}(B), \|(I - UU^\mathsf{T})B\|_2)$, there is a rank k matrix X such that $\|AX - B\|_2 < s$. Thus $\mathrm{Opt} = \max(\sigma_{k+1}(B), \|(I - UU^\mathsf{T})B\|_2)$.

It is interesting and perhaps surprising that the above theorem implies we can obtain a solution that has a value $\max(\sigma_{k+1}(B), \|(I - UU^\mathsf{T})B\|_2)$, which is a simple lower bound on the optimum. This shows that if $\|(I - UU^\mathsf{T})B\|_2 \leq \sigma_{k+1}(B)$, there is a rank k matrix in the column span of matrix A that is as good of an approximation to B in spectral norm as $[B]_k$. Also, if $\|(I - UU^\mathsf{T})B\|_2 \geq \sigma_{k+1}(B)$, then there is a rank-$k$ matrix in the column space of A that is as good of an approximation to B in spectral norm as $AA^+B = UU^\mathsf{T}B$, the projection of B onto the column span of A.

We thus have the following corollary summarizing the discussion above. The corollary was also observed in [see 20, Section 4] in terms of a different parameter they call the critical rank.

Corollary 5 *Given matrices $A \in \mathbb{R}^{n \times c}, B \in \mathbb{R}^{n \times d}$ and a parameter k,*

$$\inf_{\text{rank-}k \; X} \|AX - B\|_2 = \max(\|(I - AA^+)B\|_2, \sigma_{k+1}(B)).$$

We give a proof of Theorem 4 for completeness in Appendix A.1. Our proof is similar to the proof of Sou and Rantzer [27] with some minor changes.

4. Reduced-Rank Regression in Operator Norm

We first consider the case when c, d are small. In this case, we could assume that we can compute matrices U and Δ, where U is an orthonormal basis for the column span of matrix A, and the matrix $\Delta = B^\mathsf{T}(I - UU^\mathsf{T})B$. We give a simple algorithm that demonstrates our techniques. We then extend these ideas to the case when c, d are large, for which computing an orthonormal basis for A and computing Δ is prohibitively expensive.

From Corollary 5, we have that $\mathrm{Opt} = \max(\|(I - UU^\mathsf{T})B\|_2, \sigma_{k+1}(B))$. Let β be such that $(1 + \varepsilon)\mathrm{Opt} \leq \beta \leq (1 + 2\varepsilon)\mathrm{Opt}$, which can be found using the Block Krylov algorithm. Throughout the paper we assume we know the value β.

Lemma 6 *If there exists a rank-k matrix X such that $\|UX - B\|_2 < \beta$, then $\sigma_{k+1}(U^\mathsf{T}B(\beta^2 I - \Delta)^{-1/2}) < 1$.*

Algorithm 1: Low Rank Approximation with Approximate Matrix Multiplication

---
Input: $M \in \mathbb{R}^{n \times d}, k \in \mathbb{Z}, \varepsilon > 0, \mathrm{Oracle}_M : \mathbb{R}^d \times \varepsilon \to \mathbb{R}^n, \mathrm{Oracle}_{M^\mathsf{T}} : \mathbb{R}^n \times \varepsilon \to \mathbb{R}^d$
Output: $Z \in \mathbb{R}^{n \times k}$
$G \sim \mathcal{N}(0,1)^{d \times k}, \kappa \leftarrow \sigma_1(M)/\sigma_{k+1}(M), q \leftarrow O((1/\sqrt{\varepsilon}) \log(d/\varepsilon))$
$\varepsilon_\circ \leftarrow O\left(\varepsilon/(\kappa^{2+5q}k^7 C^q)\right), \varepsilon_\bullet \leftarrow O\left(\varepsilon^2/(48\kappa(\kappa^2(\sqrt{qk})k))\right)$
```
/* Let ∘ and ● denote approximate matrix-vector products using the Oracles
   with accuracy ε∘ and ε●, respectively                                  */
```
$K' \leftarrow [(MM^\mathsf{T})^{\circ(q-1)/2}M \circ G, (MM^\mathsf{T})^{\circ(q-3)/2}M \circ G, \ldots, M \circ G]$
$Q' \leftarrow$ Orthonormal basis for K'
$[\bar{U}, \bar{\Sigma}^2, \bar{U}^\mathsf{T}] \leftarrow \mathrm{SVD}(Q'^\mathsf{T}(M \bullet (M^\mathsf{T} \bullet Q')))$
$\bar{U}_k \leftarrow$ First k columns of $\bar{U}$
$Z \leftarrow Q'\bar{U}_k$

The proof of this lemma is in Appendix B.1. The proof of the above lemma also shows that if we can find a matrix Y of rank k such that $\|Y - U^\mathsf{T}B(\beta^2 I - \Delta)^{-1/2}\|_2 \leq 1$, then we can obtain a matrix $X = Y(\beta^2 I - \Delta)^{1/2}$ such that $\|UX - B\|_2 < \beta$. Thus, we can compute the SVD of the matrix $U^\mathsf{T}B(\beta^2 I - \Delta)^{-1/2}$ and obtain $[U^\mathsf{T}B(\beta^2 I - \Delta)^{-1/2}]_k$ and obtain a solution $[U^\mathsf{T}B(\beta^2 I - \Delta)^{-1/2}]_k(\beta^2 I - \Delta)^{1/2}$ of cost β.

Computing an exact SVD, as required in the proof of above Lemma, is much slower than computing a rank k matrix that satisfies the guarantees of the best rank k matrices approximately. The following lemma shows that we can obtain a solution of cost close to β even if we can compute a rank k matrix Y such that $\|Y - U^\mathsf{T}B(\beta^2 I - \Delta)^{-1/2}\|_2 \leq 1 + \varepsilon$.

Lemma 7 *If Y is a rank k matrix such that $\|Y - U^\mathsf{T}B(\beta^2 I - \Delta)^{-1/2}\|_2 \leq 1 + \varepsilon$, then we obtain that $\|UY(\beta^2 I - \Delta)^{1/2} - B\|_2 \leq (1+\varepsilon)\beta$. Furthermore, $\|UY(UY)^+B - B\|_2 \leq (1+\varepsilon)\beta$.*

The proof of this lemma is in Appendix B.2. The above lemma states that a $1 + \varepsilon$ approximation to the best rank-k approximation of the matrix $U^\mathsf{T}B(\beta^2 I - \Delta)^{-1/2}$ in operator norm is sufficient to find a solution of cost $(1+\varepsilon)\beta$ to the reduced-rank regression problem. We can use the Block Krylov algorithm to compute such an approximation. The Block Krylov algorithm of Musco and Musco [18] only needs an oracle to compute matrix-vector products. In the case when c, d are small, we can compute the matrices $U, (\beta^2 I - \Delta)^{-1/2}$ and then given arbitrary vectors v, v' we can compute $U^\mathsf{T}B(\beta^2 I - \Delta)^{-1/2}v$ and $(\beta^2 I - \Delta)^{-1/2}B^\mathsf{T}Uv'$ and hence run the Block Krylov Algorithm. This gives a $1 + O(\varepsilon)$ approximation to the reduced-rank regression problem.

When r, d are large, it is expensive to compute the matrices U, Δ and $(\beta^2 I - \Delta)^{-1/2}$. As the analysis of Musco and Musco [18] works only when exact matrix-vector products can be computed, we cannot run the Block Krylov algorithm unless we compute the matrices U, Δ or at least are able to compute exact matrix vector products with the matrix $U^\mathsf{T}B(\beta^2 I - \Delta)^{-1/2}$. So we analyze their algorithm and show that it works even using approximate matrix products instead of exact matrix products, given that the error is low enough.

9

5. Block Krylov Iteration with Approximate Multiplication Oracle

Given a parameter k and an oracle to approximately compute Mv and $M^{\mathsf{T}}v'$, given arbitrary vectors v and v', we would like to compute a matrix Z with k orthonormal columns such that

$$\|M - ZZ^{\mathsf{T}}M\|_2 \leq (1 + \varepsilon)\sigma_{k+1}(M). \tag{2}$$

Specifically, suppose we have an oracle that, given an arbitrary vector v and approximation parameter $\varepsilon_\circ$, can compute in time $T(\varepsilon_\circ)$ a vector $M \circ v$ such that $\|Mv - (M \circ v)\|_2 \leq \varepsilon_\circ \|M\|_2 \|v\|_2$, and also given an arbitrary vector v' and accuracy parameter $\varepsilon_\circ$ can compute in time $T(\varepsilon_\circ)$ a vector $M^{\mathsf{T}} \circ v'$ such that $\|M^{\mathsf{T}}v' - M^{\mathsf{T}} \circ v'\|_2 \leq \varepsilon_\circ \|M\|_2 \|v'\|_2$. We are also given $\kappa = \sigma_1(M)/\sigma_{k+1}(M)$ and we want to compute a matrix Z as in (2).

Our algorithm to compute such a matrix Z is Algorithm 1. It is essentially the same as the Block Krylov algorithm of [18] with exact matrix-vector multiplication replaced by approximate matrix-vector multiplication with accuracy parameters as defined in our algorithm. Our main result for this section is the following theorem that states that the Block Krylov algorithm of [18] works even with approximate matrix-vector products.

Theorem 8 *Let $M \in \mathbb{R}^{n \times d}$, $k \leq d$ be a rank parameter, and $\varepsilon > 0$ be an accuracy parameter. Let $\kappa = \sigma_1(M)/\sigma_{k+1}(M)$. Given access to an oracle that can in time $T(\varepsilon_\circ)$ compute vectors $M \circ v$ and $M^{\mathsf{T}} \circ v'$ such that*

$$\|M \circ v - Mv\|_2 \leq \varepsilon_\circ \|M\|_2 \|v\|_2 \quad \text{and} \quad \|M^{\mathsf{T}} \circ v' - M^{\mathsf{T}}v'\|_2 \leq \varepsilon_\circ \|M\|_2 \|v'\|_2,$$

for any vectors v and v', Algorithm 1 computes a matrix $Z \in \mathbb{R}^{n \times k}$ with k orthonormal columns such that, with probability $\geq 3/5$, $\|(I - ZZ^{\mathsf{T}})M\|_2 \leq (1 + \varepsilon)\sigma_{k+1}(M)$. The running time is

$$O\left(T\left(\frac{\varepsilon}{2\kappa^{5q}k^{11}D^q}\right)qk + T\left(\frac{\varepsilon^2}{192\kappa^2(\sqrt{qk})k}\right)qk\right),$$

where $q = O\left((1/\sqrt{\varepsilon})\log(d/\varepsilon)\right)$ and D is an absolute constant. Further, if the approximations $M \circ v$ are spanned by M for all v, then the columns of the matrix Z are also spanned by the matrix M.

Proof sketch Proof of Block Krylov Iteration Algorithm of [18] first shows that there is a polynomial $p(x)$ that has only odd degree monomials such that the k-dimensional column space of the matrix $p(M)G$, where G is a Gaussian matrix with k columns, spans a $(1 + \varepsilon)$ approximation. As we do not now how to compute this polynomial $p(x)$, they show that the Krylov Space K spans this matrix $p(M)G$ and then show that the rank k Frobenius norm approximation of the matrix M inside the Krylov Subspace K is also a $1 + \varepsilon$ spectral norm rank k approximation.

We adapt their proof for the case when we can compute matrix-vector products only approximately. We first show that the approximate Krylov Matrix K' computed by Algorithm 1 is close to the actual Krylov Matrix K in Lemma 14. However, this lemma isn't sufficient to directly prove that rank-k Frobenius Norm Approximation of M inside K' is a $1 + \varepsilon$ rank-k spectral approximation as the matrices K and K' can be very poorly conditioned. Therefore, similar to the matrix $p(M)G$ in [18], we define a rank-k matrix Apx (see Equation 7) and show that the matrix Apx is spanned by K'. Then we show in Lemma 21 that the matrix Apx is close to $p(M)G$. Using an upper bound on

condition number of the matrix $p(M)G$ (see Lemma 18), we conclude in Equation 8 that projection matrices onto column spaces of the matrices Apx and $p(M)G$ are close.

Similar to the argument of [18], we encounter the issue that this matrix Apx cannot be computed as we don't know the parameters of the polynomial $p(x)$ but we do have that this matrix Apx is spanned by column space of K'. Using this fact, we show that an approximate rank k Frobenius norm approximation of M in column space of K' is also a $1 + \varepsilon$ spectral norm rank k approximation for the matrix M. We also show that this approximate rank k Frobenius norm approximation can be computed using approximate matrix-vector product oracles. ∎

6. Approximate Oracles and Reduced Rank Regression

Lemma 7 shows that if Y is a rank k matrix such that $\|Y - U^\mathsf{T} B(\beta^2 I - \Delta)^{-1/2}\|_2 \leq 1 + \varepsilon$, then $\|UY(UY)^+B - B\|_2 \leq (1+\varepsilon)\beta$. Based on this result, we prove the following lemma which shows that a low rank-approximation of the matrix $AA^+B(\beta^2 I - \Delta)^{-1/2}$ suffices.

Lemma 9 *Let $\widetilde{Z} \in \mathbb{R}^{n \times k}$ be a matrix with orthonormal columns such that*

$$\|AA^+B(\beta^2 I - \Delta)^{-1/2} - \widetilde{Z}\widetilde{Z}^\mathsf{T}AA^+B(\beta^2 I - \Delta)^{-1/2}\|_2 \leq 1 + \varepsilon,$$

then $\|(AA^+\widetilde{Z})(AA^+\widetilde{Z})^+B - B\|_2 \leq (1 + \varepsilon)\beta$.

Proof of the lemma is in Appendix D.1. Hence, if we can get a good k-dimensional space $\widetilde{Z}$ for approximating the matrix $AA^+B(\beta^2 I - \Delta)^{-1/2}$, we can then obtain a good k dimensional space for B. We first show that we can instead find a low rank approximation for a matrix AA^+BM/β, for a suitable matrix M, which will also be a good low rank approximation for $AA^+B(\beta^2 I - \Delta)^{-1/2}$.

Lemma 10 *Given that $\beta \geq (1 + \varepsilon)\mathrm{Opt}$, there exists a polynomial $r(x)$ of degree at most $t = O\left(1/\sqrt{\varepsilon}\log(\kappa/\varepsilon)\right)$ such that for $M = r(\Delta/\beta^2)$, if $\widetilde{Z}$ is a matrix such that*

$$\|AA^+BM/\beta - \widetilde{Z}\widetilde{Z}^\mathsf{T}(AA^+BM/\beta)\|_2 \leq 1 + \varepsilon,$$

then $\|AA^+B(\beta^2 I - \Delta)^{-1/2} - \widetilde{Z}\widetilde{Z}^\mathsf{T}AA^+B(\beta^2 I - \Delta)^{-1/2}\|_2 \leq 1 + O(\varepsilon)$. Furthermore, $\|r\|_1 = O((1 + \sqrt{2})^{O(\sqrt{1/\varepsilon}\log(\kappa/\varepsilon))}\log(\kappa/\varepsilon)/\varepsilon)$, $\|M\|_2 \leq 2/\sqrt{\varepsilon}$ and $\sigma_{\min}(M) \geq 1/2$.

The proof of the above lemma is in Appendix D.2. From Theorem 8, to find a $1 + \varepsilon$ approximation for rank k spectral norm low rank approximation (LRA) of the matrix $\mathcal{M}'$, we need only a way to compute the products $\mathcal{M}'v$ and $\mathcal{M}'^\mathsf{T}v'$ for any vectors v, v'. As $r(\Delta/\beta^2)$ is a polynomial in the matrix Δ/β^2, it is much easier to design approximate multiplication oracles for the matrix AA^+BM/β than for the matrix $AA^+B(\beta^2 I - \Delta)^{-1/2}$. The following lemma shows that we can compute good approximations to the matrix vector products and then compute a $1+\varepsilon$ approximation to the LRA of matrix $\mathcal{M}' = AA^+B\frac{r(\Delta/\beta^2)}{\beta}$.

Lemma 11 *Given arbitrary vectors v, v' and an accuracy parameter ε_f, Algorithms 3 and 4 compute vectors y, y' such that $\|\mathcal{M}'v - y\|_2 \leq \varepsilon_\mathsf{f}\|v\|_2$ and $\|\mathcal{M}'^\mathsf{T}v' - y'\|_2 \leq \varepsilon_\mathsf{f}\|y\|_2$ in time*

$$T(\varepsilon_\mathsf{f}) := O(t \cdot (\mathsf{nnz}(B) + (\mathsf{nnz}(A) + c^2)\log\left(\kappa(B)^2\|r\|_1/(\varepsilon_\mathsf{f}\varepsilon)\right)))$$
$$+ O((\mathsf{nnz}(A) + c^2)\log(\kappa(B)/(\varepsilon_\mathsf{f}\varepsilon)))$$

where $t = O(\sqrt{1/\varepsilon}\log(\kappa/\varepsilon))$ and $\|r\|_1 = (1 + \sqrt{2})^{O(1/\sqrt{\varepsilon}\log(\kappa/\varepsilon))}\log(\kappa/\varepsilon)/\varepsilon$.

Algorithm 2: Operator Norm Regression

Input: $A \in \mathbb{R}^{n \times c}, B \in \mathbb{R}^{n \times d}, k \in \mathbb{Z}, \varepsilon > 0$

Output: $X' \in \mathbb{R}^{c \times k}, X'' \in \mathbb{R}^{k \times d}$

$\beta \leftarrow (1 + \varepsilon/2) \max(\sigma_{k+1}(B), \|(I - AA^+)B\|_2)$

$\Delta \leftarrow B^\mathsf{T}(I - AA^+)B$ /* Not computed explicitly */

/* Let $r(x)$ be the polynomial given by Lemma 10   */

$\mathcal{M}' \leftarrow (AA^+B/\beta)r(\Delta/\beta^2)$ /* Not computed explicitly */

$Z \leftarrow$ Algorithm 1($\mathcal{M}', k, \varepsilon/2$, APXPRODUCT, APXPRODCUTTRANSPOSE)

$X' \leftarrow$ HIGHPRECISIONREGRESSION($A, Z, 1/2$)

$X'' \leftarrow Z^\mathsf{T} \cdot B$

6.1. Main Theorem

We finally have our main Theorem that shows that Algorithm 2 outputs a $1 + \varepsilon$ approximation in factored form. Proof of the Theorem is in Appendix D.4.

Theorem 12 *Given matrices $A \in \mathbb{R}^{n \times c}$ and $B \in \mathbb{R}^{n \times d}$, a rank parameter $k \leq c$ and an accuracy parameter ε Algorithm 2 runs in time*

$$O\left(\left(\frac{\mathsf{nnz}(B) \cdot k}{\varepsilon} + \frac{\mathsf{nnz}(A) \cdot k}{\varepsilon^{1.5}} + \frac{c^2 k}{\varepsilon^{1.5}}\right) \cdot \mathrm{polylog}(\kappa, \kappa(AA^+B), d, k, 1/\varepsilon) + c^\omega\right).$$

and with probability $4/5$ outputs a matrix Z with k orthonormal columns and $\mathrm{colspan}(Z) \subseteq \mathrm{colspan}(A)$ such that $\|ZZ^\mathsf{T}B - B\|_2 \leq (1 + \varepsilon)\mathsf{Opt}$ and outputs matrices $X' \in \mathbb{R}^{c \times k}$ and $X'' \in \mathbb{R}^{k \times d}$ such that $\|A(X' \cdot X'') - B\|_2 = \|ZZ^\mathsf{T}B - B\|_2 \leq (1 + \varepsilon)\mathsf{Opt}$.

6.2. Removing $\kappa(AA^+B)$ dependence

We observe that we can add a random rank $k + 1$ matrix to B to obtain a matrix $\widetilde{B}$ for which $\kappa(AA^+\widetilde{B})$ is bounded in terms of $\kappa(B)$. We also show that any arbitrary vector v can be multiplied with the matrix $\widetilde{B}$ in time comparable to $\mathsf{nnz}(B)$.

Lemma 13 *Given any matrices $A \in \mathbb{R}^{n \times c}$ and $B \in \mathbb{R}^{n \times d}$, if $\mathrm{rank}(A) \geq k + 1$, then there exists a matrix $\widetilde{B}$ such that if*

$$\|A\widetilde{X} - \widetilde{B}\|_2 \leq (1 + \varepsilon/2) \min_{\text{rank-}k \ X} \|AX - \widetilde{B}\|_2 \tag{3}$$

for a rank k matrix $\widetilde{X}$, then

$$\|A\widetilde{X} - B\|_2 \leq (1 + \varepsilon)\mathsf{Opt}.$$

Additionally, $\kappa(AA^\dagger\widetilde{B}) = \sigma_1(AA^+\widetilde{B})/\sigma_{k+1}(AA^+\widetilde{B}) \leq (Cn/\varepsilon)\sigma_1(B)/\sigma_{k+1}(B)$ and given a vector v, $\widetilde{B}v$ can be computed in $O(\mathsf{nnz}(B) + (n + d)k)$ time.

The proof of this lemma is in Appendix D.5. Therefore we run Algorithm 2 on matrix $\widetilde{B}$ and can compute a $(1 + \varepsilon)$ approximate solution to the problem $\min_{\text{rank-}k \ X} \|AX - B\|_2$ in time

$$O\left(\left(\frac{\mathsf{nnz}(B) \cdot k}{\varepsilon} + \frac{(n + d)k^2}{\varepsilon} + \frac{\mathsf{nnz}(A) \cdot k}{\varepsilon^{1.5}} + \frac{c^2 k}{\varepsilon^{1.5}}\right) \cdot \mathrm{polylog}(\kappa, n, d, k, 1/\varepsilon) + c^\omega\right). \tag{4}$$

7. Experiments and Implementation

It is evident that our algorithm is faster than the algorithm of Sou and Rantzer [27] for large matrices A and B as their algorithm cannot make use of the sparsity of the matrices and also has to compute eigen value decomposition of a dense and large $d \times d$ matrix. Let $n = d = 7000, c = 100$ and $k = 30$. We instantiate an $n \times d$ matrix B with 5% of the entries being nonzero where each nonzero entry is sampled independently from a uniform distribution on $[0, 1]$. The $n \times c$ matrix A is obtained by taking first c columns of the matrix B. With $\varepsilon = 0.05$, our algorithm runs in less than 20 seconds whereas an implementation of Sou and Rantzer's algorithm runs in around 10 minutes. For even larger values of n and d, our algorithm will be faster by an even larger factor. An implementation of our algorithm and the above example is available here [2].

Acknowledgments

References

[1] Maria-Florina Balcan, Simon Shaolei Du, Yining Wang, and Adams Wei Yu. An improved gap-dependency analysis of the noisy power method. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 284–309, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR. URL http://proceedings.mlr.press/v49/balcan16a.html.

[2] Christos Boutsidis. *Topics in Matrix Sampling Algorithms*. PhD thesis, Rensselaer Polytechnic Institute, USA, 2011.

[3] Cajo JF Ter Braak and Caspar WN Looman. Biplots in reduced-rank regression. *Biometrical journal*, 36(8):983–1003, 1994.

[4] Mark Braverman, Elad Hazan, Max Simchowitz, and Blake Woodworth. The gradient complexity of linear regression. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 627–647. PMLR, 09–12 Jul 2020. URL http://proceedings.mlr.press/v125/braverman20a.html.

[5] James B. Carrell. *Groups, Matrices, and Vector Spaces: A Group Theoretic Approach to Linear Algebra*. Springer New York, New York, NY, 2017. doi: 10.1007/978-0-387-79428-0. URL http://link.springer.com/10.1007/978-0-387-79428-0.

[6] Kenneth L. Clarkson and David P. Woodruff. Low rank approximation and regression in input sparsity time. In *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013*, pages 81–90, 2013. doi: 10.1145/2488608.2488620. URL https://doi.org/10.1145/2488608.2488620.

---

2. https://gitlab.com/praneeth10/operator-norm-reduced-rank-regression

[7] Shmuel Friedland and Anatoli Torokhti. Generalized rank-constrained matrix approximations. *SIAM Journal on Matrix Analysis and Applications*, 29(2):656–659, 2007.

[8] M. Gu. Subspace iteration randomization and singular value problems. *SIAM Journal on Scientific Computing*, 37(3):A1139–A1173, 2015. doi: 10.1137/130938700. URL https://doi.org/10.1137/130938700.

[9] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

[10] Moritz Hardt and Eric Price. The noisy power method: A meta algorithm with applications. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, pages 2861–2869. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/729c68884bd359ade15d5f163166738a-Paper.pdf.

[11] Moritz Hardt and Aaron Roth. Beyond worst-case analysis in private singular vector computation. In *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, STOC '13, page 331–340, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450320290. doi: 10.1145/2488608.2488650. URL https://doi.org/10.1145/2488608.2488650.

[12] Haotian Jiang, Yin Tat Lee, Zhao Song, and Sam Chiu-wai Wong. An improved cutting plane method for convex optimization, convex-concave games, and its applications. In *Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020, Chicago, IL, USA, June 22-26, 2020*, pages 944–953, 2020. doi: 10.1145/3357713.3384284. URL https://doi.org/10.1145/3357713.3384284.

[13] Zohar Karnin and Edo Liberty. Online pca with spectral bounds. In *Conference on Learning Theory*, pages 1129–1140, 2015.

[14] Dmitry Kobak, Yves Bernaerts, Marissa A Weis, Federico Scala, Andreas Tolias, and Philipp Berens. Sparse reduced-rank regression for exploratory visualization of multimodal data sets. *bioRxiv*, page 302208, 2019.

[15] Pierre Legendre and Marti J Anderson. Distance-based redundancy analysis: testing multi-species responses in multifactorial ecological experiments. *Ecological monographs*, 69(1):1–24, 1999.

[16] PJ Maher. Some norm inequalities concerning generalized inverses, 2. *Linear algebra and its applications*, 420(2-3):517–525, 2007.

[17] Xiangrui Meng and Michael W. Mahoney. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In Dan Boneh, Tim Roughgarden, and Joan Feigenbaum, editors, *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013*, pages 91–100. ACM, 2013. doi: 10.1145/2488608.2488621. URL https://doi.org/10.1145/2488608.2488621.

[18] Cameron Musco and Christopher Musco. Randomized block krylov methods for stronger and faster approximate singular value decomposition. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1396–1404, Cambridge, MA, USA, 2015. MIT Press.

[19] Cameron Musco, Christopher Musco, and Aaron Sidford. Stability of the lanczos method for matrix function approximation. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '18, page 1605–1624, USA, 2018. Society for Industrial and Applied Mathematics. ISBN 9781611975031.

[20] Srinivas Nambirajan. *Topics in Matrix Approximation*. PhD thesis, Rensselaer Polytechnic Institute, USA, 2015.

[21] Jelani Nelson and Huy L. Nguyen. OSNAP: faster numerical linear algebra algorithms via sparser subspace embeddings. In *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2013, 26-29 October, 2013, Berkeley, CA, USA*, pages 117–126, 2013.

[22] Mark Rudelson and Roman Vershynin. Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics*, 62(12):1707–1739, 2009. doi: 10.1002/cpa.20294. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.20294.

[23] Mark Rudelson and Roman Vershynin. Non-asymptotic theory of random matrices: Extreme singular values. *Proceedings of the International Congress of Mathematicians 2010, ICM 2010*, 03 2010. doi: 10.1142/9789814324359_0111.

[24] Sushant Sachdeva and Nisheeth K. Vishnoi. Faster algorithms via approximation theory. *Foundations and Trends® in Theoretical Computer Science*, 9(2):125–210, 2014. ISSN 1551-305X. doi: 10.1561/0400000065. URL http://dx.doi.org/10.1561/0400000065.

[25] Yiyuan She and Kun Chen. Robust reduced-rank regression. *Biometrika*, 104(3):633–647, 2017.

[26] Max Simchowitz, Ahmed El Alaoui, and Benjamin Recht. Tight query complexity lower bounds for pca via finite sample deformed wigner law. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1249–1259, 2018.

[27] Kin Cheong Sou and Anders Rantzer. On generalized matrix approximation problem in the spectral norm. *Linear Algebra and its Applications*, 436(7):2331–2341, 2012.

[28] Arthur Szlam, Yuval Kluger, and Mark Tygert. An implementation of a randomized algorithm for principal component analysis. *arXiv preprint arXiv:1412.3510*, 2014.

[29] Raja Velu and Gregory C Reinsel. *Multivariate reduced-rank regression: theory and applications*, volume 136. Springer Science & Business Media, 2013.

[30] Van H. Vu and Terence Tao. The condition number of a randomly perturbed matrix. In *Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing*, STOC '07, page 248–255, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595936318. doi: 10.1145/1250790.1250828. URL https://doi.org/10.1145/1250790.1250828.

[31] David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends®
in Theoretical Computer Science*, 10(1–2):1–157, 2014. ISSN 1551-305X. doi: 10.1561/
0400000060. URL http://dx.doi.org/10.1561/0400000060.

Appendix A. Omitted Proofs from Section 3

A.1. Proof of Theorem 4

Proof Without loss of generality, we prove the theorem assuming A has orthonormal columns.
Thus $U = A$. Let X be an arbitrary matrix such that $\|UX - B\|_2 < 1$. We will give a series of
statements equivalent to $\|UX - B\|_2 < 1$ that prove the theorem. Using the fact that for any matrix
A, $\|A\|_2 < 1$ if and only if $A^\mathsf{T}A \prec I$, we obtain the equivalent statement

$$(UX - B)^\mathsf{T}(UX - B) \prec I.$$

Writing B as $UU^\mathsf{T}B + (I - UU^\mathsf{T})B$, we get another equivalent statement

$$(UX - UU^\mathsf{T}B)^\mathsf{T}(UX - UU^\mathsf{T}B) \prec I - B^\mathsf{T}(I - UU^\mathsf{T})B = I - \Delta.$$

As the LHS of the above relation is a positive semi-definite matrix, we obtain that $I - \Delta \succ 0$ and
hence invertible. Thus the above condition can be equivalently written as

$$(I - \Delta)^{-1/2}(UX - UU^\mathsf{T}B)^\mathsf{T}(UX - UU^\mathsf{T}B)(I - \Delta)^{-1/2} \prec I.$$

Using the fact that Δ is symmetric and $U^\mathsf{T}U = I$, we get that the above condition is the same as

$$\|X(I - \Delta)^{-1/2} - U^\mathsf{T}B(I - \Delta)^{-1/2}\|_2 < 1.$$

Thus we obtain that in the case that $\|(I - UU^\mathsf{T})B\|_2 < 1$, for an arbitrary matrix X, the condition
that $\|UX - B\|_2 < 1$ is equivalent to $\|X(I - \Delta)^{-1/2} - U^\mathsf{T}B(I - \Delta)^{-1/2}\|_2 < 1$. Let $\hat{B} :=
U^\mathsf{T}B(I - \Delta)^{-1/2}$. It is easy to see that $X = [\hat{B}]_{\mathrm{sve}(\hat{B})}(I - \Delta)^{1/2}$ satisfies $\|UX - B\|_2 < 1$ and
that any matrix X that satisfies $\|UX - B\|_2 < 1$ must have rank at least $\mathrm{sve}(\hat{B})$. All that remains to
show is that $\mathrm{sve}(\hat{B}) = \mathrm{sve}(B)$. We will show that $k^-(I - \hat{B}^\mathsf{T}\hat{B}) = k^-(I - B^\mathsf{T}B)$ which completes
the proof.

$$\begin{aligned}
I - \hat{B}^\mathsf{T}\hat{B} &= I - (I - \Delta)^{-1/2}B^\mathsf{T}UU^\mathsf{T}B(I - \Delta)^{-1/2} \\
&= I - (I - \Delta)^{-1/2}(B^\mathsf{T}B - \Delta)(I - \Delta)^{-1/2} \\
&= I - (I - \Delta)^{-1/2}(B^\mathsf{T}B - I + I - \Delta)(I - \Delta)^{-1/2} \\
&= I - I + (I - \Delta)^{-1/2}(I - B^\mathsf{T}B)(I - \Delta)^{-1/2} \\
&= (I - \Delta)^{-1/2}(I - B^\mathsf{T}B)(I - \Delta)^{-1/2}.
\end{aligned}$$

Thus $k^-(I - \hat{B}^\mathsf{T}\hat{B}) = k^-((I - \Delta)^{-1/2}(I - B^\mathsf{T}B)(I - \Delta)^{-1/2})$. By Sylvester's law of inertia [5,
p313], $k^-((I - \Delta)^{-1/2}(I - B^\mathsf{T}B)(I - \Delta)^{-1/2}) = k^-(I - B^\mathsf{T}B)$. Therefore

$$\mathrm{sve}(\hat{B}) = k^-(I - \hat{B}^\mathsf{T}\hat{B}) = k^-(I - B^\mathsf{T}B) = \mathrm{sve}(B).$$

Thus $\mathrm{sve}(B)$ is the optimum value for (1) if it is feasible. ∎

Appendix B. Omitted Proofs from Section 4

B.1. Proof of Lemma 6

Proof The proof of this lemma is very similar to the proof of Theorem 4. Suppose there exists a rank-k matrix X such that $\|UX - B\|_2 < \beta$. We already have $\beta > \|(I - UU^\mathsf{T})B\|_2$. The statement $\|UX - B\|_2 < \beta$ implies that

$$(UX - B)^\mathsf{T}(UX - B) \preceq \beta^2 I.$$

We can write $B = UU^\mathsf{T}B + (I - UU^\mathsf{T})B$ and obtain that for any matrix X, $(UX - B)^\mathsf{T}(UX - B) = (UX - UU^\mathsf{T}B)^\mathsf{T}(UX - UU^\mathsf{T}B) + \Delta$ which implies that

$$(UX - UU^\mathsf{T}B)^\mathsf{T}(UX - UU^\mathsf{T}B) \preceq \beta^2 I - \Delta.$$

As $\|\Delta\|_2 = \|(I - UU^\mathsf{T})B\|_2^2 < \beta^2$, $\beta^2 I - \Delta$ is invertible which implies that

$$(\beta^2 I - \Delta)^{-1/2}(UX - UU^\mathsf{T}B)^\mathsf{T}(UX - UU^\mathsf{T}B)(\beta^2 I - \Delta)^{-1/2} \preceq I.$$

Thus we have $\|(UX - UU^\mathsf{T}B)(\beta^2 I - \Delta)^{-1/2}\|_2 = \|X(\beta^2 I - \Delta)^{-1/2} - U^\mathsf{T}B(\beta^2 I - \Delta)^{-1/2}\|_2$ is less than or equal to 1. As X is a matrix of rank k, the matrix $X(\beta^2 I - \Delta)^{-1/2}$ also has rank k. Therefore

$$
\begin{aligned}
\sigma_{k+1}(U^\mathsf{T}B(\beta^2 I - \Delta)^{-1/2}) &= \|[U^\mathsf{T}B(\beta^2 I - \Delta)^{-1/2}]_k - U^\mathsf{T}B(\beta^2 I - \Delta)^{-1/2}\|_2 \\
&\leq \|X(\beta^2 I - \Delta)^{-1/2} - U^\mathsf{T}B(\beta^2 I - \Delta)^{-1/2}\|_2 \\
&\leq 1.
\end{aligned}
$$

∎

B.2. Proof of Lemma 7

Proof Suppose Y is a rank k matrix such that $\|Y - U^\mathsf{T}B(\beta^2 I - \Delta)^{-1/2}\|_2 \leq 1 + \varepsilon$. Then we have $\|Y(\beta^2 I - \Delta)^{1/2}(\beta^2 I - \Delta)^{-1/2} - U^\mathsf{T}B(\beta^2 I - \Delta)^{-1/2}\|_2 \leq 1 + \varepsilon$ and therefore

$$(\beta^2 I - \Delta)^{-1/2}(Y(\beta^2 I - \Delta)^{1/2} - U^\mathsf{T}B)^\mathsf{T}(Y(\beta^2 I - \Delta)^{1/2} - U^\mathsf{T}B)(\beta^2 I - \Delta)^{-1/2} \preceq (1 + \varepsilon)^2 I.$$

Multiplying the above relation on both sides with $(\beta^2 I - \Delta)^{1/2}$ on the left and the right, we obtain

$$(Y(\beta^2 I - \Delta)^{1/2} - U^\mathsf{T}B)^\mathsf{T}(Y(\beta^2 I - \Delta)^{1/2} - U^\mathsf{T}B) \preceq (1 + \varepsilon)^2 (\beta^2 I - \Delta).$$

Using $U^\mathsf{T}U = I$ and adding Δ to both sides, we conclude that

$$\|UY(\beta^2 I - \Delta)^{1/2} - B\|_2 \leq \sqrt{\|(1 + \varepsilon)^2 \beta^2 I\|_2} \leq (1 + \varepsilon)\beta.$$

Now Y is a matrix that has rank at most k. We also have $\|UYZ - B\|_2 \geq \|UY(UY)^+B - B\|_2$ for any matrix Z. Therefore $\|UY(UY)^+B - B\|_2 \leq \|UY(\beta^2 I - \Delta)^{1/2} - B\|_2 \leq (1 + \varepsilon)\beta$. ∎

Appendix C. Omitted Proofs from Section 5

C.1. Error in Computing Krylov Subspace

Given a matrix $M \in \mathbb{R}^{n \times d}$, an integer $k \le d$ and an odd integer $q \ge 0$, the Krylov Subspace is defined by

$$K = [(MM^{\mathsf{T}})^{(q-1)/2}MG, \ (MM^{\mathsf{T}})^{(q-3)/2}MG, \ \cdots, \ (MM^{\mathsf{T}})^1 MG, \ MG]$$

where G is a $d \times k$ matrix with i.i.d normal entries. Using the algorithm to approximately multiply a vector with the matrices M and M^{T}, we compute an approximation to the matrix K defined above. For any vector v, define $(MM^{\mathsf{T}})^{\circ 0}v := v$ and for $i > 0$, define $(MM^{\mathsf{T}})^{\circ i}v := M \circ (M^{\mathsf{T}} \circ ((MM^{\mathsf{T}})^{\circ(i-1)}v))$ (recall $M \circ v$ is the approximation for Mv computed by the oracle). The notation is similarly extended to approximate matrix multiplication using the oracle. Now we define the matrix

$$K' = [(MM^{\mathsf{T}})^{\circ(q-1)/2}M \circ G, \ (MM^{\mathsf{T}})^{\circ(q-3)/2}M \circ G, \ \cdots, \ (MM^{\mathsf{T}})^{\circ 1}M \circ G, \ M \circ G].$$

Let Q, Q' denote orthonormal bases for the matrices K and K' respectively. We now bound $\|K - K'\|_{\mathsf{F}}$ and the time required to compute K' using the following lemma.

Lemma 14 *For any matrix $M \in \mathbb{R}^{n \times d}$, matrix $G \in \mathbb{R}^{d \times k}$ and an odd integer q, let $\Delta_{i,G} := (MM^{\mathsf{T}})^{(i-1)/2}MG - (MM^{\mathsf{T}})^{\circ(i-1)/2}M \circ G$ and matrices $K, K' \in \mathbb{R}^{n \times qk}$ be as defined above. Then*

$$E_{i,G} := \|\Delta_{i,G}\|_{\mathsf{F}} \le 8\varepsilon_\circ (2^{i/2}\|M\|_2^i \|G\|_{\mathsf{F}})$$

for $i = 1, 3, 5, \ldots, q$ and $\|K - K'\|_{\mathsf{F}} \le O(\varepsilon_\circ \|G\|_{\mathsf{F}} \|M\|_2^{q+1} 2^{(q+1)/2})$. The matrix K' can be computed in $O(T(\varepsilon_\circ)qk)$.

Proof For an arbitrary vector v and i odd, let $\Delta_i := (MM^{\mathsf{T}})^{(i-1)/2}Mv - (MM^{\mathsf{T}})^{\circ(i-1)/2}M \circ v$. Let $E_i = \|\Delta_i\|_2$. We have $E_1 = \|\Delta_1\|_2 = \|Mv - M \circ v\|_2 \le \|M\|_2 \|v\|_2$. We now define a recurrence relation between E_i and E_{i-2} and then bound E_i using this recurrence. We have

$$\begin{aligned}
\Delta_i &= (MM^{\mathsf{T}})^{(i-1)/2}Mv - (MM^{\mathsf{T}})^{\circ(i-1)/2}M \circ v \\
&= (MM^{\mathsf{T}})(MM^{\mathsf{T}})^{(i-3)/2}Mv - (MM^{\mathsf{T}})^{\circ 1}(MM^{\mathsf{T}})^{\circ(i-3)/2}M \circ v \\
&= (MM^{\mathsf{T}})[(MM^{\mathsf{T}})^{(i-3)/2}Mv - (MM^{\mathsf{T}})^{\circ(i-3)/2}M \circ v] \\
&\quad + [(MM^{\mathsf{T}})^1(MM^{\mathsf{T}})^{\circ(i-3)/2}M \circ v - (MM^{\mathsf{T}})^{\circ 1}(MM^{\mathsf{T}})^{\circ(i-3)/2}M \circ v] \\
&= (MM^{\mathsf{T}})\Delta_{i-2} + [(MM^{\mathsf{T}})^1(MM^{\mathsf{T}})^{\circ(i-3)/2}M \circ v - (MM^{\mathsf{T}})^{\circ 1}(MM^{\mathsf{T}})^{\circ(i-3)/2}M \circ v].
\end{aligned}$$

Therefore by triangle inequality of $\|\cdot\|_2$,

$$\begin{aligned}
E_i &\le \|MM^{\mathsf{T}}\Delta_{i-2}\|_2 + \|(MM^{\mathsf{T}})^1(MM^{\mathsf{T}})^{\circ(i-3)/2}M \circ v - (MM^{\mathsf{T}})^{\circ 1}(MM^{\mathsf{T}})^{\circ(i-3)/2}M \circ v\|_2 \\
&\le \|M\|_2^2 E_{i-2} + \|(MM^{\mathsf{T}})^1(MM^{\mathsf{T}})^{\circ(i-3)/2}M \circ v - (MM^{\mathsf{T}})^{\circ 1}(MM^{\mathsf{T}})^{\circ(i-3)/2}M \circ v\|_2.
\end{aligned}$$

Let $v' := (MM^{\mathsf{T}})^{\circ(i-3)/2}M \circ v$. We now bound $\|MM^{\mathsf{T}}v' - (MM^{\mathsf{T}})^{\circ 1}v'\|_2$.

$$\|MM^{\mathsf{T}}v' - (MM^{\mathsf{T}})^{\circ 1}v'\|_2 = \|MM^{\mathsf{T}}v' - M \circ (M \circ v')\|_2$$

$$\leq \|MM^{\mathsf{T}}v' - M(M^{\mathsf{T}} \circ v')\|_2 + \|M(M^{\mathsf{T}} \circ v') - M \circ (M^{\mathsf{T}} \circ v')\|_2$$
$$\leq \|M\|_2 \|M^{\mathsf{T}}v' - M^{\mathsf{T}} \circ v'\|_2 + \varepsilon_\circ \|M\|_2 \|M^{\mathsf{T}} \circ v'\|_2$$
$$\leq \varepsilon_\circ \|M\|_2^2 \|v'\|_2 + \varepsilon_\circ \|M\|_2 (\varepsilon_\circ \|M\|_2 \|v'\|_2 + \|M^{\mathsf{T}}v'\|_2)$$
$$\leq 3\varepsilon_\circ \|M\|_2^2 \|v'\|_2.$$

As $v' = (MM^{\mathsf{T}})^{(i-3)/2}Mv - \Delta_{i-2}$, we get $\|v'\|_2 \leq \|(MM^{\mathsf{T}})^{(i-3)/2}Mv\|_2 + \|\Delta_{i-2}\|_2 \leq \|M\|_2^{i-2}\|v\|_2 + E_{i-2}$. Therefore we finally obtain that

$$E_i \leq \|M\|_2^2 E_{i-2} + 3\varepsilon_\circ \|M\|_2^2 \|v'\|_2 \leq \|M\|_2^2 E_{i-2} + 3\varepsilon_\circ \|M\|_2^2 (\|M\|_2^{i-2}\|v\|_2 + E_{i-2})$$
$$\leq (1 + 3\varepsilon_\circ)\|M\|_2^2 E_{i-2} + 3\varepsilon_\circ \|M\|_2^i \|v\|_2.$$

Solving this recurrence relation we obtain that

$$E_i \leq (1 + 3\varepsilon_\circ)^{(i-1)/2}\|M\|_2^{i-1}E_1 + (1 + (1 + 3\varepsilon_\circ) + \cdots + (1 + 3\varepsilon_\circ)^{(i-3)/2})(3\varepsilon_\circ \|M\|_2^i \|v\|_2)$$
$$\leq \varepsilon_\circ(1 + 2^{(i-1)/2}(3\varepsilon_\circ))\|M\|_2^i \|v\|_2 + 2^{(i-1)/2}(3\varepsilon_\circ)\|M\|_2^i \|v\|_2$$
$$\leq 8(\varepsilon_\circ 2^{i/2}\|M\|_2^i \|v\|_2).$$

In the above inequalities, we used the standard inequality $(1 + x)^n \leq 1 + 2^n x$ if $0 \leq x \leq 1$. Thus for any arbitrary vector v, $\|(MM^{\mathsf{T}})^{\circ(i-1)/2}M \circ v - (MM^{\mathsf{T}})^{(i-1)/2}Mv\|_2 \leq 8\varepsilon_\circ 2^{i/2}\|M\|_2^i \|v\|_2$ and therefore for the Gaussian matrix G,

$$E_{i,G} = \|(MM^{\mathsf{T}})^{\circ(i-1)/2}M \circ G - (MM^{\mathsf{T}})^{(i-1)/2}MG\|_{\mathsf{F}} \leq 8\varepsilon_\circ 2^{i/2}\|M\|_2^i \|G\|_{\mathsf{F}}.$$

We then have that $\|K - K'\|_{\mathsf{F}} \leq O(\varepsilon_\circ \|G\|_{\mathsf{F}} \|M\|_2^{q+1} 2^{(q+1)/2})$. In computing the matrix K' we make $O(qk)$ calls to each of the oracles and therefore take $O(T(\varepsilon_\circ)qk)$ time. ∎

Musco and Musco [18] consider a polynomial $p(x)$ such that the column space of the matrix $p(M)G$ is spanned by K. They then argue that the column span of $p(M)G$ is a "good" k-dimensional subspace to project M onto and then conclude that the best rank k approximation of M inside the span of K satisfies (2). Although we have an upper bound on $\|K - K'\|_{\mathsf{F}}$ from the above lemma, we cannot directly argue that the best rank k approximation of M inside K' satisfies the guarantee of (2), as the matrix K might be very poorly conditioned.

To overcome this issue, we first show that the matrix $p(M)G$ has a bounded condition number with $O(1)$ probability and that K' spans a matrix Apx that is close to $p(M)G$. We then show that the span of the matrix Apx is a good subspace to project the matrix M onto and then conclude that the best rank k approximation of M inside the span of K' satisfies (2).

C.2. Condition Number of the matrix $p(M)G$ and existence of good rank k subspace inside an approximate Krylov Subspace

Throughout this section let $\alpha = \sigma_{k+1}(M)$ and $\gamma = \varepsilon/2$. Let q be an odd integer and $T(x)$ be the degree q Chebyshev polynomial. Define

$$p(x) := (1 + \gamma)\alpha \frac{T(x/\alpha)}{T(1 + \gamma)}. \tag{5}$$

The following lemma bounds $\sigma_1(p(M))/\sigma_{k+1}(p(M))$ which lets us bound $\kappa(p(M)G)$.

Lemma 15 *If $M \in \mathbb{R}^{n \times d}$ is a matrix such that $\sigma_1(M)/\sigma_{k+1}(M) = \kappa$, then*

$$\sigma_1(p(M))/\sigma_{k+1}(p(M)) \leq (3\kappa)^q.$$

First, we have the following lemma that shows that $T(x) \geq 1$ for all $x \geq 1$ for the Chebyshev Polynomial T of any degree d.

Lemma 16 *If $T_d(x)$ is the degree d Chebyshev Polynomial, then for all $d \geq 0$ and for all $x \geq 1$, $T_{d+1}(x) \geq T_d(x) \geq 1$.*

Proof We prove the theorem using induction on the degree d. We have $T_0(x) = 1$ and $T_1(x) = x$. Thus $T_1(x) \geq T_0(x) \geq 1$ for $x \geq 1$. Assume that for all $d < n$ and $x \geq 1$, $T_{d+1}(x) \geq T_d(x) \geq 1$. If we now prove that $T_{n+1}(x) \geq T_n(x) \geq 1$, we are done by induction.

We have $T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x) = T_n(x) + [T_n(x) - T_{n-1}(x)] + (2x - 2)T_n(x)$. As $x \geq 1$ and by induction hypothesis $T_n(x) \geq T_{n-1}(x) \geq 1$, we obtain that $T_{n+1}(x) \geq T_n(x) \geq 1$. Thus for all $d \geq 0$ and $x \geq 1$, $T_{d+1}(x) \geq T_d(x) \geq 1$. ∎

Recall $p(x) = (1 + \gamma)\alpha \frac{T(x/\alpha)}{T(1+\gamma)} = (1 + \varepsilon/2)\sigma_{k+1}\frac{T(x/\alpha)}{T(1+\gamma)}$.

Lemma 17 *If $x \geq \alpha > 0$, then $p(x) \leq (1 + \gamma)\alpha \frac{3^q(x/\alpha)^q}{T(1+\gamma)}$.*

Proof By a standard property, the sum of absolute values of coefficients of the degree-q Chebyshev polynomial is bounded above by 3^q. Thus $T(x/\alpha) = \sum_{i=1}^{q} T_i(x/\alpha)^i \leq \sum_{i=1}^{q} |T_i|(x/\alpha)^i \leq (x/\alpha)^q \sum_{i=1}^{q} |T_i| \leq 3^q(x/\alpha)^q$, where we use the fact that $(x/\alpha) \geq 1$. Thus $p(x) = (1 + \gamma)\alpha T(x/\alpha)/T(1+\gamma) \leq (1+\gamma)\alpha 3^q(x/\alpha)^q/T(1+\gamma)$. ∎

Proof [Proof of Lemma 15] We bound $\sigma_1(p(M))$ and $\sigma_{k+1}(p(M))$ then infer an upper bound on $\frac{\sigma_1(p(M))}{\sigma_{k+1}(p(M))}$. Let $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d \geq 0$ be the singular values of the matrix M. Then we have that $|p(\sigma_1)|, |p(\sigma_2)|, \ldots, |p(\sigma_d)|$ are the singular values of the matrix $p(M)$. Consider any $i \leq k+1$. We have $\sigma_i \geq \sigma_{k+1} = \alpha$. Therefore,

$$p(\sigma_i) = (1+\gamma)\sigma_{k+1}\frac{T(\sigma_i/\sigma_{k+1})}{T(1+\gamma)} \geq \frac{(1+\gamma)\sigma_{k+1}}{T(1+\gamma)}.$$

Here we use Lemma 16 to lower bound the value of $T(\sigma_i/\sigma_{k+1})$ by 1. Therefore at least $k+1$ singular values of $p(M)$ are at least $\frac{(1+\gamma)\sigma_{k+1}}{T(1+\gamma)}$ which implies $\sigma_{k+1}(p(M)) \geq \frac{(1+\gamma)\sigma_{k+1}}{T(1+\gamma)}$.

Now for any $i \leq k+1$, $p(\sigma_i) \leq (1+\gamma)\sigma_{k+1}(3^q\kappa^q)/T(1+\gamma)$ by Lemma 17. For any $i \geq k+1$, we have that $\sigma_i \leq \sigma_{k+1}$ and $|p(\sigma_i)| = (1+\gamma)\sigma_{k+1}|T(\sigma_i/\sigma_{k+1})|/T(1+\gamma) \leq (1+\gamma)\sigma_{k+1}/T(1+\gamma)$ by a well known property of Chebyshev polynomials that $|T(x)| \leq 1$ for all $x \in [-1, 1]$. Therefore

$$\|p(M)\|_2 = \sigma_1(p(M)) = \max_i |p(\sigma_i(M))| \leq (1+\gamma)\sigma_{k+1}\frac{3^q\kappa^q}{T(1+\gamma)}. \tag{6}$$

Thus, $\sigma_1(p(M))/\sigma_{k+1}(p(M)) \leq 3^q\kappa^q$. ∎

We now bound the condition number of the matrix $p(M)G$ where G is a Gaussian matrix with k columns. We use results from [23] to bound the maximum and minimum singular values of G with $O(1)$ probability and then use the above lemma to obtain bounds on extreme singular values of $p(M)G$.

Lemma 18 *If $G \in \mathbb{R}^{d \times k}$ is a matrix of i.i.d. normal entries and $M \in \mathbb{R}^{n \times d}$ is a matrix such that $\sigma_1(M)/\sigma_{k+1}(M) = \kappa$, then with probability $\geq 4/5$,*

$$\kappa(p(M)G) = \sigma_{\max}(p(M)G)/\sigma_{\min}(p(M)G) \leq Ck3^q\kappa^q,$$

for an absolute constant $C > 0$.

Lemma 19 *If $A \in \mathbb{R}^{n \times d}$ is a matrix with $\sigma_1(A)/\sigma_k(A) \leq \kappa_1$ and $G \in \mathbb{R}^{d \times k}$ is a matrix with i.i.d. normal entries, then for d greater than a constant, with probability $\geq 4/5$, the matrix AG has full rank and has $\sigma_1(AG)/\sigma_k(AG) \leq Ck(\sigma_1(A)/\sigma_k(A))$ where $C > 0$ is an absolute constant.*

Proof Let $A = U\Sigma V^\mathsf{T}$ be the singular value decomposition of A with $U \in \mathbb{R}^{n \times n}, \Sigma \in \mathbb{R}^{n \times d}$ and $V^\mathsf{T} \in \mathbb{R}^{d \times d}$. Let $G' = V^\mathsf{T}G$. As rows of V^T are orthonormal and entries of G are i.i.d. normal random variables, we obtain that G' is a also a matrix of i.i.d. normal random variables of size $d \times k$. Let $\Sigma_k \in \mathbb{R}^{k \times d}$ be the first k rows of Σ. For any vector x,

$$\|AGx\|_2 = \|U\Sigma V^\mathsf{T}Gx\|_2 = \|\Sigma G'x\|_2 \geq \|\Sigma_k G'x\|_2.$$

Thus, $\min_{x:\|x\|_2=1}\|AGx\|_2 \geq \min_{x:\|x\|_2=1}\|\Sigma_k G'x\|_2$. We have that

$$\Pr[\sigma_{\min}(G') \leq \frac{1}{20C}(\sqrt{d} - \sqrt{k-1})] \leq \left(\frac{1}{20}\right)^{d-k+1} + e^{-cd}$$

for some absolute constants c and C by Theorem 1.1 of [22]. Thus for large enough d, with probability $\geq 9/10$, we have

$$\sigma_{\min}(G') \geq \frac{1}{20C}(\sqrt{d} - \sqrt{k-1}).$$

Thus $\min_{x:\|x\|_2=1}\|\Sigma_k G'x\|_2 \geq \sigma_{\min}(\Sigma_k)\sigma_{\min}(G') \geq \frac{\sigma_k(A)}{20C}(\sqrt{d} - \sqrt{k-1})$. Similarly, for large enough d, we have with probability $\geq 9/10$ that $\sigma_{\max}(G') \leq D(\sqrt{d} + \sqrt{k})$ for an absolute constant D by Proposition 2.4 of [23] and therefore $\max_{x:\|x\|_2=1}\|AGx\|_2 = \max_{x:\|x\|_2=1}\|\Sigma G'x\|_2 \leq D\sigma_1(A)(\sqrt{d} + \sqrt{k})$. Therefore with probability $\geq 4/5$,

$$\kappa(AG) = \frac{\sigma_{\max}(AG)}{\sigma_{\min}(AG)} \leq 20CD\frac{\sigma_1(A)}{\sigma_k(A)}\frac{\sqrt{d} + \sqrt{k}}{\sqrt{d} - \sqrt{k-1}}.$$

The maximum of this expression occurs at $d = k$ and is at most $4k$. Therefore with probability $\geq 4/5$, for d at least some constant, $\kappa(AG) \leq 40CDk(\sigma_1(A)/\sigma_k(A))$. ∎

Proof [Proof of Lemma 18] Using the above lemma, we have that with probability $\geq 4/5$,

$$\kappa(p(M)G) = \frac{\sigma_{\max}(p(M)G)}{\sigma_{\min}(p(M)G)} \leq Ck\frac{\sigma_1(p(M))}{\sigma_k(p(M))} \leq Ck\frac{\sigma_1(p(M))}{\sigma_{k+1}(p(M))}$$

for an absolute constant C. The last inequality follows from $\sigma_k(p(M)) \geq \sigma_{k+1}(p(M))$. From Lemma 15, we have $\frac{\sigma_1(p(M))}{\sigma_{k+1}(p(M))} \leq 3^q\kappa^q$. Therefore $\kappa(p(M)G) \leq Ck3^q\kappa^q$ with probability $\geq 4/5$. ∎

The bound on the condition number of $p(M)G$ enables us to conclude that if the Frobenius norm error between $p(M)G$ and a matrix Apx is small, the projection matrices onto the column spaces of the matrices $p(M)G$ and Apx are close. Specifically, we use the following lemma.

Lemma 20 *Let A and B be full column rank matrices such that $\|A - B\|_2 \leq \delta\|A\|_2$. Let $\kappa(A)$ denote the condition number of the matrix A i.e., $\kappa(A) = \sigma_{\max}(A)/\sigma_{\min}(A)$. Let U and V denote orthonormal basis for matrices A and B respectively. If $\delta \leq 1/(2\kappa(A)) \leq 1$, then $\|AA^+ - BB^+\|_2 = \|UU^\mathsf{T} - VV^\mathsf{T}\|_2 \leq 20\delta\kappa(A)^4$.*

Proof As A and B are full rank matrices, we have $A^+ = (A^\mathsf{T}A)^{-1}A^\mathsf{T}$ and $B^+ = (B^\mathsf{T}B)^{-1}B^\mathsf{T}$. Let $A - B = \Delta$. We have $\|\Delta\|_2 \leq \delta\|A\|_2$. We first have

$$
\begin{aligned}
\|AA^+ - BB^+\|_2 &= \|AA^+ - (A - \Delta)B^+\|_2 \\
&\leq \|A\|_2\|A^+ - B^+\|_2 + \|\Delta\|_2\|B^+\|_2 \\
&\leq \|A\|_2\|A^+ - B^+\|_2 + \frac{\|\Delta\|_2}{\sigma_{\min}(B)}.
\end{aligned}
$$

Note that $A^\mathsf{T}A = (B + \Delta)^\mathsf{T}(B + \Delta) = B^\mathsf{T}B + \Delta^\mathsf{T}B + B^\mathsf{T}\Delta + \Delta^\mathsf{T}\Delta$. Now,

$$
\begin{aligned}
\|A^+ - B^+\|_2 &= \|(A^\mathsf{T}A)^{-1}A^\mathsf{T} - (B^\mathsf{T}B)^{-1}B^\mathsf{T}\|_2 \\
&= \|(A^\mathsf{T}A)^{-1}A^\mathsf{T} - (B^\mathsf{T}B)^{-1}(A^\mathsf{T} - \Delta^\mathsf{T})\|_2 \\
&\leq \|(A^\mathsf{T}A)^{-1} - (B^\mathsf{T}B)^{-1}\|_2\|A\|_2 + \|(B^\mathsf{T}B)^{-1}\|_2\|\Delta\|_2 \\
&\leq \|(A^\mathsf{T}A)^{-1} - (B^\mathsf{T}B)^{-1}\|_2\|A\|_2 + \frac{\|\Delta\|_2}{\sigma_{\min}(B)^2}.
\end{aligned}
$$

We finally bound $\|(A^\mathsf{T}A)^{-1} - (B^\mathsf{T}B)^{-1}\|_2$.

$$
\begin{aligned}
\|(A^\mathsf{T}A)^{-1} - (B^\mathsf{T}B)^{-1}\|_2 &\leq \frac{1}{\sigma_{\min}(A^\mathsf{T}A)}\|(A^\mathsf{T}A)((A^\mathsf{T}A)^{-1} - (B^\mathsf{T}B)^{-1})\|_2 \\
&\leq \frac{1}{\sigma_{\min}(A^\mathsf{T}A)}\|I - (A^\mathsf{T}A)(B^\mathsf{T}B)^{-1}\|_2 \\
&\leq \frac{1}{\sigma_{\min}(A^\mathsf{T}A)}\|I - (B^\mathsf{T}B + \Delta^\mathsf{T}B + B^\mathsf{T}\Delta + \Delta^\mathsf{T}\Delta)(B^\mathsf{T}B)^{-1}\|_2 \\
&\leq \frac{1}{\sigma_{\min}(A^\mathsf{T}A)}\|I - I - (\Delta^\mathsf{T}B + B^\mathsf{T}\Delta + \Delta^\mathsf{T}\Delta)(B^\mathsf{T}B)^{-1}\|_2 \\
&\leq \frac{2\|\Delta\|_2\|B\|_2 + \|\Delta\|_2^2}{\sigma_{\min}(A^\mathsf{T}A)\sigma_{\min}(B^\mathsf{T}B)}.
\end{aligned}
$$

We therefore obtain

$$
\begin{aligned}
\|AA^+ - BB^+\|_2 &\leq \|A\|_2^2\|(A^\mathsf{T}A)^{-1} - (B^\mathsf{T}B)^{-1}\|_2 + \frac{\|\Delta\|_2\|A\|_2}{\sigma_{\min}(B^\mathsf{T}B)} + \frac{\|\Delta\|_2}{\sigma_{\min}(B)} \\
&\leq \frac{\|A\|_2^2}{\sigma_{\min}(A^\mathsf{T}A)}\frac{2\|\Delta\|_2\|B\|_2 + \|\Delta\|_2^2}{\sigma_{\min}(B^\mathsf{T}B)} + \frac{\|\Delta\|_2\|A\|_2}{\sigma_{\min}(B^\mathsf{T}B)} + \frac{\|\Delta\|_2}{\sigma_{\min}(B)}.
\end{aligned}
$$

As $\|A - B\|_2 \leq \delta\|A\|_2$, we get that $(1 - \delta)\|A\|_2 \leq \|B\|_2 \leq (1 + \delta)\|A\|_2$. We also have that $\sigma_{\min}(B) \geq \sigma_{\min}(A) - \|A - B\|_2 \geq \|A\|_2/\kappa(A) - \delta\|A\|_2 \geq \|A\|_2/2\kappa(A) = \sigma_{\min}(A)/2$ if $\delta < 1/2\kappa(A)$. We can therefore conclude that $\|AA^+ - BB^+\|_2 \leq 20\delta\kappa(A)^4$. ∎

The condition that δ must be less than $1/2\kappa(A)$ in the above lemma makes sense as otherwise $20\delta\kappa(A)^4 \geq 10\kappa(A)^3 \geq 10$, which is a trivial upper bound on the norm.

Now we construct a matrix Apx that has its columns spanned by K' and is close to the matrix $p(M)G$. Using the bound on the condition number of the matrix $p(M)G$, we can conclude that projection matrices onto column spans of Apx and $p(M)G$, respectively, are close.

Recall $p(x)$ from (5). For q odd, the Chebyshev polynomial of degree q contains only odd degree monomials. So we have $T(x) = T_q x^q + T_{q-2} x^{q-2} + \ldots + T_1 x$ and therefore the polynomial $p(x) = \frac{(1+\gamma)\alpha}{T(1+\gamma)} \left(\frac{T_q}{\alpha^q} x^q + \frac{T_{q-2}}{\alpha^{q-2}} x^{q-2} + \cdots + \frac{T_1}{\alpha_1} x \right)$ which implies

$$p(M)G = \frac{(1+\gamma)\alpha}{T(1+\gamma)} \left(\frac{T_q}{\alpha^q} (MM^\mathsf{T})^{(q-1)/2} MG + \cdots + \frac{T_1}{\alpha_1} MG \right).$$

We now define

$$Apx = \frac{(1+\gamma)\alpha}{T(1+\gamma)} \left(\frac{T_q}{\alpha^q} (MM^\mathsf{T})^{\circ(q-1)/2} M \circ G + \cdots + \frac{T_1}{\alpha_1} M \circ G \right). \tag{7}$$

Clearly, the matrix Apx is spanned by the columns of the matrix K'. Using Lemma 14 and properties of Gaussian matrices, the following lemma bounds $\|Apx - p(M)G\|_2$.

Lemma 21 *For the matrices $p(M)G$ and Apx defined above, we have with probability $\geq 3/5$*

$$\|p(M)G - Apx\|_2 \leq \|p(M)G - Apx\|_\mathsf{F} \leq 64 C \varepsilon_\circ k^{3/2} (3\sqrt{2}\kappa)^q \|p(M)G\|_2.$$

Proof By the triangle inequality,

$$\|p(M)G - Apx\|_\mathsf{F}$$
$$\leq \frac{(1+\gamma)\alpha}{T(1+\gamma)} \sum_{\text{odd } i \leq q} \frac{|T_i|}{\alpha^i} \|(MM^\mathsf{T})^{(i-1)/2} MG - (MM^\mathsf{T})^{\circ(i-1)/2} M \circ G\|_\mathsf{F}$$
$$\leq \frac{(1+\gamma)\alpha}{T(1+\gamma)} \sum_{\text{odd } i \leq q} \frac{|T_i|}{\alpha^i} E_{i,G}$$
$$\leq \frac{(1+\gamma)\alpha}{T(1+\gamma)} \sum_{\text{odd } i \leq q} \frac{|T_i|}{\alpha^i} 8\varepsilon_\circ (2^{i/2} \|M\|_2^i \|G\|_\mathsf{F}) \qquad \text{(Lemma 14)}$$
$$\leq \frac{(1+\gamma)\sigma_{k+1}(M)}{T(1+\gamma)} 8\varepsilon_\circ \|G\|_\mathsf{F} \sum_{\text{odd } i \leq q} |T_i| (\sqrt{2}\kappa)^i \qquad (\alpha = \sigma_{k+1}(M))$$
$$\leq \frac{(1+\gamma)\sigma_{k+1}(M)}{T(1+\gamma)} 8\varepsilon_\circ \|G\|_\mathsf{F} (3\sqrt{2}\kappa)^q \qquad (\sum_i |T_i| \leq 3^q)$$
$$\leq \|p(M)\|_2 8\varepsilon_\circ \|G\|_\mathsf{F} (3\sqrt{2}\kappa)^q. \qquad \text{(Equation 6)}$$

We also condition on the following events both of which hold simultaneously with probability $\geq 4/5$.

- $\|G\|_\mathsf{F} \leq 4\sqrt{dk}$, and

- $\|p(M)G\|_2 \geq (1/C)\|p(M)\|_2(\sqrt{d} - \sqrt{k-1}) \geq (1/2C)\|p(M)\|_2\sqrt{d}$.

23

Thus, with probability $\geq 4/5$, if $d \geq 4k$,

$$\|p(M)G - Apx\|_{\mathsf{F}} \leq \|p(M)\|_2 (32\varepsilon_\circ)\sqrt{dk}(3\sqrt{2}\kappa)^q \leq 64C\varepsilon_\circ\sqrt{k}(3\sqrt{2}\kappa)^q\|p(M)G\|_2.$$

If $k \leq d \leq 4k$, then $\|p(M)G\|_2 \geq (1/C)\|p(M)\|_2(\sqrt{d} - \sqrt{k-1}) \geq (1/2C)\|p(M)\|_2(1/\sqrt{k})$ and $\|p(M)G - Apx\|_{\mathsf{F}} \leq 64C\varepsilon_\circ k^{3/2}(3\sqrt{2}\kappa)^q\|p(M)G\|_2$. ∎

Let $Y_1 \in \mathbb{R}^{n \times k}$ be an orthonormal basis for the column span of $p(M)G$ and $Y \in \mathbb{R}^{n \times k}$ be an orthonormal basis for the matrix Apx. We now have from Lemmas 20 and 21 that

$$\|YY^{\mathsf{T}} - Y_1Y_1^{\mathsf{T}}\|_2 \leq O(\varepsilon_\circ k^{3/2}(3\sqrt{2}\kappa)^q\kappa(p(M)G)^4) = O(\varepsilon_\circ k^{3/2}(3\sqrt{2}\kappa)^q(k^4 3^{4q}\kappa^{4q}))$$
$$= \varepsilon_\circ C^q k^6 \kappa^{5q}$$

for some constant C. Let $\delta := \varepsilon_\circ C^q k^6 \kappa^{5q}$. Hence

$$\|Y_1Y_1^{\mathsf{T}} - YY^{\mathsf{T}}\|_2 \leq \delta. \tag{8}$$

For $l \leq k$ such that $\sigma_l(M) \geq (1+\varepsilon)\sigma_{k+1}(M)$, let $\mathcal{E}_l = \|[M]_l\|_{\mathsf{F}}^2 - \|Y_1Y_1^{\mathsf{T}}[M]_l\|_{\mathsf{F}}^2$ and $\mathcal{E}_l' = \|[M]_l\|_{\mathsf{F}}^2 - \|YY^{\mathsf{T}}[M]_l\|_{\mathsf{F}}^2$. Musco and Musco [18, Equation 7] show that

$$\mathcal{E}_l = \|[M]_l\|_{\mathsf{F}}^2 - \|Y_1Y_1^{\mathsf{T}}[M]_l\|_{\mathsf{F}}^2 \leq (\varepsilon/2)\sigma_{k+1}(M)^2.$$

Bounding $\mathcal{E}_l$ is one of the important steps in the analysis of [18]. We obtain a similar bound on $\mathcal{E}_l'$. We further show that if $M_{K',l}$ is the best rank l Frobenius norm approximation of M in colspan(K'), then $\|[M]_l\|_{\mathsf{F}}^2 - \|M_{K',l}\|_{\mathsf{F}}^2 \leq (3\varepsilon/4)\sigma_{k+1}(M)^2$, showing that there is a very good rank-l approximation for M in colspan(K'). We have the following lemma.

Lemma 22 *Given a matrix A and a parameter k, let Y_1 be an orthonormal basis for a k dimensional subspace such that $\mathcal{E}_l = \|[M]_l\|_{\mathsf{F}}^2 - \|Y_1Y_1^{\mathsf{T}}[M]_l\|_{\mathsf{F}}^2 \leq (\varepsilon/2)\sigma_{k+1}^2$ for all $l \leq k$ satisfying $\sigma_l(M) \geq (1+\varepsilon)\sigma_{k+1}(M)$. If Y is an orthonormal basis for another k dimensional subspace for which $\|YY^{\mathsf{T}} - Y_1Y_1^{\mathsf{T}}\|_2 \leq \varepsilon/(16\kappa^2\sqrt{k})$, where $\kappa = \sigma_1(M)/\sigma_{k+1}(M)$, then for all such l,*

$$\mathcal{E}_l' = \|[M]_l\|_{\mathsf{F}}^2 - \|YY^{\mathsf{T}}[M]_l\|_{\mathsf{F}}^2 \leq (3\varepsilon/4)\sigma_{k+1}^2.$$

There also exists a matrix Y^l with l orthonormal columns with colspan(Y^l) $\subseteq$ colspan(K') such that $\|[M]_l\|_{\mathsf{F}}^2 - \|Y^l(Y^l)^{\mathsf{T}}M\|_{\mathsf{F}}^2 \leq (3\varepsilon/4)\sigma_{k+1}^2$.

Proof For any $1 > \varepsilon_{\mathsf{s}} > 0$

$$\|Y_1Y_1^{\mathsf{T}}M_l\|_{\mathsf{F}}^2 \leq (1+\varepsilon_{\mathsf{s}})\|YY^{\mathsf{T}}M_l\|_{\mathsf{F}}^2 + (1 + \frac{1}{\varepsilon_{\mathsf{s}}})\|(YY^{\mathsf{T}} - Y_1Y_1^{\mathsf{T}})M_l\|_{\mathsf{F}}^2$$
$$\leq (1+\varepsilon_{\mathsf{s}})\|YY^{\mathsf{T}}M_l\|_{\mathsf{F}}^2 + (2/\varepsilon_{\mathsf{s}})2k\delta^2\sigma_1(M)^2.$$

The last inequality follows from the fact that $YY^{\mathsf{T}} - Y_1Y_1^{\mathsf{T}}$ has rank at most $2k$. Therefore

$$\|YY^{\mathsf{T}}M_l\|_{\mathsf{F}}^2 \geq \frac{1}{1+\varepsilon_{\mathsf{s}}}\|Y_1Y_1^{\mathsf{T}}M_l\|_{\mathsf{F}}^2 - \frac{4k\sigma_1(M)^2}{\varepsilon_{\mathsf{s}}}\delta^2$$

which implies that

$$\mathcal{E}_l' = \|M_l\|_{\mathsf{F}}^2 - \|YY^{\mathsf{T}}M_l\|_{\mathsf{F}}^2$$

24

$$\leq \|M_l\|_{\mathsf{F}}^2 - \frac{1}{1+\varepsilon_{\mathrm{s}}}\|Y_1 Y_1^{\mathsf{T}} M_l\|_{\mathsf{F}}^2 + \frac{4k\sigma_1(M)^2}{\varepsilon_{\mathrm{s}}}\delta^2$$

$$\leq \frac{1}{1+\varepsilon_{\mathrm{s}}}(\|M_l\|_{\mathsf{F}}^2 - \|Y_1 Y_1^{\mathsf{T}} M_l\|_{\mathsf{F}}^2) + \varepsilon_{\mathrm{s}}\|M_l\|_{\mathsf{F}}^2 + \frac{4k\sigma_1(M)^2}{\varepsilon_{\mathrm{s}}}\delta^2$$

$$\leq \frac{1}{1+\varepsilon_{\mathrm{s}}}\frac{\varepsilon}{2}\sigma_{k+1}(M)^2 + \varepsilon_{\mathrm{s}}k\sigma_1(M)^2 + \frac{4k\sigma_1(M)^2}{\varepsilon_{\mathrm{s}}}\delta^2.$$

Picking $\varepsilon_{\mathrm{s}} = \varepsilon/(8k\kappa^2)$ and if $\delta \leq \varepsilon/(16\kappa^2\sqrt{k})$, we obtain that

$$\mathcal{E}'_l = \|M_l\|_{\mathsf{F}}^2 - \|YY^{\mathsf{T}} M_l\|_{\mathsf{F}}^2 \leq \frac{3\varepsilon}{4}\sigma_{k+1}^2.$$

Recall here that $\kappa = \sigma_1(M)/\sigma_{k+1}(M)$. The matrix $YY^{\mathsf{T}} M_l$ is a rank l approximation for matrix M inside the column span of Y and hence in the column span of K'. Let Y^l be a rank l matrix that forms a basis for the best rank l approximation of M inside the column space of K' i.e.,

$$\min_{\text{rank-}l\ B:\text{colspan}(B)\subseteq\text{colspan}(K')}\|M - B\|_{\mathsf{F}}^2 = \|M - Y^l Y^l M\|_{\mathsf{F}}^2.$$

From Lemma 3, note that if $\bar{U}\bar{\Sigma}^2\bar{V}^{\mathsf{T}}$ is the singular value decomposition of the matrix $Q'^{\mathsf{T}}MM^{\mathsf{T}}Q'$ (recall Q' denotes an orthonormal basis for the matrix K'), then $Y^l = Q'\bar{U}_l$ where $\bar{U}_l$ denotes the first l columns of the matrix $\bar{U}$. By the optimality of Y^l, $\|M - Y^l(Y^l)^{\mathsf{T}} M\|_{\mathsf{F}}^2 \leq \|M - YY^{\mathsf{T}} M_l\|_{\mathsf{F}}^2$ which implies that $\|YY^{\mathsf{T}} M_l\|_{\mathsf{F}}^2 \leq \|Y^l(Y^l)^{\mathsf{T}} M\|_{\mathsf{F}}^2$. Thus $\|M_l\|_{\mathsf{F}}^2 - \|Y^l(Y^l)^{\mathsf{T}} M\|_{\mathsf{F}}^2 \leq \|M_l\|_{\mathsf{F}}^2 - \|YY^{\mathsf{T}} M_l\|_{\mathsf{F}}^2 = \mathcal{E}'_l \leq (3\varepsilon/4)\sigma_{k+1}^2$. ∎

The proof also shows that if $\bar{U}\bar{\Sigma}^2\bar{U}^{\mathsf{T}}$ is the singular value decomposition of the positive semi-definite matrix $Q'^{\mathsf{T}}MM^{\mathsf{T}}Q'$, then $Y^l = Q'\bar{U}_l$ where $\bar{U}_l$ denotes the matrix that contains the first l columns of $\bar{U}$. Let $m \leq k$ be the largest integer for which $\sigma_m(M) \geq (1+\varepsilon)\sigma_{k+1}(M)$. From the above lemma, the matrix Y^m satisfies $\|M_l\|_{\mathsf{F}}^2 - \|Y^m(Y^m)^{\mathsf{T}} M\|_{\mathsf{F}}^2 \leq (3\varepsilon/4)\sigma_{k+1}(M)^2$. We later show that this implies $\|M - Y^m(Y^m)^{\mathsf{T}} M\|_2 \leq (1 + 3\varepsilon/2)\sigma_{k+1}(M)$. Unfortunately, we cannot compute the matrix $Q'^{\mathsf{T}}MM^{\mathsf{T}}Q'$ exactly as we only have access to an oracle that computes vector products with matrices M, M^{T} approximately. Nevertheless we show that we can compute a matrix $\hat{Y}^m$ based on an approximation to the matrix $Q'^{\mathsf{T}}MM^{\mathsf{T}}Q'$ and it still satisfies the desired guarantees approximately.

First we have the following lemma that shows if a subspace Y^m is a good approximation for Frobenius norm low rank approximation of M in m dimensions, then the subspace Y^m is also a good subspace for spectral norm rank-k approximation of matrix M. It also shows that even if $\hat{Y}^m$ only approximately satisfies the properties of Y^m, the matrix $\hat{Y}^m$ spans a good low rank approximation for M.

Lemma 23 *Given an arbitrary matrix M, if an orthonormal basis Y^m to an m-dimensional subspace, where $m \leq k$ is the largest integer such that $\sigma_m(M) \geq (1+\varepsilon)\sigma_{k+1}(M)$, satisfies*

$$\|M_m\|_{\mathsf{F}}^2 - \|Y^m(Y^m)^{\mathsf{T}} M\|_{\mathsf{F}}^2 \leq \varepsilon\sigma_{k+1}(M)^2,$$

then $\|M - Y^m(Y^m)^{\mathsf{T}} M\|_2 \leq (1 + 2\varepsilon)\sigma_{k+1}(M)$. Additionally if $\hat{Y}^m$ is a matrix with m orthonormal columns such that

$$\|M - \hat{Y}^m(\hat{Y}^m)^{\mathsf{T}} M\|_{\mathsf{F}}^2 \leq \|M - Y^m(Y^m)^{\mathsf{T}} M\|_{\mathsf{F}}^2 + \delta,$$

then $\|M - \hat{Y}^m(\hat{Y}^m)^{\mathsf{T}} M\|_2 \leq (1 + 2\varepsilon)\sigma_{k+1}(M) + \sqrt{\delta}$.

Proof As $\|M_m\|_\mathsf{F}^2 - \|Y^m(Y^m)^\mathsf{T}M\|_\mathsf{F}^2 = \|M\|_\mathsf{F}^2 - \|M - M_m\|_\mathsf{F}^2 - \|Y^m(Y^m)^\mathsf{T}M\|_\mathsf{F}^2 = \|M - Y^m(Y^m)^\mathsf{T}M\|_\mathsf{F}^2 - \|M - M_m\|_\mathsf{F}^2$, we obtain that

$$\|M - Y^m(Y^m)^\mathsf{T}M\|_\mathsf{F}^2 \leq \|M - M_m\|_\mathsf{F}^2 + \varepsilon\sigma_{k+1}(M)^2.$$

As an additive error in Frobenius norm translates to additive error in spectral norm for the above case (see Theorem 3.2 from [8]), we obtain

$$\|M - Y^m(Y^m)^\mathsf{T}M\|_2^2 \leq \|M - M_m\|_2^2 + \varepsilon\sigma_{k+1}(M)^2 \leq \sigma_{m+1}(M)^2 + \varepsilon\sigma_{k+1}(M)^2$$
$$\leq (1 + 4\varepsilon)\sigma_{k+1}(M)^2.$$

Thus $\|M - Y^m(Y^m)^\mathsf{T}M\|_2 \leq (1 + 2\varepsilon)\sigma_{k+1}(M)$. Similarly, we have that

$$\|M - \hat{Y}^m(\hat{Y}^m)^\mathsf{T}M\|_\mathsf{F}^2 \leq \|M - M_m\|_\mathsf{F}^2 + \varepsilon\sigma_{k+1}(M)^2 + \delta$$

which implies that

$$\|M - \hat{Y}^m(\hat{Y}^m)^\mathsf{T}M\|_2^2 \leq \|M - M_m\|_2^2 + \varepsilon\sigma_{k+1}(M)^2 + \delta \leq (1 + 4\varepsilon)\sigma_{k+1}(M)^2 + \delta$$

which shows $\|M - \hat{Y}^m(\hat{Y}^m)^\mathsf{T}M\|_2 \leq (1 + 2\varepsilon)\sigma_{k+1}(M) + \sqrt{\delta}$. ∎

The above lemma shows that we need only compute a matrix $\hat{Y}^m$ such that $\|M - \hat{Y}^m(\hat{Y}^m)^\mathsf{T}M\|_\mathsf{F}^2 \approx \|M - Y^m(Y^m)^\mathsf{T}M\|_\mathsf{F}^2$.

We show that using an approximation to matrix $Q'^\mathsf{T}MM^\mathsf{T}Q'^\mathsf{T}$ we can compute such a matrix $\hat{Y}^m$ which shows that $\|M - \hat{Y}^m(\hat{Y}^m)^\mathsf{T}M\|_2 \leq (1 + O(\varepsilon))\sigma_{k+1}(M)$. As the value of $m \leq k$ is not known, we further show that we can compute a matrix $\hat{Y}^k$ with k orthonormal columns such that $\mathrm{colspan}(M) \supseteq \mathrm{colspan}(K') \supseteq \mathrm{colspan}(\hat{Y}^k) \supseteq \mathrm{colspan}(\hat{Y}^m)$. Therefore we can conclude that $\|M - \hat{Y}^k(\hat{Y}^k)^\mathsf{T}M\|_2 \leq \|M - \hat{Y}^k(\hat{Y}^k)^\mathsf{T}M\|_2 \leq (1 + O(\varepsilon))\sigma_{k+1}(M)$. We thus have our final result for low rank approximation.

C.3. Proof of Theorem 8

Computing top k singular vectors of the matrix $Q'^\mathsf{T}MM^\mathsf{T}Q'$ We now show that if $\hat{Y}^m$ are top m singular vectors of the matrix $Q'^\mathsf{T}((MM^\mathsf{T}) \circ Q')$, then

$$\|M - \hat{Y}^m(\hat{Y}^m)^\mathsf{T}M\|_\mathsf{F}^2 \approx \|M - Y^m(Y^m)^\mathsf{T}M\|_\mathsf{F}^2.$$

Lemma 24 *If Z_m are the top m orthonormal eigenvectors of the matrix MM^T, then for any matrix Y with m orthonormal columns,*

$$\mathrm{tr}(Z_m^\mathsf{T}MM^\mathsf{T}Z_m) \geq \mathrm{tr}(Y^\mathsf{T}MM^\mathsf{T}Y).$$

Proof We have $\mathrm{tr}(Z_m^\mathsf{T}MM^\mathsf{T}Z_m) = \|Z_mZ_m^\mathsf{T}M\|_\mathsf{F}^2$ and $\mathrm{tr}(Y^\mathsf{T}MM^\mathsf{T}Y) = \|YY^\mathsf{T}M\|_\mathsf{F}^2$. We are given that Z_m are top m eigenvectors of the matrix MM^T and therefore Z_m are top m singular vectors of the matrix M. Therefore for any matrix Y with m orthonormal columns, we have that $\|Z_mZ_m^\mathsf{T}M\|_\mathsf{F}^2 \geq \|YY^\mathsf{T}M\|_\mathsf{F}^2$ and therefore that $\mathrm{tr}(Z_m^\mathsf{T}MM^\mathsf{T}Z_m) \geq \mathrm{tr}(Y^\mathsf{T}MM^\mathsf{T}Y)$. ∎

Lemma 25 *Let M be a matrix and Q be an orthonormal basis for an arbitrary r dimensional space. Let B be a positive semi-definite matrix such that $B - Q^\mathsf{T} M M^\mathsf{T} Q = \Delta$. Let Z be a matrix whose columns are the top k eigenvectors of the matrix B. Then if Z_m denotes the matrix with first m columns of Z for $m = 1, \ldots, k$ we have*

$$\|M - (QZ_m)(QZ_m)^\mathsf{T} M\|_\mathsf{F}^2 \le \|M - Q(Q^\mathsf{T} M)_m\|_\mathsf{F}^2 + 2m\|\Delta\|_\mathsf{F}.$$

Proof Let Z^* be the matrix whose columns are the top k eigenvectors of the matrix $Q^\mathsf{T} M M^\mathsf{T} Q$ and Z_m^* be the first m columns of Z^*. Thus $Q(Q^\mathsf{T} M)_m = Q(Z_m^*(Z_m^*)^\mathsf{T} Q^\mathsf{T} M) = (QZ_m^*)(QZ_m^*)^\mathsf{T} M$. Now,

$$
\begin{aligned}
\|(QZ_m)(QZ_m)^\mathsf{T} M\|_\mathsf{F}^2 &= \|(QZ_m)^\mathsf{T} M\|_\mathsf{F}^2 \\
&= \operatorname{tr}(Z_m^\mathsf{T} Q^\mathsf{T} M M^\mathsf{T} Q Z_m) \\
&= \operatorname{tr}(Z_m^\mathsf{T}(Q^\mathsf{T} M M^\mathsf{T} Q + \Delta)Z_m) - \operatorname{tr}(Z_m^\mathsf{T} \Delta Z_m) \\
&= \operatorname{tr}(Z_m^\mathsf{T} B Z_m) - \operatorname{tr}(Z_m^\mathsf{T} \Delta Z_m) \\
&\ge \operatorname{tr}((Z_m^*)^\mathsf{T} B Z_m^*) - m\|\Delta\|_\mathsf{F} \\
&\quad (\text{Since } \operatorname{tr}(Z_m^\mathsf{T} \Delta Z_m) = \operatorname{tr}(\Delta Z_m Z_m^\mathsf{T}) \le \|\Delta\|_\mathsf{F}\|Z_m Z_m^\mathsf{T}\|_\mathsf{F} \le \|\Delta\|_\mathsf{F} \cdot m) \\
&= \operatorname{tr}((Z_m^*)^\mathsf{T}(Q^\mathsf{T} M M^\mathsf{T} Q)Z_m^*) - \operatorname{tr}((Z_m^*)^\mathsf{T} \Delta Z_m^*) - m\|\Delta\|_\mathsf{F} \\
&= \operatorname{tr}(QZ_m^*(Z_m^*)^\mathsf{T} Q^\mathsf{T} M M^\mathsf{T} Q Z_m^*(Z_m^*)^\mathsf{T} Q^\mathsf{T}) - \operatorname{tr}((Z_m^*)^\mathsf{T} \Delta Z_m^*) - m\|\Delta\|_\mathsf{F} \\
&\ge \|(QZ_m^*)(QZ_m^*)^\mathsf{T} M\|_\mathsf{F}^2 - 2m\|\Delta\|_\mathsf{F}.
\end{aligned}
$$

Thus,

$$\|M - (QZ_m)(QZ_m)^\mathsf{T} M\|_\mathsf{F}^2 \le \|M - (QZ_m^*)(QZ_m^*)^\mathsf{T} M\|_\mathsf{F}^2 + 2m\|\Delta\|_\mathsf{F},$$

which concludes the proof. ∎

Hence if $\widetilde{Apx}$ is a positive semi-definite matrix such that $\|\widetilde{Apx} - Q'^\mathsf{T} M M^\mathsf{T} Q'\|_\mathsf{F}$ is small and if Z_m denotes the top m singular vectors of the matrix $\widetilde{Apx}$, we can conclude by Lemma 23 that $\|M - (Q'Z_m)(Q'Z_m)^\mathsf{T} M\|_2$ is close to $\sigma_{k+1}(M)$.

We now show that we can compute such a matrix $\widetilde{Apx}$. Let $\Xi = Q'^\mathsf{T}((MM^\mathsf{T}) \circ Q')$ (recall that $\circ$ denotes matrix multiplication using the noisy oracle). Let $\widetilde{Apx} = \operatorname{psd}((\Xi + \Xi^\mathsf{T})/2)$. Then the following lemma shows that $\widetilde{Apx}$ is close to $Q'^\mathsf{T} M M^\mathsf{T} Q'$.

Lemma 26 *Given matrices $M \in \mathbb{R}^{n \times d}$ and $Q' \in \mathbb{R}^{n \times t}$ where Q' is a matrix with t orthonormal columns, if for all vectors v, v', $\|M \circ v - Mv\|_2 \le \varepsilon_\circ\|M\|_2\|v\|_2$ and $\|M^\mathsf{T} \circ v' - M^\mathsf{T} v'\|_2 \le \varepsilon_\circ\|M\|_2\|v'\|_2$, and $\Xi := Q'^\mathsf{T}(MM^\mathsf{T}) \circ Q'$, then*

$$\|\operatorname{psd}((\Xi + \Xi^\mathsf{T})/2) - Q'^\mathsf{T} M M^\mathsf{T} Q'\|_\mathsf{F} \le (6\varepsilon_\circ\|M\|_2^2)\sqrt{t}.$$

Let $\widetilde{Apx} = \operatorname{psd}((\Xi + \Xi^\mathsf{T})/2)$. The matrix $\widetilde{Apx}$ can be computed in time $O(2tT(\varepsilon_\circ) + t^3)$.

Proof Let k_i be the ith column of the matrix K' and $E_i = \|Q'^\mathsf{T}(MM^\mathsf{T}) \circ k_i - Q'^\mathsf{T}(MM^\mathsf{T})k_i\|_2$. Then

$$E_i = \|Q'^\mathsf{T}(MM^\mathsf{T}) \circ k_i - Q'^\mathsf{T}(MM^\mathsf{T})k_i\|_2$$

$$\leq \|(MM^{\mathsf{T}}) \circ k_i - (MM^{\mathsf{T}})k_i\|_2$$

$$= \|M \circ (M^{\mathsf{T}} \circ k_i) - M(M^{\mathsf{T}}k_i)\|_2$$

$$\leq \|M \circ (M^{\mathsf{T}} \circ k_i) - M(M^{\mathsf{T}} \circ k_i) + M(M^{\mathsf{T}} \circ k_i) - M(M^{\mathsf{T}}k_i)\|_2$$

$$\leq \|M \circ (M^{\mathsf{T}} \circ k_i) - M(M^{\mathsf{T}} \circ k_i)\|_2 + \|M(M^{\mathsf{T}} \circ k_i) - M(M^{\mathsf{T}}k_i)\|_2$$

$$\leq \varepsilon_\circ \|M\|_2 \|M^{\mathsf{T}} \circ k_i\|_2 + \|M\|_2 \varepsilon_\circ \|M\|_2 \|k_i\|_2$$

$$\leq \varepsilon_\circ \|M\|_2 (\|M^{\mathsf{T}}k_i\|_2 + \varepsilon_\circ \|M\|_2 \|k_i\|_2) + \|M\|_2^2 \varepsilon_\circ \|k_i\|_2$$

$$\leq 3\varepsilon_\circ \|M\|_2^2. \qquad\qquad\qquad \text{(Since } \|k_i\|_2 = 1)$$

Thus $\|Q'^{\mathsf{T}}MM^{\mathsf{T}}Q' - \Xi\|_{\mathsf{F}}^2 = \sum_{i=1}^{t} \|Q'^{\mathsf{T}}MM^{\mathsf{T}}k_i - Q'^{\mathsf{T}}(MM^{\mathsf{T}}) \circ k_i\|_2^2 \leq (3\varepsilon_\circ \|M\|_2^2)^2 t$ which implies that $\|Q'^{\mathsf{T}}MM^{\mathsf{T}}Q' - \Xi\|_{\mathsf{F}} \leq (3\varepsilon_\circ \|M\|_2^2)\sqrt{t}$. Now as $Q'^{\mathsf{T}}MM^{\mathsf{T}}Q'$ is a symmetric matrix, $\|Q'^{\mathsf{T}}MM^{\mathsf{T}}Q' - (\Xi + \Xi^{\mathsf{T}})/2\|_{\mathsf{F}} \leq (3\varepsilon_\circ \|M\|_2^2)\sqrt{t}$. As $Q'^{\mathsf{T}}MM^{\mathsf{T}}Q'$ is itself a positive semidefinite matrix,

$$\|\mathrm{psd}((\Xi + \Xi^{\mathsf{T}})/2) - (\Xi + \Xi^{\mathsf{T}})/2\|_{\mathsf{F}} \leq \|Q'^{\mathsf{T}}MM^{\mathsf{T}}Q' - (\Xi + \Xi^{\mathsf{T}})/2\|_{\mathsf{F}} \leq (3\varepsilon_\circ \|M\|_2^2)\sqrt{t}.$$

Finally, by the triangle inequality we obtain that $\|Q'^{\mathsf{T}}MM^{\mathsf{T}}Q' - \widetilde{Apx}\|_{\mathsf{F}} = \|Q'^{\mathsf{T}}MM^{\mathsf{T}}Q' - \mathrm{psd}((\Xi + \Xi^{\mathsf{T}})/2)\|_{\mathsf{F}} \leq 6\varepsilon_\circ \|M\|_2^2 \sqrt{t}$. The time required to compute matrix Ξ is $2tT(\varepsilon_\circ) + nt^2$ and $\mathrm{psd}((\Xi + \Xi^{\mathsf{T}})/2)$ can be computed in time $O(t^3)$. Thus the matrix $\widetilde{Apx}$ can be computed in time $O(2tT(\varepsilon_\circ) + t^3)$. ∎

Proof [Proof of Theorem 8] Let $q = O((1/\sqrt{\varepsilon}) \log(d/\varepsilon))$. Algorithm 1 computes the Krylov Subspace K' with

$$\varepsilon_\circ = \frac{\varepsilon}{16\kappa^{2+5q}k^7 C^q}$$

for an absolute constant C. Let Y_1 be an orthonormal basis for $p(M)G$ and Y be an orthonormal basis for the matrix Apx(defined in (7)). Then by (8) we have that $\|YY^{\mathsf{T}} - Y_1 Y_1^{\mathsf{T}}\|_2 \leq \varepsilon/(16\kappa^2 \sqrt{k})$. If $m \leq k$ is the largest integer such that $\sigma_m(M) \geq (1 + \varepsilon)\sigma_{k+1}(M)$, by Lemma 22, there exists a d dimensional subspace Y^m inside the column span of K' such that

$$\|M_m\|_{\mathsf{F}}^2 - \|Y^m(Y^m)^{\mathsf{T}}A\|_{\mathsf{F}}^2 \leq (3\varepsilon/4)\sigma_{k+1}^2.$$

If Ξ is now computed with $\varepsilon_o = \varepsilon^2/(48\kappa^2(\sqrt{qk})k)$, then by Lemma 26,

$$\|Q'^{\mathsf{T}}MM^{\mathsf{T}}Q' - \widetilde{Apx}\|_{\mathsf{F}} \leq \frac{\varepsilon}{8k}\sigma_{k+1}^2.$$

Now if Z_k denote the first k singular vectors of the matrix $\widetilde{Apx}$ and Z_m denote the first m columns of Z_k, then by Lemma 25, we get that

$$\|M - (Q'Z_m)(Q'Z_m)^{\mathsf{T}}M\|_{\mathsf{F}}^2 \leq \|M - Q'(Q'^{\mathsf{T}}M)_m\|_{\mathsf{F}}^2 + 2m(\frac{\varepsilon^2}{8k}\sigma_{k+1}^2)$$

$$\leq \|M - Q'(Q'^{\mathsf{T}}M)_m\|_{\mathsf{F}}^2 + \frac{\varepsilon^2}{4}\sigma_{k+1}^2.$$

Finally, by Lemma 23, we obtain that

$$\|M - (Q'Z_m)(Q'Z_m)^{\mathsf{T}}M\|_2 \leq (1 + 3\varepsilon/2)\sigma_{k+1} + \sqrt{(\varepsilon^2/4)\sigma_{k+1}^2} \leq (1 + 2\varepsilon)\sigma_{k+1}.$$

Also $\|M - (Q'Z_k)(Q'Z_k)^\mathsf{T} M\|_2 \le \|M - (Q'Z_m)(Q'Z_m)^\mathsf{T} M\|_2 \le (1+2\varepsilon)\sigma_{k+1}(M)$ since $Q'Z_k$ has orthonormal columns and $\mathrm{colspan}(Q'Z_k) \supseteq \mathrm{colspan}(Q'Z_m)$. Thus in time

$$T\left(\frac{\varepsilon}{\kappa^{5q}k^7 C^q}\right)qk + T\left(\frac{\varepsilon^2}{48\kappa^2(\sqrt{qk})k}\right)qk$$

Algorithm 1 computes a $1 + 2\varepsilon$ approximation. Scaling the value of ε gives us the result. If the approximations $M \circ v$ are spanned by column space of M for all vectors v, then the columns of K' are spanned by the matrix M. Thus the columns of Q' are also spanned by M which implies that columns of the matrix $Q'Z_m$ is spanned by M. ∎

Appendix D. Omitted Proofs in Section 6

D.1. Proof of Lemma 9

Proof Define $Z := U^T \widetilde{Z}$. We have

$$
\begin{aligned}
1 + \varepsilon &\ge \|AA^+B(\beta^2 I - \Delta)^{-1/2} - \widetilde{Z}\widetilde{Z}^\mathsf{T} AA^+B(\beta^2 I - \Delta)^{-1/2}\|_2 \\
&\ge \|UU^\mathsf{T} B(\beta^2 I - \Delta)^{-1/2} - \widetilde{Z}\widetilde{Z}^\mathsf{T} UU^\mathsf{T} B(\beta^2 I - \Delta)^{-1/2}\|_2 \\
&\ge \|UU^\mathsf{T} B(\beta^2 I - \Delta)^{-1/2} - UU^\mathsf{T}\widetilde{Z}\widetilde{Z}^\mathsf{T} UU^\mathsf{T} B(\beta^2 I - \Delta)^{-1/2}\|_2 \\
&= \|UU^\mathsf{T} B(\beta^2 I - \Delta)^{-1/2} - UZZ^\mathsf{T} U^\mathsf{T} B(\beta^2 I - \Delta)^{-1/2}\|_2 \\
&= \|U^\mathsf{T} B(\beta^2 I - \Delta)^{-1/2} - ZZ^\mathsf{T} U^\mathsf{T} B(\beta^2 I - \Delta)^{-1/2}\|_2
\end{aligned}
$$

which implies using Lemma 7 that $UZ = UU^\mathsf{T}\widetilde{Z} = AA^+\widetilde{Z}$ is a good space to project the columns of B onto i.e.,

$$\|(AA^+\widetilde{Z})(AA^+\widetilde{Z})^+ B - B\|_2 \le (1+\varepsilon)\beta.$$

∎

D.2. Proof of Lemma 10

Polynomial Approximation of $(1-x)^{-1/2}$**.** We want to obtain a polynomial $p(x)$ such that $|p(x) - (1-x)^{-1/2}| \le \delta$ in the interval $x \in [0, 1/(1+\varepsilon)]$. Consider the Taylor expansion of $(1-x)^{-1/2}$:

$$(1-x)^{-1/2} = \sum_{j=0}^{\infty} \frac{(2j)!}{2^{2j}j!^2} x^j.$$

The above series converges for all $|x| < 1$. Consider $q(x)$ to be the Taylor series up to T terms. Then for $1 > x \ge 0$, we have $0 \le q(x) \le (1-x)^{-1/2}$ and for $0 \le x \le 1/(1+\varepsilon)$

$$(1-x)^{-1/2} - q(x) = \sum_{j=T}^{\infty} \frac{(2j)!}{2^{2j}j!^2} x^j \le \sum_{j=T}^{\infty} x^j = \frac{x^T}{1-x} \le \frac{(1+\varepsilon)}{\varepsilon(1+\varepsilon)^T} = \frac{1}{\varepsilon(1+\varepsilon)^{T-1}}.$$

Thus, if $T - 1 \ge 4\log(1/(\varepsilon\delta))/\varepsilon \ge \log(1/\varepsilon\delta)/\log(1+\varepsilon)$, we have $(1+\varepsilon)^{T-1} \ge 1/\varepsilon\delta$ which implies that

$$0 \le (1-x)^{-1/2} - q(x) \le \delta$$

for all $0 \leq x \leq 1/(1 + \varepsilon)$. So, there is a degree $t = O(\log(1/\varepsilon\delta)/\varepsilon)$ polynomial that uniformly approximates $(1 - x)^{-1/2}$ upto an error δ in the interval $[0, 1/(1 + \varepsilon)]$. Now, we further approximate the degree t polynomial $q(x)$ with a degree $\widetilde{O}(\sqrt{t})$ polynomial.

First we have the following Theorem.

Theorem 27 (Theorem 3.3 of [24]) *For any positive integers s and d, there is a degree d polynomial $p_{s,d}(x)$ that satisfies*

$$\sup_{x \in [-1,1]} |p_{s,d}(x) - x^s| \leq 2e^{-d^2/2s}.$$

Further, this polynomial $p_{s,d}$ is defined as follows

$$p_{s,d}(x) = \mathbb{E}_{Y_1,\ldots,Y_s}[T_{|D|}(x)\mathbb{I}[|D| \leq d]]$$

where $Y_1, \ldots, Y_s$ are independent Rademacher random variables, $D = \sum_{i=1}^{s} Y_i$ and $\mathbb{I}$ denotes the indicator function.

Clearly the polynomial $p_{s,d}$ is defined as a weighted linear combination of Chebyshev Polynomials of various degrees at most d. With $d = \sqrt{2s\log(1/\delta)}$, we have that

$$\sup_{x \in [-1,1]} |p_{s,d}(x) - x^s| \leq 2e^{-\log(1/\delta)} \leq 2\delta.$$

Thus, given an arbitrary degree t polynomial $q(x) = \sum_{i=0}^{t} q_i x^i$ where $q_0, \ldots, q_t$ are the coefficients of the polynomial, then the degree d polynomial $r(x) = \sum_{i=0}^{t} q_i p_{i,d}(x)$ with $d = \sqrt{2t\log(1/\delta)}$ satisfies

$$
\begin{aligned}
\sup_{x \in [-1,1]} |q(x) - r(x)| &= \sup_{x \in [-1,1]} |\sum_{i=0}^{t} q_i x^i - \sum_{i=0}^{t} q_i p_{i,d}(x)| \\
&\leq \sup_{x \in [-1,1]} \sum_{i=0}^{t} |q_i||x^i - p_{i,d}(x)| \\
&\leq \sup_{x \in [-1,1]} \sum_{i=0}^{t} |q_i| 2\delta \\
&= 2\|q\|_1 \delta.
\end{aligned}
$$

We now bound $\|r\|_1$. We have

$$
\begin{aligned}
\|r\|_1 = \|\sum_i q_i p_{i,d}(x)\| &\leq \sum_i |q_i| \|p_{i,d}(x)\|_1 \\
&= \sum_i |q_i| \|\mathbb{E}_{Y_1,\ldots,Y_s}[T_{|D|}(x)\mathbb{I}[|D| \leq d]]\|_1 \\
&\leq \sum_i |q_i| \mathbb{E}_{Y_1,\ldots,Y_s}[\|T_{|D|}(x)\mathbb{I}[|D| \leq d]]\|_1] \\
&\leq \sum_i |q_i| \frac{1}{2}(1 + \sqrt{2})^d = \frac{1}{2}(1 + \sqrt{2})^d \|q\|_1.
\end{aligned}
$$

Here we use the fact that $\| \cdot \|_1$ is convex over polynomials and that the sum of absolute values of coefficients of a Chebyshev polynomial of degree d is bounded by $(1 + \sqrt{2})^d$. Thus we have the following lemma.

Lemma 28 *Given any polynomial $q(x)$ of degree t, there exists a polynomial $r(x)$ of degree $d = \sqrt{2t \log(2\|q\|_1/\delta)}$ such that*

$$\sup_{x \in [-1,1]} |q(x) - r(x)| \le \delta$$

and $\|r\|_1 \le (1 + \sqrt{2})^d \|q\|_1$.

We already saw that the polynomial $q(x) = \sum_{j=0}^{t} \frac{(2j)!}{2^{2j} j!^2} x^j$ satisfies $|q(x) - (1-x)^{-1/2}| \le \delta$ for $x \in [0, 1/(1+\varepsilon)]$ if $t = O(\log(1/\varepsilon\delta)/\varepsilon)$. We also have $\|q\|_1 = \sum_{j=0}^{t} |(2j)!/(2^{2j}(j!)^2)| \le t + 1$. Thus by the above lemma, we can compute a polynomial $r(x)$ of degree $d = O(\sqrt{t \log(t/\delta)}) = O(\frac{1}{\sqrt{\varepsilon}} \log(1/\varepsilon\delta))$ such that

$$\sup_{x \in [0, 1/(1+\varepsilon)]} |r(x) - (1-x)^{-1/2}| \le \sup_{x \in [0, 1/(1+\varepsilon)]} |(1-x)^{-1/2} - q(x)| + \sup_{x \in [-1,1]} |q(x) - r(x)| \le 2\delta.$$

and we also have $\|r\|_1 = O((1 + \sqrt{2})^d t) = O((1 + \sqrt{2})^{O(\sqrt{1/\varepsilon} \log(1/\varepsilon\delta))} \log(1/\varepsilon\delta)/\varepsilon)$. We summarize this in the following lemma.

Lemma 29 *Given $\varepsilon, \delta > 0$, there exists a polynomial $r(x)$ of degree $O(\frac{1}{\sqrt{\varepsilon}} \log(1/\varepsilon\delta))$ and $\|r\|_1 = O((1 + \sqrt{2})^{O(\sqrt{1/\varepsilon} \log(1/\varepsilon\delta))} \log(1/\varepsilon\delta)/\varepsilon)$ such that*

$$\sup_{x \in [0, 1/(1+\varepsilon)]} |r(x) - (1-x)^{-1/2}| \le \delta.$$

Lemma 30 (Matrix Approximation Lemma) *If $A \in \mathbb{R}^{n \times n}$ is a positive semidefinite matrix with $\lambda_{\max}(A) < 1$ and if $r(x)$ is a polynomial such that*

$$\sup_{x \in [0, \lambda_{\max}(A)]} |r(x) - (1-x)^{-1/2}| \le \delta$$

then $\|r(A) - (I - A)^{-1/2}\|_2 \le \delta$.

Proof Let $A = VDV^{\mathsf{T}}$ be the eigenvalue decomposition of D with $D = \text{diag}(\lambda_1, \ldots, \lambda_n)$ where $\lambda_{\max} = \lambda_1 \ge \ldots \ge \lambda_n \ge 0$. Then $(I - A)^{-1/2} = V(I - D)^{-1/2} V^{\mathsf{T}}$ and $r(A) = Vr(D)V^{\mathsf{T}}$. Therefore

$$\begin{aligned}
\|r(A) - (I - A)^{-1/2}\|_2 &= \|V((I - D)^{-1/2} - r(D))V^{\mathsf{T}}\|_2 \\
&= \|(I - D)^{-1/2} - r(D)\|_2 \\
&= \max_i |(1 - \lambda_i)^{-1/2} - r(\lambda_i)| \\
&\le \sup_{x \in [0, \lambda_{\max}(A)]} |(1 - x)^{-1/2} - r(x)| \le \delta.
\end{aligned}$$

Here we use the fact that $0 \le \lambda_1, \ldots, \lambda_n \le \lambda_{\max}(A)$. ∎

As Δ is a positive semidefinite matrix such that $\beta^2 \ge (1+\varepsilon)\|\Delta\|_2$, then $\|\Delta/\beta^2\|_2 \le 1/(1+\varepsilon)$ and hence we can compute a polynomial $r(x)$ of degree $O(\frac{1}{\sqrt{\varepsilon}} \log(1/\varepsilon\delta))$ such that

$$\|r(\Delta/\beta^2) - (I - \Delta/\beta^2)^{-1/2}\|_2 \le \delta.$$

Modified Problem. Instead of considering the matrix $\mathcal{M} = AA^+B(\beta^2 I - \Delta)^{-1/2}$ for low rank approximation, we consider the matrix $\mathcal{M}' = AA^+BM/\beta$ for $M = r(\Delta/\beta^2)$, where $r(x)$ is a low degree polynomial, and argue that a $1 + \varepsilon$ LRA solution for the matrix $\mathcal{M}'$ is a $1 + 2\varepsilon$ approximation for the LRA problem on matrix $\mathcal{M}$.

Proof [Proof of Lemma 10] Recall $\Delta = B^\mathsf{T}(I - AA^+)B$ and therefore $\|\Delta\|_2 = \|(I - AA^+)B\|_2^2$. Given that $\beta \geq (1 + \varepsilon) \max(\|(I - AA^+)B\|_2, \sigma_{k+1}(B))$, we have $\beta^2 \geq (1 + \varepsilon)^2 \|\Delta\|_2 \geq (1 + \varepsilon)\|\Delta\|_2$. Thus, $\|\Delta/\beta^2\|_2 \leq 1/(1 + \varepsilon)$.

As $\|\Delta/\beta^2\|_2 \leq 1/(1 + \varepsilon)$, we approximate $(I - \Delta/\beta^2)^{-1/2}$ with the matrix $M = r(\Delta/\beta^2)$ where $r(x) = \sum_{i=0}^t r_i x^i$ is a polynomial of degree $t = O(\frac{1}{\sqrt{\varepsilon}}\log(\frac{1}{\varepsilon\delta}))$ given by Lemma 29. By Lemma 30, the matrix $r(\Delta/\beta^2) = \sum_{i=0}^t r_i(\Delta/\beta^2)^i$ satisfies

$$\|(I - \Delta/\beta^2)^{-1/2} - M\|_2 = \|(I - \Delta/\beta^2)^{-1/2} - r(\Delta/\beta^2)\|_2$$
$$= \|(I - \Delta/\beta^2)^{-1/2} - \sum_{i=0}^t r_i\left(\frac{\Delta}{\beta^2}\right)^i\|_2 \leq \delta.$$

As $\|\Delta/\beta^2\|_2 \leq 1/(1 + \varepsilon)$ and Δ/β^2 is a positive semidefinite matrix, we have $\sigma_{\max}(I - \Delta/\beta^2) \leq 1$ and $\sigma_{\min}(I - \Delta/\beta^2) \geq \varepsilon/(1 + \varepsilon)$. Therefore $\sigma_{\max}((I - \Delta/\beta^2)^{-1/2}) \leq \sqrt{(1 + \varepsilon)/\varepsilon}$ and $\sigma_{\min}((I - \Delta/\beta^2)^{-1/2}) \geq 1$. By Weyl's inequality, we obtain that

$$\sigma_{\max}(M) \leq \sqrt{(1 + \varepsilon)/\varepsilon} + \delta \quad \text{and} \quad \sigma_{\min}(M) \geq 1 - \delta.$$

By sub-multiplicativity of the spectral norm

$$\|AA^+B(\beta^2 I - \Delta)^{-1/2} - \frac{AA^+BM}{\beta}\|_2 \leq \frac{\|AA^+B\|_2}{\beta}\|(I - (\Delta/\beta^2))^{-1/2} - M\|_2$$
$$\leq \frac{\|AA^+B\|_2}{\beta}\delta.$$

Using Weyl's inequality, we obtain

$$\sigma_{k+1}\left(\frac{AA^+BM}{\beta}\right) \leq \sigma_{k+1}(AA^+B(\beta^2 I - \Delta)^{-1/2}) + \frac{\|AA^+B\|_2}{\beta}\delta \leq 1 + \frac{\|AA^+B\|_2}{\beta}\delta. \quad (9)$$

The last inequality follows as there exists a rank k matrix with $\|AX - B\|_2 \leq \beta$. If we can now find a rank k matrix Z with orthonormal columns such that

$$\|ZZ^\mathsf{T}\frac{AA^+BM}{\beta} - \frac{AA^+BM}{\beta}\|_2 \leq (1 + \varepsilon)\sigma_{k+1}\left(\frac{AA^+BM}{\beta}\right), \quad (10)$$

then

$$\|ZZ^\mathsf{T}AA^+B(\beta^2 I - \Delta)^{-1/2} - AA^+B(\beta^2 I - \Delta)^{-1/2}\|_2$$
$$\leq \|ZZ^\mathsf{T}\frac{AA^+BM}{\beta} - \frac{AA^+BM}{\beta}\|_2 + \|(I - ZZ^\mathsf{T})\left(\frac{AA^+BM}{\beta} - AA^+B(\beta^2 I - \Delta)^{-1/2}\right)\|_2$$
$$\leq (1 + \varepsilon)\sigma_{k+1}\left(\frac{AA^+BM}{\beta}\right) + \frac{\|AA^+B\|_2}{\beta}\delta$$

Algorithm 3: $\text{Oracle}_{\mathcal{M}'}$

Input: $v \in \mathbb{R}^d$, $\varepsilon_{\mathrm{r}} > 0$
Output: $y \in \mathbb{R}^n$
```
/* Let r(x) be the polynomial as in Lemma 10                        */
```
$t \leftarrow \text{degree}(r)$
$\varepsilon_{\mathrm{reg}} \leftarrow O\left(\varepsilon_{\mathrm{r}}/\kappa \|r\|_1\right)$
$y \leftarrow 0$
$Apx_0 \leftarrow v$
for $i = 0, \ldots, t$ **do**
$\quad$ $y \leftarrow y + r_i Apx_i$
$\quad$ $Apx_{i+1} \leftarrow B^{\mathsf{T}} B \cdot Apx_i - B^{\mathsf{T}} \cdot (\text{HIGHPRECISIONREGRESSION}(A, B \cdot Apx_i, \varepsilon_{\mathrm{reg}}))$
end
$y \leftarrow (\text{HighPrecisionRegression}(A, B \cdot y, \varepsilon_{\mathrm{reg}}))/\beta$

$$\leq (1+\varepsilon)(1 + 2\|AA^+B\|_2(\delta/\beta)).$$

The last inequality follows from (9). If δ is chosen to be less than $\varepsilon/4\kappa$ where $\kappa = \sigma_1(B)/\sigma_{k+1}(B)$, then

$$
\begin{aligned}
&\|ZZ^{\mathsf{T}} AA^+B(\beta^2 I - \Delta)^{-1/2} - AA^+B(\beta^2 I - \Delta)^{-1/2}\|_2 \\
&\leq (1+\varepsilon)\left(1 + 2\|AA^+B\|_2(\delta/\beta)\right) \\
&\leq (1+\varepsilon)\left(1 + 2\frac{\|AA^+B\|_2}{\beta}\frac{\varepsilon\sigma_{k+1}(B)}{4\sigma_1(B)}\right) \\
&\leq 1 + 2\varepsilon
\end{aligned}
$$

as $\|AA^+B\|_2 \leq \sigma_1(B)$ and $\beta \geq (1+\varepsilon)\sigma_{k+1}(B)$. This implies that if $(1-x)^{-1/2}$ is approximated by a polynomial $r(x)$ uniformly in the interval $[0, 1/(1+\varepsilon)]$ with an error at most $\varepsilon/4\kappa$ and if matrix Z is an orthonormal basis for a space that spans a $1 + \varepsilon$ rank k approximation in spectral norm for the matrix $AA^+B\frac{r(\Delta/\beta^2)}{\beta}$, then

$$\|AA^+Z(AA^+Z)^+B - B\|_2 \leq (1+6\varepsilon)\beta = (1+O(\varepsilon))\text{Opt}.$$

We obtain the proof by appropriately scaling ε. $\blacksquare$

D.3. Proof of Lemma 11

Throughout the analysis, we assume $\|AA^+B\|_2 \geq \varepsilon\|B\|_2$. Suppose if $\|AA^+B\|_2 \leq \varepsilon\|B\|_2$. Let z be the top singular vector of matrix B. Then

$$
\begin{aligned}
\|B\|_2^2 &= \|Bz\|_2^2 \\
&= \|AA^+Bz\|_2^2 + \|(I - AA^+)Bz\|_2^2 \\
&\leq \varepsilon^2\|B\|_2^2 + \|(I - AA^+)Bz\|_2^2.
\end{aligned}
$$

Algorithm 4: Oracle$_{\mathcal{M}'^\top}$

Input: $v \in \mathbb{R}^d$, $\varepsilon_\mathrm{r} > 0$
Output: $y \in \mathbb{R}^n$
```
/* Let r(x) be the polynomial as in Lemma 10                              */
```
$t \leftarrow \mathrm{degree}(r)$
$\varepsilon_\mathrm{reg} \leftarrow O\left(\varepsilon_\mathrm{r}/\kappa\|r\|_1\right)$
$y \leftarrow 0$
$Apx_0 \leftarrow B^\top \cdot (\textsc{HighPrecisionRegression}(A, v, \varepsilon_\mathrm{reg}))$
for $i = 0, \ldots, t$ **do**
$\quad \mid \quad y \leftarrow y + r_i Apx_i$
$\quad \mid \quad Apx_{i+1} \leftarrow B^\top B \cdot Apx_i - B^\top \cdot (\textsc{HighPrecisionRegression}(A, B \cdot Apx_i, \varepsilon_\mathrm{reg}))$
end
$y \leftarrow y/\beta$

Thus, $\|(I - AA^+)B\|_2^2 \geq \|(I - AA^+)Bz\|_2^2 \geq (1 - \varepsilon^2)\|B\|_2^2$. Therefore Opt $\geq \sqrt{1 - \varepsilon^2}\|B\|_2$ which implies $\|B\|_2 \leq (1/\sqrt{1 - \varepsilon^2})$Opt $\leq (1 + \varepsilon)$Opt for $\varepsilon \leq 1/2$. Thus $\|A(0) - B\|_2 \leq (1 + \varepsilon)$Opt and hence we have a trivial $(1 + \varepsilon)$ approximate solution. Thus we can assume $\|AA^+B\|_2 \geq \varepsilon\|B\|_2$.

Based on Theorem 1, we compute approximate projections onto the column span of A. The following lemma states that a matrix-vector product with the matrix (Δ/β^2) can be approximated well.

Lemma 31 *Given an arbitrary vector $v \in \mathbb{R}^d$, we can compute a vector $y \in \mathbb{R}^d$ such that*

$$\|y - (1/\beta^2)\Delta v\|_2 \leq \varepsilon_\mathrm{reg}\kappa\|v\|_2$$

in time $O(\mathsf{nnz}(B) + (\mathsf{nnz}(A) + c^2)\log(1/\varepsilon_\mathrm{reg}))$.

Proof Recall that $\Delta = B^\top(I - AA^+)B$. Therefore, for a vector v, $\Delta v = B^\top Bv - B^\top AA^+ Bv$. After computing Bv exactly, we can compute $\widetilde{y}$ by Theorem 1 in $O((\mathsf{nnz}(A) + c^2)\log(1/\varepsilon_\mathrm{reg}))$ time such that

$$\|AA^+Bv - \widetilde{y}\|_2 \leq \varepsilon_\mathrm{reg}\|(I - AA^+)Bv\|_2.$$

Let $y = B^\top Bv - B^\top\widetilde{y}$ which can be computed in $O(\mathsf{nnz}(B))$ time. Then $\Delta v - y = B^\top(\widetilde{y} - AA^+Bv)$ which implies $\|\Delta v - y\|_2 \leq \|B\|_2\|\widetilde{y} - AA^+Bv\|_2 \leq \varepsilon_\mathrm{reg}\|B\|_2\|(I - AA^+)B\|_2\|v\|_2$.

Thus given a vector v, we can compute $(\Delta/\beta^2)v$ up to an error of

$$\varepsilon_\mathrm{reg}\|B\|_2\|(I - AA^+)B\|_2\|v\|_2/\beta^2 \leq \varepsilon_\mathrm{reg}\kappa\|v\|_2$$

since $\beta \geq \max(\|(I - AA^+)B\|_2, \sigma_{k+1}(B))$. ∎

Lemma 32 *Given an arbitrary vector $v \in \mathbb{R}^d$, for matrix $M = r\left(\Delta/\beta^2\right) = \sum_{j=0}^t r_j\left(\Delta/\beta^2\right)^j$ where the degree $t = O((1/\sqrt{\varepsilon})\log(\kappa/\varepsilon))$ and $\|r\|_1 = O((1 + \sqrt{2})^{O(\sqrt{1/\varepsilon}\log(\kappa/\varepsilon))}\log(\kappa/\varepsilon))$, we can compute a vector y such that $\|Mv - y\|_2 \leq \varepsilon_\mathrm{r}\|v\|_2$ in time*

$$O\left(t \cdot \left(\mathsf{nnz}(B) + (\mathsf{nnz}(A) + c^2)\log\left(\kappa\|r\|_1/\varepsilon_\mathrm{r}\right)\right)\right).$$

Proof Let $Apx_0 := v$ and for $i \geq 1$, define Apx_i to be the approximation computed for the product $(\Delta/\beta^2)Apx_{i-1}$ by Lemma 31. Define

$$E_i := \|(\Delta/\beta^2)^i v - Apx_i\|_2.$$

We have the following recurrence

$$E_i = \|\left(\frac{\Delta}{\beta^2}\right)^i v - Apx_i\|_2 \leq \|(\Delta/\beta^2)^i v - (\Delta/\beta^2)Apx_{i-1}\|_2 + \|(\Delta/\beta^2)Apx_{i-1} - Apx_i\|_2$$

$$\leq \|(\Delta/\beta^2)\|_2 E_{i-1} + \varepsilon_{\mathrm{reg}}\kappa\|Apx_{i-1}\|_2$$

$$\leq \|(\Delta/\beta^2)\|_2 E_{i-1} + \varepsilon_{\mathrm{reg}}\kappa \cdot (\|\Delta/\beta^2\|_2^{i-1}\|v\|_2 + E_{i-1})$$

$$\leq (\|\Delta\|_2/\beta^2 + \varepsilon_{\mathrm{reg}}\kappa)E_{i-1} + \varepsilon_{\mathrm{reg}}\kappa\|\Delta/\beta^2\|_2^{i-1}\|v\|_2.$$

As $\beta \geq (1+\varepsilon)\|(I - AA^+)B\|_2$, we have that $\|\Delta/\beta^2\|_2 \leq 1/(1+\varepsilon)^2$. If $\varepsilon_{\mathrm{reg}}\kappa \leq \varepsilon/4$, then $\|\Delta/\beta^2\|_2 + \varepsilon_{\mathrm{reg}}\kappa \leq 1/(1+\varepsilon)^2 + \varepsilon/4 \leq 1/(1+\varepsilon)$. Therefore

$$E_i \leq \frac{E_{i-1}}{1+\varepsilon} + \frac{\varepsilon_{\mathrm{reg}}\kappa}{(1+\varepsilon)^{2(i-1)}}\|v\|_2.$$

which implies upon solving the recurrence that

$$E_i \leq \varepsilon_{\mathrm{reg}}\kappa\|v\|_2$$

for all i. Then

$$\|Mv - \sum_{j=0}^t r_j Apx_j\|_2 \leq \sum_{j=0}^t |r_j|\|(\Delta/\beta^2)^j v - Apx_j\|_2$$

$$\leq \sum_{j=0}^t |r_j| E_j \leq \varepsilon_{\mathrm{reg}}\kappa\|v\|_2 \sum_{j=0}^t |r_j| = \varepsilon_{\mathrm{reg}}\kappa\|v\|_2\|r\|_1.$$

So for any arbitrary vector v, we can compute a vector y such that

$$\|Mv - y\|_2 \leq \varepsilon_{\mathrm{r}}\|v\|_2.$$

by setting $\varepsilon_{\mathrm{reg}} = O(\frac{\varepsilon_{\mathrm{r}}}{\kappa\|r\|_1}) \leq \varepsilon/4\kappa$ for all t approximate products and and thus y can be computed by Lemma 31 in time

$$O(t \cdot (\mathsf{nnz}(B) + (\mathsf{nnz}(A) + r^2)\log(\kappa\|r\|_1/\varepsilon_{\mathrm{r}}))).$$

This concludes the proof of the lemma. ∎

Thus for an arbitrary vector v, we can compute a vector y such that $\|Mv - y\|_2 \leq \varepsilon_{\mathrm{r}}\|v\|_2$.
Proof [Proof of Lemma 11] Recall that $\mathcal{M}' = (AA^+BM)/\beta$, $\|M\|_2 \leq 2/\sqrt{\varepsilon}$ and $\sigma_{\min}(M) \geq 1/2$ from Lemma 10. We have $\|AA^+BM\|_2 \geq \|AA^+B\|_2\sigma_{\min}(M) \geq \|AA^+B\|_2/2 \geq \varepsilon\|B\|_2/2$ where the last inequality follows from our assumption that $\|AA^+B\|_2 \geq \varepsilon\|B\|_2$. Thus $\|\mathcal{M}'\|_2 \geq \varepsilon\|B\|_2/2\beta \geq \varepsilon/4$ as $\beta \leq (1+\varepsilon)\|B\|_2$.

Now we show how to compute approximations to $\mathcal{M}'v$ and $\mathcal{M}'^\mathsf{T}v'$ for arbitrary vectors v, v'.

To compute an approximation to $\mathcal{M}'v$, we first obtain a vector y_1 using the above lemma such that $\|Mv - y_1\|_2 \le \varepsilon_{\mathrm{r}}\|v\|_2$. Then we compute the product By_1 exactly in time $O(\mathsf{nnz}(B))$. Thereafter we compute a vector y_2 by Theorem 1 such that

$$\|y_2 - AA^+ By_1\|_2 \le \varepsilon_{\mathrm{reg}}\|(I - AA^+)By_1\|_2 \le \varepsilon_{\mathrm{reg}}\|(I - AA^+)B\|_2\|y_1\|_2.$$

We also have $\|AA^+ By_1 - AA^+ BMv\|_2 \le \varepsilon_{\mathrm{r}}\|AA^+ B\|_2\|v\|_2$. Therefore by triangle inequality, $\|AA^+ BMv - y_2\|_2 \le \varepsilon_{\mathrm{r}}\|AA^+ B\|_2\|v\|_2 + \varepsilon_{\mathrm{reg}}\|(I - AA^+)B\|_2\|y_1\|_2$. Hence

$$\|\mathcal{M}'v - (y_2/\beta)\|_2 \le \varepsilon_{\mathrm{r}}\frac{\|AA^+ B\|_2}{\beta}\|v\|_2 + \varepsilon_{\mathrm{reg}}\|y_1\|_2 \le \varepsilon_{\mathrm{r}}\kappa\|v\|_2 + \varepsilon_{\mathrm{reg}}(\varepsilon_{\mathrm{r}}\|v\|_2 + \|M\|_2 v)$$

$$\le \varepsilon_{\mathrm{r}}(\kappa + 1)\|v\|_2 + \frac{2\varepsilon_{\mathrm{reg}}}{\sqrt{\varepsilon}}\|v\|_2.$$

Thus if $\varepsilon_{\mathrm{r}} = O(\varepsilon_{\mathrm{f}}\varepsilon/\kappa)$ and $\varepsilon_{\mathrm{reg}} = O(\varepsilon_{\mathrm{f}}\varepsilon^{3/2})$, we have that $\|\mathcal{M}'v - (y_2/\beta)\|_2 \le \varepsilon_{\mathrm{f}}\varepsilon\|v\|_2 \le \varepsilon_{\mathrm{f}}\|\mathcal{M}'\|_2\|v\|_2$. Therefore a vector y_2/β can be computed in time $O(t \cdot (\mathsf{nnz}(B) + (\mathsf{nnz}(A) + c^2)\log\left(\frac{\kappa^2\|r\|_1}{\varepsilon_{\mathrm{f}}\varepsilon}\right))) + O((\mathsf{nnz}(A) + c^2)\log(\frac{1}{\varepsilon_{\mathrm{f}}\varepsilon}))$.

Now we compute an approximation for $\mathcal{M}'^{\mathsf{T}}v = (M^{\mathsf{T}}B^{\mathsf{T}}AA^+/\beta)v$ for an arbitrary vector v. We first compute a vector y_1 such that

$$\|AA^+ v - y_1\|_2 \le \varepsilon_{\mathrm{reg}}\|(I - AA^+)v\|_2 \le \varepsilon_{\mathrm{reg}}\|v\|_2.$$

Then we compute $B^{\mathsf{T}}y_1$ exactly. Then we compute a vector y_2 such that $\|MB^{\mathsf{T}}y_1 - y_2\|_2 \le \varepsilon_{\mathrm{r}}\|B^{\mathsf{T}}y_1\|_2 \le \varepsilon_{\mathrm{r}}\|B\|_2(1 + \varepsilon_{\mathrm{reg}})\|v\|_2$. We further have

$$\|MB^{\mathsf{T}}AA^+ v - MB^{\mathsf{T}}y_1\|_2 \le \varepsilon_{\mathrm{reg}}\|MB^{\mathsf{T}}\|_2\|v\|_2 \le \varepsilon_{\mathrm{reg}}\frac{2\|B\|_2}{\sqrt{\varepsilon}}\|v\|_2.$$

Thus

$$\|y_2 - MB^{\mathsf{T}}AA^+ v\|_2 \le 2\varepsilon_{\mathrm{r}}\|B\|_2\|v\|_2 + \varepsilon_{\mathrm{reg}}\frac{2\|B\|_2}{\sqrt{\varepsilon}}\|v\|_2$$

and hence

$$\|y_2/\beta - \mathcal{M}'^{\mathsf{T}}v\|_2 \le 2\varepsilon_{\mathrm{r}}\kappa\|v\|_2 + \varepsilon_{\mathrm{reg}}\frac{2\kappa}{\sqrt{\varepsilon}}\|v\|_2$$

Now picking $\varepsilon_{\mathrm{r}} = O(\varepsilon_{\mathrm{f}}\varepsilon/\kappa)$ and $\varepsilon_{\mathrm{reg}} = O(\varepsilon_{\mathrm{f}}\varepsilon^{3/2}/\kappa)$, we obtain that

$$\|(y_2/\beta) - \mathcal{M}'^{\mathsf{T}}v\|_2 \le \varepsilon_{\mathrm{f}}\varepsilon\|v\|_2 \le \varepsilon_{\mathrm{f}}\|\mathcal{M}'\|_2\|v\|_2.$$

Thus this approximation can be computed in time $O(t \cdot (\mathsf{nnz}(B) + (\mathsf{nnz}(A) + c^2)\log\left(\frac{\kappa^2\|r\|_1}{\varepsilon_{\mathrm{f}}\varepsilon}\right))) + O((\mathsf{nnz}(A) + c^2)\log(\frac{\kappa}{\varepsilon_{\mathrm{f}}\varepsilon}))$. So, given an accuracy parameter ε_{f} we can compute approximate matrix-vector products with $\mathcal{M}'$ and $\mathcal{M}'^{\mathsf{T}}$ in time at most

$$T(\varepsilon_{\mathrm{f}}) = O(t \cdot (\mathsf{nnz}(B) + (\mathsf{nnz}(A) + c^2)\log\left(\kappa(B)^2\|r\|_1/(\varepsilon_{\mathrm{f}}\varepsilon)\right)))$$
$$+ O((\mathsf{nnz}(A) + c^2)\log(\kappa(B)/(\varepsilon_{\mathrm{f}}\varepsilon))).$$

■

D.4. Proof of Theorem 12

Proof From Lemma 10,

$$\sigma_1(\mathcal{M}') \leq \sigma_1\left(\frac{AA^+B}{\beta}\right)\|M\|_2 \leq \sigma_1\left(\frac{AA^+B}{\beta}\right)\frac{2}{\sqrt{\varepsilon}}$$

and

$$\sigma_{k+1}(\mathcal{M}') \geq \sigma_{k+1}(AA^+B/\beta) \cdot \sigma_{\min}(M) \geq \sigma_{k+1}(AA^+B/\beta)(1/2).$$

Therefore $\kappa(\mathcal{M}') \leq \sigma_1(AA^+B/\beta)(2/\sqrt{\varepsilon})/\sigma_{k+1}(AA^+B/\beta)/2 \leq (4/\sqrt{\varepsilon})\kappa(AA^+B)$. By Theorem 8, we can compute a matrix $Z \in \mathbb{R}^{n \times k}$ such that $\|(I - ZZ^\mathsf{T})\mathcal{M}'\|_2 \leq (1 + 2\varepsilon)\sigma_{k+1}(\mathcal{M}')$ in time

$$T\left(\frac{\varepsilon}{\kappa(\mathcal{M}')^{5q}k^{11}C^q}\right)qk + T\left(\frac{\varepsilon^2}{48\kappa(\mathcal{M}'^2(\sqrt{qk})k)}\right)qk$$

where $q = O((1/\sqrt{\varepsilon})\log(d/\varepsilon))$. Thus the total time required is

$$O\left(tqk \cdot \left(\mathsf{nnz}(B) + (\mathsf{nnz}(A) + c^2)\log\left(\frac{\kappa^2\|r\|_1\kappa(\mathcal{M}')^{5q}k^{11}C^q}{\varepsilon^2}\right)\right)\right).$$

As $\|r\|_1 = (1 + \sqrt{2})^{O(1/\sqrt{\varepsilon}\log(\kappa/\varepsilon))}\log(\kappa/\varepsilon)/\varepsilon$ and $\kappa(\mathcal{M}') = \kappa(AA^+B)/\sqrt{\varepsilon}$, we obtain that the total time required is $O(tqk \cdot \mathsf{nnz}(B) + tqk \cdot (\frac{1}{\sqrt{\varepsilon}}\log(\kappa/\varepsilon) + q)\log(\frac{\kappa \cdot \kappa(\mathcal{M}') \cdot k}{\varepsilon}) \cdot (\mathsf{nnz}(A) + c^2))$. Substituting $t = O(\sqrt{1/\varepsilon}\log(\kappa/\varepsilon))$, we obtain that the total running time is

$$O\left(\left(\frac{\mathsf{nnz}(B) \cdot k}{\varepsilon} + \frac{\mathsf{nnz}(A) \cdot k}{\varepsilon^{1.5}} + \frac{c^2k}{\varepsilon^{1.5}}\right) \cdot \mathrm{polylog}(\kappa, \kappa(AA^+B), d, k, 1/\varepsilon)\right) \qquad (11)$$

and an additional c^ω for computing a preconditioner. By Lemmas 9 and 10, we obtain that

$$\|(AA^+Z)(AA^+Z)^+B - B\|_2 = \|ZZ^\mathsf{T}B - B\|_2 \leq (1 + O(\varepsilon))\mathrm{Opt}.$$

The equality is from the fact that Z is spanned by columns of matrix A by Theorem 8 and therefore $AA^+Z = Z$. Thus there exists a matrix $X_1 \in \mathbb{R}^{c \times k}$ such that $AX_1 = Z$ and the matrix X_1 can be computed in time $O((\mathsf{nnz}(A) + c^2)k + c^\omega)$ using sketching-based preconditioning techniques. Let $Y_1 = Z^\mathsf{T}B$ which can be computed in time $O(\mathsf{nnz}(B) \cdot k)$. Therefore,

$$\|AX_1Y_1 - B\|_2 = \|ZZ^\mathsf{T}B - B\|_2 \leq (1 + O(\varepsilon))\mathrm{Opt}.$$

Thus $X_1 \cdot Y_1$ is a $1 + O(\varepsilon)$ approximation for the regression problem. By appropriately scaling ε, we obtain the proof. $\blacksquare$

D.5. Proof of Lemma 13

Proof Let $G \sim N(0,1)^{n \times (k+1)}$ and $F^\mathsf{T} \in \mathbb{R}^{(k+1) \times d}$ be a matrix with $k + 1$ orthonormal rows. Let α be a parameter to be chosen later and $\widetilde{B} := B + \alpha GF^\mathsf{T}$. For all matrices X, by the triangle inequality,

$$\|AX - \widetilde{B}\|_2 \in \|AX - B\|_2 \pm \alpha\|GF^\mathsf{T}\|_2.$$

With probability $\geq 9/10$, $\|G\|_2 \leq 2\sqrt{n}$. Thus $\|AX - \widetilde{B}\|_2 \in \|AX - B\|_2 \pm 2\alpha\sqrt{n}$. Therefore, if $\widetilde{X}$ is a $1 + \varepsilon$ approximation for $\min_{\text{rank-}k\ X} \|AX - \widehat{B}\|_2$, then $\|A\widetilde{X} - B\|_2 \leq (1+\varepsilon)\text{Opt} + 6\alpha\sqrt{n}$.

We now have $\sigma_1(AA^+\widetilde{B}) \leq \|\widetilde{B}\|_2 \leq \|B\|_2 + 2\alpha\sqrt{n}$ from above discussion. We now lower bound $\sigma_{k+1}(AA^+\widetilde{B})$. Let U be an orthonormal basis for columns of A. Therefore $AA^+ = UU^\mathsf{T}$.

$$
\begin{aligned}
\sigma_{k+1}(AA^+\widetilde{B}) &= \sigma_{k+1}(UU^\mathsf{T}\widetilde{B}) \\
&= \sigma_{k+1}(U^\mathsf{T}\widetilde{B}) \\
&= \sigma_{k+1}(U^\mathsf{T}B + \alpha U^\mathsf{T}GF^\mathsf{T}) \\
&\geq \sigma_{k+1}(U^\mathsf{T}BFF^\mathsf{T} + \alpha U^\mathsf{T}GF^\mathsf{T}) \\
&\geq \sigma_{k+1}(U^\mathsf{T}BF + \alpha U^\mathsf{T}G).
\end{aligned}
$$

As the rows of U^T are orthonormal, the matrix $G' = U^\mathsf{T}G$ is a matrix of i.i.d. normal random variables. Assuming A is of full rank, then G' is a $c \times (k+1)$ matrix. Assuming $c \geq k+1$, let E be the top $(k+1) \times (k+1)$ sub-matrix of $U^\mathsf{T}BF + \alpha G'$. Then E can be seen as a fixed $(k+1)\times(k+1)$ matrix where each entry is perturbed by a Gaussian random variable of variance α^2. From Theorem 2.2 of [30], we obtain that $\sigma_{\min}(E) \geq \alpha/(C\sqrt{k})$ for a constant C with probability $\geq 9/10$. Thus $\sigma_{k+1}(U^\mathsf{T}BF + \alpha U^\mathsf{T}G) \geq \sigma_{\min}(E) \geq \alpha/(C\sqrt{k})$.

Thus, $\sigma_1(AA^+\widetilde{B})/\sigma_{k+1}(AA^+\widetilde{B}) \leq (\|B\|_2 + 2\alpha\sqrt{n})/(\alpha/(C\sqrt{k}))$. For $\alpha = \frac{\varepsilon\sigma_{k+1}(B)}{(6\sqrt{n})}$, we obtain that

$$
\sigma_1(AA^+\widetilde{B})/\sigma_{k+1}(AA^+\widetilde{B}) \leq \frac{Cn}{\varepsilon}\kappa
$$

for a constant C with probability $\geq 4/5$. Also, if $\widetilde{X}$ is a $1 + \varepsilon$ approximation as mentioned above, $\|A\widetilde{X} - B\|_2 \leq (1+\varepsilon)\text{Opt} + \varepsilon\sigma_{k+1}(B) \leq (1+2\varepsilon)\text{Opt}$. We obtain the proof by scaling ε appropriately. $\blacksquare$