

Spectrum Approximation Beyond Fast Matrix Multiplication: Algorithms and Hardness

Cameron Musco
MIT
cnmusco@mit.edu

Praneeth Netrapalli
Microsoft Research
praneeth@microsoft.com

Aaron Sidford
Stanford University
sidford@stanford.edu

Shashanka Ubaru
University of Minnesota
ubaru001@umn.edu

David P. Woodruff
Carnegie Mellon University
dwoodruf@cs.cmu.edu

Abstract

Understanding the singular value spectrum of a matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is a fundamental task in countless numerical computation and data analysis applications. In matrix multiplication time, it is possible to perform a full SVD of $\mathbf{A}$ and directly compute the singular values $\sigma_1, \dots, \sigma_n$. However, little is known about algorithms that break this runtime barrier.

Using tools from stochastic trace estimation, polynomial approximation, and fast linear system solvers, we show how to efficiently isolate different ranges of $\mathbf{A}$'s spectrum and approximate the number of singular values in these ranges. We thus effectively compute an *approximate histogram* of the spectrum, which can stand in for the true singular values in many applications.

We use our histogram primitive to give the first algorithms for approximating a wide class of symmetric matrix norms and spectral sums *faster than the best known runtime for matrix multiplication*. For example, we show how to obtain a $(1 + \epsilon)$ approximation to the Schatten 1-norm (i.e. the nuclear or trace norm) in just $\tilde{O}((\text{nnz}(\mathbf{A})n^{1/3} + n^2)\epsilon^{-3})$ time for $\mathbf{A}$ with uniform row sparsity or $\tilde{O}(n^{2.18}\epsilon^{-3})$ time for dense matrices. The runtime scales smoothly for general Schatten p -norms, notably becoming $\tilde{O}(p \text{nnz}(\mathbf{A})\epsilon^{-3})$ for any real $p \geq 2$.

At the same time, we show that the complexity of spectrum approximation is inherently tied to fast matrix multiplication in the small ϵ regime. We use fine-grained complexity to give conditional lower bounds for spectrum approximation, showing that achieving milder ϵ dependencies in our algorithms would imply triangle detection algorithms for general graphs running in faster than state of the art matrix multiplication time. This further implies, through a reduction of [WW10], that highly accurate spectrum approximation algorithms running in subcubic time can be used to give subcubic time matrix multiplication. As an application of our bounds, we show that precisely computing all effective resistances in a graph in less than matrix multiplication time is likely difficult, barring a major algorithmic breakthrough.

1 Introduction

Given $\mathbf{A} \in \mathbb{R}^{n \times d}$, a central primitive in numerical computation and data analysis is to compute $\mathbf{A}$'s spectrum: the singular values $\sigma_1(\mathbf{A}) \geq \dots \geq \sigma_d(\mathbf{A}) \geq 0$. These values can reveal matrix structure and low effective dimensionality, which can be exploited in a wide range of spectral data analysis methods [Jol02, US16]. The singular values are also used as tuning parameters in many numerical algorithms performed on $\mathbf{A}$ [GVL12], and in general, to determine some of the most well-studied matrix functions [Hig08]. For example, for any $f : \mathbb{R}^+ \rightarrow \mathbb{R}^+$, we can define the *spectral sum*:

$$\mathcal{S}_f(\mathbf{A}) \stackrel{\text{def}}{=} \sum_{i=1}^d f(\sigma_i(\mathbf{A})).$$

Spectral sums often serve as snapshots of $\mathbf{A}$'s spectrum and are important in many applications. They encompass, for example, the log-determinant, the trace inverse, the Schatten p -norms, including the important nuclear norm, and general Orlicz norms (see Section 1.2 for details).

While the problem of computing a few of the largest or smallest singular values of $\mathbf{A}$ has been exhaustively studied [Par98, Saa11], much less is known about algorithms that approximate the full spectrum, and in particular, allow for the computation of summary statistics such as spectral sums. In n^ω time, it is possible to perform a full SVD and compute the singular values exactly.¹ Here, and throughout, $\omega \approx 2.3729$ denotes the *current* best exponent of fast matrix multiplication [Wil12]. However, even if one simply desires, for example, a constant factor approximation to the nuclear norm $\|\mathbf{A}\|_1$, no $o(n^\omega)$ time algorithm is known. We study the question of spectrum approximation, asking whether obtaining an accurate picture of $\mathbf{A}$'s spectrum is truly as hard as matrix multiplication, or if it is possible to break this barrier. We focus on spectral sums as a motivating application.

1.1 Our Contributions

Upper Bounds: On the upper bound side, we show that significant information about $\mathbf{A}$'s spectrum can be determined in $o(n^\omega)$ time, for the current value of ω . We show how to compute a histogram of the spectrum, which gives approximate counts of the number of squared singular values in the ranges $[(1-\alpha)^t \sigma_1^2(\mathbf{A}), (1-\alpha)^{t-1} \sigma_1^2(\mathbf{A})]$ for some width parameter α and for t ranging from 0 to some maximum T . Specifically our algorithm satisfies the following:

Theorem 1 (Histogram Approximation – Informal). *Given $\mathbf{A} \in \mathbb{R}^{n \times d}$, let b_t be the number of squared singular values of $\mathbf{A}$ on the range $[(1-\alpha)^t \sigma_1^2(\mathbf{A}), (1-\alpha)^{t-1} \sigma_1^2(\mathbf{A})]$. Then given error parameter $\epsilon > 0$, with probability 99/100, Algorithm 1 outputs for all $t \in \{0, \dots, T\}$, $\tilde{b}_t$ satisfying:*

$$(1 - \epsilon)b_t \leq \tilde{b}_t \leq (1 + \epsilon)b_t + \epsilon(b_{t-1} + b_{t+1}).$$

For input parameter $k \in \{1, \dots, d\}$, let $\bar{\kappa} \stackrel{\text{def}}{=} \frac{k\sigma_k^2(\mathbf{A}) + \sum_{i=k+1}^d \sigma_i^2(\mathbf{A})}{d \cdot (1-\alpha)^T}$ and $\hat{\kappa} \stackrel{\text{def}}{=} \frac{\sigma_{k+1}^2(\mathbf{A})}{(1-\alpha)^T}$. Let $d_s(\mathbf{A})$ be the maximum number of nonzeros in a row of $\mathbf{A}$. The algorithm's runtime can be bounded by:

$$\tilde{O}\left(\frac{\text{nnz}(\mathbf{A})k + dk^{\omega-1} + \sqrt{\text{nnz}(\mathbf{A})[d \cdot d_s(\mathbf{A}) + dk]\bar{\kappa}}}{\text{poly}(\epsilon, \alpha)}\right) \quad \text{or} \quad \tilde{O}\left(\frac{\text{nnz}(\mathbf{A})k + dk^{\omega-1} + (\text{nnz}(\mathbf{A}) + dk)[\sqrt{\hat{\kappa}}]}{\text{poly}(\epsilon, \alpha)}\right)$$

for sparse $\mathbf{A}$ or $\tilde{O}\left(\frac{nd^{\gamma-1} + n^{1/2}d^{3/2}\sqrt{\bar{\kappa}}}{\text{poly}(\epsilon, \alpha)}\right)$ for dense $\mathbf{A}$, where d^γ is the time it takes to multiply a $d \times d$ matrix by a $d \times k$ matrix using fast matrix multiplication.

¹ Note that an exact SVD is incomputable even with exact arithmetic [TB97]. Nevertheless, direct methods for the SVD obtain superlinear convergence rates and hence are often considered to be ‘exact’.

This primitive is useful on its own, giving an accurate summary of $\mathbf{A}$'s spectrum which can be used in many downstream applications. Setting the parameter k appropriately to balance costs (see overview in Section 1.3), we use it to give the first $o(n^\omega)$ algorithms for computing $(1 \pm \epsilon)$ relative error approximations to a broad class of spectral sums for functions f , which are a) smooth and b) quickly growing, so that very small singular values cannot make a significant contribution to $\mathcal{S}_f(\mathbf{A})$. This class includes for example the Schatten p -norms for all $p > 0$, the SVD entropy, the Ky Fan norms, and many general Orlicz norms.

For a summary of our p -norm results see Table 1. Focusing for simplicity on square matrices, with uniformly sparse rows, and assuming ϵ, p are constants, our algorithms approximate $\|\mathbf{A}\|_p^p$ in $\tilde{O}(\text{nnz}(\mathbf{A}))$ time for any real $p \geq 2$.² For $p \leq 2$, we achieve $\tilde{O}\left(\text{nnz}(\mathbf{A})n^{\frac{1/p-1/2}{1/p+1/2}} + n^{\frac{4/p-1}{2/p+1}}\sqrt{\text{nnz}(\mathbf{A})}\right)$ runtime. In the important case of $p = 1$, this becomes $\tilde{O}\left(\text{nnz}(\mathbf{A})n^{1/3} + n\sqrt{\text{nnz}(\mathbf{A})}\right)$. Note that $n\sqrt{\text{nnz}(\mathbf{A})} \leq n^2$, and for sparse enough $\mathbf{A}$, this bound is subquadratic. For dense $\mathbf{A}$, we use fast matrix multiplication, achieving time $\tilde{O}\left(n^{\frac{2.3729-.0994p}{1+.0435p}}\right)$ for all $p < 2$. For $p = 1$, this gives $\tilde{O}(n^{2.18})$. Even without fast matrix multiplication, the runtime is $\tilde{O}(n^{2.33})$, and so $o(n^\omega)$ for $\omega \approx 2.3729$.

Lower Bounds: On the lower bound side, we show that obtaining $o(n^\omega)$ time spectrum approximation algorithms with very high accuracy may be difficult. Our runtimes all depend polynomially on the error ϵ , and we show that improving this dependence, e.g., to $\log(1/\epsilon)$, or even to a better polynomial, would give faster algorithms for the well studied Triangle Detection problem.

Specifically, for a broad class of spectral sums, including all Schatten p -norms with $p \neq 2$, SVD entropy, $\log \det(\mathbf{A})$, $\text{tr}(\mathbf{A}^{-1})$, and $\text{tr}(\exp(\mathbf{A}))$, we show that any $(1 \pm \epsilon)$ approximation algorithm running in $O(n^\gamma \epsilon^{-c})$ time yields an algorithm for triangle detection running in $O(n^{\gamma+O(c)})$ time. For $\gamma < \omega$ and sufficiently small c , such an algorithm would improve the state of the art in triangle detection, which currently requires $\Theta(n^\omega)$ time on dense graphs. Furthermore, through a reduction of [WW10], any subcubic time triangle detection algorithm yields a subcubic time algorithm for Boolean Matrix Multiplication (BMM). Thus, any spectral sum algorithm achieving subcubic runtime and $\frac{1}{\epsilon^c}$ accuracy for small enough constant c , must (implicitly) implement fast matrix multiplication. This is in stark contrast to the fact that, for $c = 3$, for many spectral sums, including all Schatten p -norms with $p \geq 1/2$, we obtain subcubic, and in fact $o(n^\omega)$ for $\omega = 2.3729$, runtimes without using fast matrix multiplication (see Table 1 for precise ϵ dependencies).

Our lower bounds hold even for well-conditioned matrices and structured matrices like symmetric diagonally dominant (SDD) systems, both of which admit nearly linear time algorithms for system solving [ST04]. This illustrates a dichotomy between linear algebraic primitives like applying $\mathbf{A}^{-1}$ to a vector and spectral summarization tasks like precisely computing $\text{tr}(\mathbf{A}^{-1})$, which in some sense require more global information about the matrix. Our analysis has ramifications regarding natural open problems in graph theory and numerical computation. For example, for graph Laplacians, we show that accurately computing all *effective resistances* yields an accurate algorithm for computing $\text{tr}(\mathbf{A}^{-1})$ of certain matrices, which is enough to give triangle detection.

1.2 Related Work on Spectral Sums

The applications of approximate spectral sum computation are very broad. When $\mathbf{A}$ is positive semidefinite (PSD) and $f(x) = \log(x)$, $\mathcal{S}_f(\mathbf{A})$ is the log-determinant, which is important in machine learning and inference applications [Ras04, DKJ⁺07, FHT08]. For $f(x) = 1/x$, $\mathcal{S}_f(\mathbf{A})$ is the trace of the inverse, used in uncertainty quantification [BCF09] and quantum chromodynamics [SLO13].

²For any $\mathbf{A} \in \mathbb{R}^{n \times d}$, $\text{nnz}(\mathbf{A})$ denotes the number of nonzero entries in $\mathbf{A}$.

p range	Sparsity	Approx.	Runtime	Theorem
$p > 2$	uniform	$(1 + \epsilon)$	$\tilde{O}(\text{nnz}(\mathbf{A}) \cdot p/\epsilon^3)$	Thm 32
$p \leq 2$	uniform	$(1 + \epsilon)$	$\tilde{O}\left(\frac{1}{f(p,\epsilon)} \left[\text{nnz}(\mathbf{A}) n^{\frac{1/p-1/2}{1/p+1/2}} + n^{\frac{4/p-1}{2/p+1}} \sqrt{\text{nnz}(\mathbf{A})} \right]\right)$	Thm 33
$p \leq 2$	dense	$(1 + \epsilon)$	$\tilde{O}\left(\frac{1}{f(p,\epsilon)} n^{\frac{2.3729-.0994p}{1+.0435p}}\right)$ or $\tilde{O}\left(\frac{1}{f(p,\epsilon)} \cdot n^{\frac{3+p/2}{1+p/2}}\right)$ w/o FMM	Thm 31
$p > 0$	general	$(1 + \epsilon)$	$\tilde{O}\left(\frac{1}{f(p,\epsilon)} \left[\text{nnz}(\mathbf{A}) n^{\frac{1}{1+p}} + n^{1+\frac{2}{1+p}} \right]\right)$	Thms 32, 33
$p > 2$	general	$1/\gamma$	$O(p \text{nnz}(\mathbf{A}) \cdot n^\gamma)$	Thm 34
$p < 2$	general	$1/\gamma$	$\tilde{O}\left(\frac{1}{p^\gamma} \left[\text{nnz}(\mathbf{A}) n^{\frac{1/p-1/2}{1/p+1/2} + \gamma/2} + \sqrt{\text{nnz}(\mathbf{A})} \cdot n^{\frac{4/p-1}{2/p-1}} \right]\right)$	Thm 34

Table 1: Summary of our results for approximating the Schatten p -norms. We define $f(p, \epsilon) = \min\{1, p^3\} \cdot \epsilon^{\max\{3, 1+1/p\}}$, which appears as a factor in many of the bounds. The uniform sparsity assumption is that the maximum row sparsity $d_s(\mathbf{A}) \leq \frac{\xi}{n} \text{nnz}(\mathbf{A})$ for some constant ξ . In our theorems, we give general runtimes, parameterized by ξ . When we do not have the uniform sparsity assumption, we are still able to give a $(1 + \epsilon)$ approximation in $\tilde{O}(\epsilon^{-3} \text{nnz}(\mathbf{A}) \sqrt{n} + n^2)$ time for example for $\|\mathbf{A}\|_1$. We can also give $1/\gamma$ approximation for any constant $\gamma < 1$ by paying an $n^{\gamma/2}$ factor in our runtime. Note that for dense matrices, for all p we obtain $o(n^\omega)$ runtime, or $o(n^3)$ runtime if we do not use fast matrix multiplication.

When $f(x) = x^p$, $S_f(\mathbf{A}) = \|\mathbf{A}\|_p^p$ where $\|\mathbf{A}\|_p$ is the Schatten p -norm of $\mathbf{A}$. Computation of the Schatten 1-norm, also known as the nuclear or trace norm, is required in a wide variety of applications. It is often used in place of the matrix rank in matrix completion algorithms and other convex relaxations of rank-constrained optimization problems [CR12, DTV11, JNS13, NNS⁺14]. It appears as the ‘graph energy’ in theoretical chemistry [Gut92, Gut01], the ‘singular value bound’ in differential privacy [HLM12, LM12], and in rank aggregation and collaborative ranking [LN15].

Similar to the nuclear norm, general Schatten p -norms are used in convex relaxations for rank-constrained optimization [NHD12, NWC⁺12]. They also appear in image processing applications such as denoising and background subtraction [XGL⁺16], classification [LYCG14], restoration [XQT⁺16], and feature extraction [DHJZ15].

When $f(x) = -x \log x$ (after $\mathbf{A}$ is normalized by $\|\mathbf{A}\|_1$), $S_f(\mathbf{A})$ is the SVD entropy [ABB00], which is used in feature selection [VGLH06, BP14], financial data analysis [Car14, GXL15], and genomic data [ABB00] applications.

Despite their importance, prior to our work, few algorithms for fast computation of spectral sums existed. Only a few special cases of the Schatten p -norms were known to be computable efficiently in $o(n^\omega)$ time. These include the Frobenius norm ($p = 2$) which is trivially computed in $O(\text{nnz}(\mathbf{A}))$ time, the spectral norm ($p = \infty$) which can be estimated via the Lanczos method in $\tilde{O}(\text{nnz}(\mathbf{A}) \epsilon^{-\frac{1}{2}})$ time [KW92], and the Schatten- p norms for *even integers* $p > 2$, or general integers with PSD $\mathbf{A}$. These norms can be approximated in $O(\text{nnz}(\mathbf{A}) \epsilon^{-2})$ time via trace estimation [Woo14, BCKY16], since when p is even or $\mathbf{A}$ is PSD, $\mathbf{A}^p$ is PSD and so its trace equals $\|\mathbf{A}\|_p^p$.

There are a number of works which consider estimating matrix norms in sublinear space and with a small number of passes over $\mathbf{A}$ [LNW14, AKR15, LW16a, LW16b, BCKY16, LW17]. However, in these works, the main focus is on space complexity, and no non-trivial runtime bounds are given. We seem to be the first to tackle the arguably more fundamental problem of obtaining the best time complexity for simple norms like the Schatten- p norms, irrespective of space and pass complexity.

Another interesting line of works tries to estimate the Schatten- p norms of an underlying covariance matrix from a small number of samples from the distribution [KV16], or from entrywise sampling under various incoherence assumptions [KO17]. This model is different from ours, as we do not assume an underlying distribution or any incoherence properties. Moreover, even with such

assumptions, these algorithms also only give non-trivial sample complexity when either $\mathbf{A}$ is PSD and p is an integer, or $\mathbf{A}$ is a general matrix but p is an even integer, which as mentioned above are easy to handle from the perspective of time complexity alone.

A number of works have focused on computing spectral sums when $\mathbf{A}$ has bounded condition number, and relative error results exist for example for the log-determinant, $\text{tr}(\exp(\mathbf{A}))$, $\text{tr}(\mathbf{A}^{-1})$, and the Schatten p -norms [BDKZ15, HMAS17, UCS17]. We are the first to give relative error results in $o(\omega)$ time for general matrices, without the condition number dependency. Our histogram approach resembles spectral filtering and spectral density estimation techniques that have been considered in the numerical computation literature [ZWJ15, DNPS16, LSY16, US16, UCS17, USS17]. However, this literature typically requires assuming gaps between the singular values and existing work is not enough to give relative error spectral sum approximation for general matrices

1.3 Algorithmic Approach

Spectral Sums via Trace Estimation: A common approach to spectral sum approximation is to reduce to a trace estimation problem involving the PSD matrix $\mathbf{A}^T \mathbf{A}$, using the fact that the trace of this matrix equals the sum of its singular values. In fact, this has largely been the only known technique, other than the full SVD, for obtaining aggregate information about $\mathbf{A}$'s singular values [Hut90, SLO13, WLK⁺16, WWZ14, RKA15, FORF16, BDKZ15, HMAS17]. The idea is, letting $g(x) = f(x^{1/2})$, we have $\mathcal{S}_f(\mathbf{A}) = \mathcal{S}_g(\mathbf{A}^T \mathbf{A})$. Writing the SVD $\mathbf{A}^T \mathbf{A} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^T$, and defining the matrix function $g(\mathbf{A}^T \mathbf{A}) \stackrel{\text{def}}{=} \mathbf{U} g(\mathbf{\Lambda}) \mathbf{U}^T$ where $[g(\mathbf{\Lambda})]_{i,i} = g([\mathbf{\Lambda}]_{i,i})$, we have $\mathcal{S}_g(\mathbf{A}^T \mathbf{A}) = \text{tr}(g(\mathbf{A}^T \mathbf{A}))$ since, if $g(\cdot)$ is nonnegative, $g(\mathbf{A}^T \mathbf{A})$ is PSD and its trace equals the sum of its singular values.

It is well known that this trace can be approximated up to $(1 \pm \epsilon)$ accuracy by averaging $\tilde{O}(\epsilon^{-2})$ samples of the form $\mathbf{x}^T g(\mathbf{A}^T \mathbf{A}) \mathbf{x}$ where $\mathbf{x}$ is a random Gaussian or sign vector [Hut90, AT11]. While $g(\mathbf{A}^T \mathbf{A})$ cannot be explicitly computed without a full SVD, a common approach is to approximate g with a low-degree polynomial ϕ [BDKZ15, HMAS17]. If ϕ has degree q , one can apply $\phi(\mathbf{A}^T \mathbf{A})$ to any vector $\mathbf{x}$ in $O(\text{nnz}(\mathbf{A}) \cdot q)$ time, and so estimate its trace in just $O(\text{nnz}(\mathbf{A}) \cdot \frac{q}{\epsilon^2})$ time. Unfortunately, for many of the functions most important in applications, e.g., $f(x) = x^p$ for odd p , $f(x) = x \log x$, $f(x) = x^{-1}$, $g(x)$ has a discontinuity at $x = 0$ and *cannot* be approximated well by a low-degree polynomial near zero. While the approximation only needs to be good on the range $[\sigma_n^2(\mathbf{A}), \sigma_1^2(\mathbf{A})]$, the required degree q will still typically depend on $\sqrt{\kappa}$ where $\kappa \stackrel{\text{def}}{=} \frac{\sigma_1^2(\mathbf{A})}{\sigma_n^2(\mathbf{A})}$ is the condition number, which can be unbounded for general matrices.

Singular Value Deflation for Improved Conditioning: Our first observation is that, for many functions, it is not necessary to approximate $g(x)$ on the full spectral range. For example, for $g(x) = x^{p/2}$ (i.e., when $\mathcal{S}_g(\mathbf{A}^T \mathbf{A}) = \|\mathbf{A}\|_p^p$), setting $\lambda = (\frac{\epsilon}{n} \|\mathbf{A}\|_p^p)^{1/p}$, we have:

$$\sum_{\{i | \sigma_i(\mathbf{A}) \leq \lambda\}} \sigma_i(\mathbf{A})^p \leq n \cdot \frac{\epsilon}{n} \|\mathbf{A}\|_p^p \leq \epsilon \|\mathbf{A}\|_p^p.$$

Hence we can safely ‘ignore’ any $\sigma_i(\mathbf{A}) \leq \lambda$ and still obtain a relative error approximation to $\mathcal{S}_g(\mathbf{A}^T \mathbf{A}) = \|\mathbf{A}\|_p^p$. The larger p is, the larger we can set λ (corresponding to $(1 - \alpha)^T$ in Theorem 1) to be, since, after powering, the singular values below this threshold do not contribute significantly to $\|\mathbf{A}\|_p^p$. For $\|\mathbf{A}\|_p^p$, our ‘effective condition number’ for approximating $g(x)$ becomes $\hat{\kappa} = \frac{\sigma_1^2(\mathbf{A})}{\lambda^2} = (\frac{n}{\epsilon})^{2/p} \cdot \frac{\sigma_1^2(\mathbf{A})}{\|\mathbf{A}\|_p^2}$. Unfortunately, in the worst case, we may have $\sigma_1(\mathbf{A}) \approx \|\mathbf{A}\|_p$ and hence $\sqrt{\hat{\kappa}} = (\frac{n}{\epsilon})^{1/p}$.

Hiding ϵ dependences, this gives runtime $\tilde{O}(\text{nnz}(\mathbf{A}) \cdot n)$ when $p = 1$.

To improve the effective condition number, we can apply *singular vector deflation*. Our above bound on $\hat{\kappa}$ is only tight when the first singular value is very large and so dominates $\|\mathbf{A}\|_p$. We can remedy this by flattening $\mathbf{A}$'s spectrum by deflating off the top k singular vectors (corresponding to k in Theorem 1), and including their singular values in the spectral sum directly.

Specifically, letting $\mathbf{P}_k$ be the projection onto the top k singular vectors of $\mathbf{A}$, we consider the deflated matrix $\bar{\mathbf{A}} \stackrel{\text{def}}{=} \mathbf{A}(\mathbf{I} - \mathbf{P}_k)$, which has $\sigma_1(\bar{\mathbf{A}}) = \sigma_{k+1}(\mathbf{A})$. Importantly, $\sigma_{k+1}^p(\mathbf{A}) \leq \frac{1}{k} \|\mathbf{A}\|_p^p$, and thus this singular value cannot dominate the p -norm. As an example, considering $p = 1$ and ignoring ϵ dependencies, our effective condition number after deflation is

$$\hat{\kappa} = \frac{n^2 \cdot \sigma_{k+1}^2(\mathbf{A})}{\|\mathbf{A}\|_1^2} \leq \frac{n^2}{k^2} \quad (1)$$

The runtime required to approximate $\mathbf{P}_k$ via an iterative method (ignoring possible gains from fast matrix multiplication) is roughly $O(\text{nnz}(\mathbf{A})k + nk^2)$. We then require $\tilde{O}(\text{nnz}(\mathbf{A})\sqrt{\hat{\kappa}} + nk\sqrt{\hat{\kappa}})$ time to approximate the polynomial trace of $\bar{\mathbf{A}}^T \bar{\mathbf{A}}$. The $nk\sqrt{\hat{\kappa}}$ term comes from projecting off the top singular directions with each application of $\bar{\mathbf{A}}^T \bar{\mathbf{A}}$. Setting $k = \sqrt{n}$ to balance the costs, we obtain runtime $\tilde{O}(\text{nnz}(\mathbf{A})\sqrt{n} + n^2)$.

For $p \neq 1$ a similar argument gives runtime $\tilde{O}(\text{nnz}(\mathbf{A})n^{\frac{1}{p+1}} + n^{2+\frac{1}{p+1}})$. This is already a significant improvement over a full SVD. As p grows larger, the runtime approaches $\tilde{O}(\text{nnz}(\mathbf{A}))$ reflecting the fact that for larger p we can ignore a larger and larger portion of the small singular values in $\mathbf{A}$ and correspondingly deflate off fewer and fewer top values.

Unfortunately, we get stuck here. Considering the important Schatten-1 norm, for a matrix with $\sqrt{n}$ singular values each equal to $\sqrt{n}$ and $\Theta(n)$ singular values each equal to 1, the tail of small singular values contributes a constant fraction of $\|\mathbf{A}\|_1 = \Theta(n)$. However, there is no good polynomial approximation to $g(x) = x^{1/2}$ on the range $[1, n]$ with degree $o(\sqrt{n})$ (recall that we pick this function since $\mathcal{S}_g(\mathbf{A}^T \mathbf{A}) = \|\mathbf{A}\|_1$). So to accurately approximate $g(\mathbf{A}^T \mathbf{A})$, we either must deflate off all $\sqrt{n}$ top singular values, requiring $\Theta(\text{nnz}(\mathbf{A})\sqrt{n})$ time, or apply a $\Theta(\sqrt{n})$ degree polynomial approximation, requiring the same amount of time.

Further Improvements with Stochastic Gradient Descent: To push beyond this barrier, we look to *stochastic gradient* methods for linear systems. When using polynomial approximation, our bounds depend on the condition number of the interval over which we must approximate $g(\mathbf{A}^T \mathbf{A})$, after ignoring the smallest singular values and deflating off the largest. This is analogous to the condition number dependence of iterative linear system solvers like conjugate gradient or accelerated gradient descent, which approximate $f(\mathbf{A}^T \mathbf{A})$ for $f = 1/x$ using a polynomial of $\mathbf{A}^T \mathbf{A}$.

However, recent advances in convex optimization offer an alternative. Stochastic gradient methods [JZ13, SSZ14] sample one row, $\mathbf{a}_i$, of $\mathbf{A}$ at a time, updating the current iterate by adding a multiple of $\mathbf{a}_i$. They trade a larger number of iterations for updates that take $O(\text{nnz}(\mathbf{a}_i))$ time, rather than $O(\text{nnz}(\mathbf{A}))$ time to multiply $\mathbf{A}$ by a vector. At the same time, these methods give much finer dependencies on the singular value spectrum. Specifically, it is possible to approximately apply $(\mathbf{A}^T \mathbf{A})^{-1}$ to a vector with the number of iterations dependent on the *average condition number*:

$$\bar{\kappa} = \frac{\frac{1}{n} \sum_{i=1}^n \sigma_i^2(\mathbf{A})}{\sigma_n^2(\mathbf{A})}.$$

$\bar{\kappa}$ is always at most the standard condition number, $\kappa = \frac{\sigma_1^2(\mathbf{A})}{\sigma_n^2(\mathbf{A})}$. It can be significantly smaller when $\mathbf{A}$ has a quickly decaying spectrum, and hence $\frac{1}{n} \sum_{i=1}^n \sigma_i^2(\mathbf{A}) \ll \sigma_1^2(\mathbf{A})$. Further, the case of a quickly decaying spectrum with a few large and many small singular values is *exactly the hard case*

for our earlier approach. If we can understand how to translate improvements on linear system solvers to spectral sum approximation, we can handle this hard case.

From Linear System Solvers to Histogram Approximation: The key idea to translating the improved average condition number bounds for linear systems to our problem of approximating $\mathcal{S}_f(\mathbf{A})$ is to note that linear system solvers can be used to apply threshold functions to $\mathbf{A}^T \mathbf{A}$.

Specifically, given any vector $\mathbf{y}$, we can first compute $\mathbf{A}^T \mathbf{A} \mathbf{y}$. We can then apply a fast system solver to approximate $(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T \mathbf{A} \mathbf{y}$. The matrix function $r_\lambda(\mathbf{A}^T \mathbf{A}) \stackrel{\text{def}}{=} (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T \mathbf{A}$ has a number of important properties. All its singular values are between 0 and 1. Further, any singular value in $\mathbf{A}^T \mathbf{A}$ with value $\geq \lambda$ is mapped to a singular value in $r_\lambda(\mathbf{A}^T \mathbf{A})$ which is $\geq 1/2$. Correspondingly, any singular value $< \lambda$ is mapped to $< 1/2$.

Thus, we can apply a low degree polynomial approximation to a step function at $1/2$ to $r_\lambda(\mathbf{A}^T \mathbf{A})$ to obtain $s_\lambda(\mathbf{A}^T \mathbf{A})$, which approximates a step function at λ [FMMS16]. For some steepness parameter γ which affects the degree of the polynomial approximation, for $x \geq (1 + \gamma)\lambda$ we have $s_\lambda(x) \approx 1$ and for $x < (1 - \gamma)\lambda$, $s_\lambda(x) \approx 0$. On the intermediate range $x \in [(1 - \gamma)\lambda, (1 + \gamma)\lambda]$, $s_\lambda(x)$ falls somewhere between 0 and 1.

By composing these approximate threshold functions at different values of λ , it is possible to ‘split’ our spectrum into a number of small spectral windows. For example, $s_a(\mathbf{A}^T \mathbf{A}) \cdot (\mathbf{I} - s_b(\mathbf{A}^T \mathbf{A}))$ is ≈ 1 on the range $[a, b]$ and ≈ 0 outside this range, with some ambiguity near a and b .

Splitting our spectrum into windows of the form $[(1 - \alpha)^{t-1}, (1 - \alpha)^t]$ for a width parameter α , and applying trace estimation on each window lets us produce an approximate spectral histogram. Of course, this histogram is not exact and in particular, the ‘blurring’ of our windows at their boundaries can introduce significant error. However, by applying a random shifting technique and setting the steepness parameter γ small enough (i.e., $1/\text{poly}(\alpha, \epsilon)$), we can ensure that most of the spectral weight falls outside these boundary regions with good probability, giving Theorem 1.

From Histogram Approximation to Spectral Sums: If α is small enough, and $f(\cdot)$ (the function in the spectral sum) and correspondingly $g(\cdot)$ (where $g(x) = f(x^{1/2})$) are smooth enough, we can approximate $\mathcal{S}_f(\mathbf{A}) = \mathcal{S}_g(\mathbf{A}^T \mathbf{A})$ by simply summing over each bucket in the histogram, approximating $g(x)$ by its value at one end of the bucket.

This technique can be applied for any spectral sum. The number of windows required (controlled by α) and the histogram accuracy ϵ scale with the smoothness of $f(\cdot)$ and the desired accuracy in computing the sum, introducing polynomial runtime dependencies on these parameters.

However, the most important factor determining the final runtime is the smallest value λ (corresponding to $(1 - \alpha)^T$ in Theorem 1) which we must include in our histogram in order to approximate $\mathcal{S}_f(\mathbf{A})$. The cost of computing the last bucket of the histogram is proportional to the cost of applying $s_\lambda(\mathbf{A}^T \mathbf{A})$, and hence of approximately computing $(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T \mathbf{A} \mathbf{y}$. Using stochastic gradient descent this depends on the average condition number of $(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})$.

Again considering the Schatten 1-norm for illustration, we can ignore any singular values with $\sigma_i(\mathbf{A}) \leq \frac{\epsilon}{n} \|\mathbf{A}\|_1$. Hiding ϵ dependence, this means that in our histogram, we must include any singular values of $\mathbf{A}^T \mathbf{A}$ with value $\sigma_i(\mathbf{A}^T \mathbf{A}) = \sigma_i^2(\mathbf{A}) \geq \frac{1}{n^2} \|\mathbf{A}\|_1^2$. This gives us effective average condition number after deflating off the top k singular values:

$$\bar{\kappa} = \frac{n^2 \sum_{i=k+1}^n \sigma_i^2(\mathbf{A})}{n \|\mathbf{A}\|_1^2} \leq \frac{n^2 \sigma_{k+1}(\mathbf{A}) \cdot \sum_{i=k+1}^n \sigma_i(\mathbf{A})}{n \|\mathbf{A}\|_1^2} \leq \frac{n}{k} \quad (2)$$

where the last inequality follows from the observation that $\sigma_{k+1}(\mathbf{A}) \leq \frac{1}{k} \|\mathbf{A}\|_1$ and $\sum_{i=k+1}^n \sigma_i(\mathbf{A}) \leq \|\mathbf{A}\|_1$. Comparing to (1), this bound is better by an n/k factor.

Ignoring details and using a simplification of the runtimes in Theorem 1, we obtain an algorithm running in $\tilde{O}(\text{nnz}(\mathbf{A})k+nk^2)$ time to deflate k singular vectors, plus $\tilde{O}\left(\text{nnz}(\mathbf{A})\sqrt{\bar{\kappa}} + \sqrt{\text{nnz}(\mathbf{A})nk\bar{\kappa}}\right)$ time to approximate the spectral sum over the deflated matrix. Choosing k to balance these costs, gives our final runtimes. For the nuclear norm, using the bound on $\bar{\kappa}$ from (2), we set $k = n^{1/3}$ which gives $\bar{\kappa} = n^{2/3}$ and runtime $\tilde{O}(\text{nnz}(\mathbf{A})n^{1/3} + n^{3/2}\sqrt{d_s})$ where $d_s \leq n$ is the maximum row sparsity. For dense $\mathbf{A}$ this is $\tilde{O}(n^{2.33})$, which is faster than state of the art matrix multiplication time. It can be further accelerated using fast matrix multiplication methods. See details in Section 7.

Returning to our original hard example for intuition, we have $\mathbf{A}$ with $\sqrt{n}$ singular values at $\sqrt{n}$ and $\Theta(n)$ singular values at 1. Even without deflation, we have (again ignoring ϵ dependencies) $\bar{\kappa} = \frac{\sum_{i=1}^n \sigma_i^2(\mathbf{A})}{n\lambda} = \frac{n\|\mathbf{A}\|_F^2}{\|\mathbf{A}\|_1^2}$. Since $\|\mathbf{A}\|_F^2 = \Theta(n^{3/2})$ and $\|\mathbf{A}\|_1^2 = \Theta(n^2)$, this gives $\bar{\kappa} = \Theta(\sqrt{n})$. Thus, we can actually approximate $\|\mathbf{A}\|_1$ in just $\tilde{O}(\text{nnz}(\mathbf{A})n^{1/4})$ time for this matrix.

With average condition number dependence, our performance is limited by a new hard case. Consider $\mathbf{A}$ with $n^{1/3}$ singular values at $n^{2/3}$ and $\Theta(n)$ at 1. The average condition number without deflation is $\frac{n\|\mathbf{A}\|_F^2}{\|\mathbf{A}\|_1^2} = \Theta\left(\frac{n^{5/3}}{n}\right) = \Theta(n^{2/3})$ giving $\sqrt{\bar{\kappa}} = \Theta(n^{1/3})$. Further, we can see that unless we deflate off nearly all $n^{1/3}$ top singular vectors, we do not improve this bound significantly.

1.4 Lower Bound Approach

We now shift focus to our lower bounds, which explore the fundamental limits of spectrum approximation using fine-grained complexity approaches. Fine-grained complexity has had much success for graph problems, string problems, and problems in other areas (see, e.g., [Wil15] for a survey), and is closely tied to understanding the complexity of matrix multiplication. However, to the best of our knowledge it has not been applied broadly to problems in linear algebra.

Existing hardness results for linear algebraic problems tend to apply to restricted computational models such as arithmetic circuits [BS83], bilinear circuits or circuits with bounded coefficients and number of divisions [Mor73, Raz03, Shp03, RS03], algorithms for dense linear systems that can only add multiples of rows to each other [KKS65, KS70], and algorithms with restrictions on the dimension of certain manifolds defined in terms of the input [Win70, Win87, Dem13]. In contrast, we obtain conditional lower bounds for arbitrary polynomial time algorithms by showing that faster algorithms for them imply faster algorithms for canonical hard problems in fine-grained complexity.

From Schatten 3-norm to Triangle Detection: We start with the fact that the number of triangles in any unweighted graph G is equal to $\text{tr}(\mathbf{A}^3)/6$, where $\mathbf{A}$ is the adjacency matrix. Any algorithm for approximating $\text{tr}(\mathbf{A}^3)$ to high enough accuracy therefore gives an algorithm for detecting if a graph has at least one triangle.

$\mathbf{A}$ is not PSD, so $\text{tr}(\mathbf{A}^3)$ is actually not a function of $\mathbf{A}$'s singular values – it depends on the signs of $\mathbf{A}$'s eigenvalues. However, the graph Laplacian given by $\mathbf{L} = \mathbf{D} - \mathbf{A}$ where $\mathbf{D}$ is the diagonal degree matrix, is PSD and we have:

$$\|\mathbf{L}\|_3^3 = \text{tr}(\mathbf{L}^3) = \text{tr}(\mathbf{D}^3) - 3\text{tr}(\mathbf{D}^2\mathbf{A}) + 3\text{tr}(\mathbf{D}\mathbf{A}^2) - \text{tr}(\mathbf{A}^3).$$

$\text{tr}(\mathbf{D}^2\mathbf{A}) = 0$ since $\mathbf{A}$ has an all 0 diagonal. Further, it is not hard to see that $\text{tr}(\mathbf{D}\mathbf{A}^2) = \text{tr}(\mathbf{D}^2)$. So this term and $\text{tr}(\mathbf{D}^3)$ are easy to compute exactly. Thus, if we approximate $\|\mathbf{L}\|_3^3$ up to additive error 6, we can determine if $\text{tr}(\mathbf{A}^3) = 0$ or $\text{tr}(\mathbf{A}^3) \geq 6$ and so detect if G contains a triangle. $\|\mathbf{L}\|_3^3 \leq 8n^4$ for any unweighted graph on n nodes, and hence computing this norm up to $(1 \pm \epsilon)$ relative error for $\epsilon = 3/(6n^4)$ suffices to detect a triangle. If we have an $O(n^\gamma \epsilon^{-c})$ time $(1 \pm \epsilon)$ approximation algorithm for the Schatten 3-norm, we can thus perform triangle detection in $O(n^{\gamma+4c})$ time.

Our strongest algorithmic result for the Schatten 3-norm requires just $\tilde{O}(n^2/\epsilon^3)$ time for dense matrices. Improving the ϵ dependence to $o(1/\epsilon^{(\omega-2)/4}) = O(1/\epsilon^{.09})$ for the current value of ω , would yield an algorithm for triangle detection running in $o(n^\omega)$ time for general graphs, breaking a longstanding runtime barrier for this problem. Even a $1/\epsilon^{1/3}$ dependence would give a sub-cubic time triangle detection algorithm, and hence could be used to give a subcubic time matrix multiplication algorithm via the reduction of [WW10].

Generalizing to Other Spectral Sums We can generalize the above approach to the Schatten 4-norm by adding λ self-loops to each node of G , which corresponds to replacing $\mathbf{A}$ with $\lambda\mathbf{I} + \mathbf{A}$. We then consider $\text{tr}((\lambda\mathbf{I} + \mathbf{A})^4) = \|\lambda\mathbf{I} + \mathbf{A}\|_4^4$. This is the sum over all vertices of the number of paths that start at v_i and return to v_i in four steps. All of these paths are either (1) legitimate four cycles, (2) triangles combined with self loops, or (3) combinations of self-loops and two-step paths from a vertex v_i to one of its neighbors and back. The number of type (3) paths is exactly computable using the node degrees and number of self loops. Additionally, if the number of self loops λ is large enough, the number of type (2) paths will dominate the number of type (1) paths, even if there is just a single triangle in the graph. Hence, an accurate approximation to $\|\lambda\mathbf{I} + \mathbf{A}\|_4^4$ will give us the number of type (2) paths, from which we can easily compute the number of triangles.

This argument extends to a very broad class of spectral sums by considering a power series expansion of $f(x)$ and showing that for large enough λ , $\text{tr}(f(\lambda\mathbf{I} + \mathbf{A}))$ is dominated by $\text{tr}(\mathbf{A}^3)$ along with some exactly computable terms. Thus, an accurate approximation to this spectral sum allows us to determine the number of triangles in G . This approach works for any $f(x)$ that can be represented as a power series, with reasonably well-behaved coefficients on some interval of $\mathbb{R}^+$, giving bounds for all $\|\mathbf{A}\|_p$ with $p \neq 2$, the SVD entropy, $\log \det(\mathbf{A})$, $\text{tr}(\mathbf{A}^{-1})$, and $\text{tr}(\exp(\mathbf{A}))$.

We further show that approximating $\text{tr}(\mathbf{A}^{-1})$ for the $\mathbf{A}$ used in our lower bound can be reduced to computing all effective resistances of a certain graph Laplacian up to $(1 \pm \epsilon)$ error. Thus, we rule out highly accurate (with $1/\epsilon^c$ dependence for small c) approximation algorithms for all effective resistances, despite the existence of linear time system solvers (with $\log(1/\epsilon)$ error dependence) for Laplacians [ST04]. Effective resistances and leverage scores are quantities that have recently been crucial to achieving algorithmic improvements to fundamental problems like graph sparsification [SS08] and regression [LMP13, CLM⁺15]. While crude multiplicative approximations to the quantities suffice for these problems, more recently computing these quantities has been used to achieve breakthroughs in solving maximum flow and linear programming [LS14], cutting plane methods [LSW15], and sampling random spanning trees [KM09, MST15]. In each of these settings having more accurate estimates would be a natural route to either simplify or possibly improve existing results; we show that this is unlikely to be successful if the precision requirements are two high.

1.5 Paper Outline

Section 2: Preliminaries. We review notations that will be used throughout.

Section 3: Spectral Window Approximation. We show how to approximately restrict the spectrum of a matrix to a small window, which will be our main primitive for accessing the spectrum.

Section 4: Spectral Histogram Approximation. We show how our spectral window algorithms can be used to compute an approximate spectral histogram. We give applications to approximating general spectral sums, including the Schatten p -norms, Orlicz norms, and Ky Fan norms.

Section 5: Lower Bounds. We prove lower bounds showing that highly accurate spectral sum algorithms can be used to give algorithms for triangle detection and matrix multiplication.

Section 6: Improved Algorithms via Polynomial Approximation. We demonstrate how to tighten ϵ dependencies in our runtimes using a more general polynomial approximation approach.

Section 7: Optimized Runtime Bounds. We instantiate the techniques of Section 6 give our best runtimes for the Schatten p -norms and SVD entropy.

2 Preliminaries

Here we outline notation and conventions used throughout the paper.

Matrix Properties: For $\mathbf{A} \in \mathbb{R}^{n \times d}$ we assume without loss of generality that $d \leq n$. We let $\sigma_1(\mathbf{A}) \geq \dots \geq \sigma_d(\mathbf{A}) \geq 0$ denote the matrix's singular values, $\text{nnz}(\mathbf{A})$ denote the number of non-zero entries, and $d_s(\mathbf{A})$ denote the maximum number of non-zero entries in any row. Note that $d_s(\mathbf{A}) \in [\text{nnz}(\mathbf{A})/n, d]$.

Fast Matrix Multiplication: Let $\omega \approx 2.3729$ denote the current best exponent of fast matrix multiplication [Wil12, GU17]. Additionally, let $\omega(\gamma)$ denote the exponent such that it takes $O(d^{\omega(\gamma)})$ time to multiply a $d \times d$ matrix by a $d \times d^\gamma$ matrix for any $\gamma \leq 1$. $\omega(\gamma) = 2$ for $\gamma < \alpha$ where $\alpha > 0.31389$ and $\omega(\gamma) = 2 + (\omega - 2)\frac{\gamma - \alpha}{1 - \alpha}$ for $\gamma \geq \alpha$ [LG12, GU17]. For $\gamma = 1$, $\omega(\gamma) = \omega$.

Asymptotic Notation: We use $\tilde{O}(\cdot)$ notation to hide poly-logarithmic factors in the input parameters, including dimension, failure probability, and error ϵ . We use ‘with high probability’ or ‘w.h.p.’ to refer to events happening with probability at least $1 - 1/d^c$ for some constant c , where d is our smaller input dimension.

Other: We denote $[d] \stackrel{\text{def}}{=} \{0, \dots, d\}$. For any $\mathbf{y} \in \mathbb{R}^d$ and PSD $\mathbf{N} \in \mathbb{R}^{d \times d}$, we denote $\|\mathbf{y}\|_{\mathbf{N}} \stackrel{\text{def}}{=} \sqrt{\mathbf{y}^T \mathbf{N} \mathbf{y}}$.

3 Approximate Spectral Windows via Ridge Regression

In this section, we give state-of-the-art results for approximating spectral windows over $\mathbf{A}$. As discussed, our algorithms will split $\mathbf{A}$'s spectrum into small slices using these window functions, performing trace estimation to estimate the number of singular values on each window and producing an approximate spectral histogram.

In Section 3.1 we show how to efficiently apply smooth approximations to threshold functions of the spectrum provided we have access to an algorithm for solving regularized regression problems with the matrix. In Section 3.2 we then provide the fastest known algorithms for the regression problems in the given parameter regimes using both stochastic gradient methods and traditional solvers. Departing from our high level description in Section 1.3, we actually incorporate singular vector deflation directly into our system solvers to reduce condition number. This simplifies our final algorithms but has the same effect as the deflation techniques discussed in Section 1.3. Finally, in Section 3.3 we provide algorithms and runtime analysis for applying smooth approximations to window functions of the spectrum, which is the main export of this section.

3.1 Step Function Approximation

To compute a window over $\mathbf{A}$'s spectrum, we will combine two threshold functions at the boundaries of the window. We begin by discussing how to compute these threshold functions.

Let $s_\lambda : [0, 1] \rightarrow [0, 1]$ be the threshold function at λ . $s_\lambda(x) = 1$ for $x \in [\lambda, 1]$ and 0 for $x \in [0, \lambda)$. For some gap γ we define a soft step function by:

Definition 2 (Soft Step Function). $s_\lambda^\gamma : [0, 1] \rightarrow [0, 1]$ is a γ -soft step at $\lambda > 0$ if:

$$s_\lambda^\gamma(x) = \begin{cases} 0 & \text{for } x \in [0, (1 - \gamma)\lambda] \\ 1 & \text{for } x \in [\lambda, 1] \end{cases} \quad \text{and } s_\lambda^\gamma(x) \in [0, 1] \text{ for } x \in [(1 - \gamma)\lambda, \lambda]. \quad (3)$$

We use the strategy from [FMMS16], which, for $\mathbf{A}$ with $\|\mathbf{A}\|_2 \leq 1$ shows how to efficiently multiply a γ -soft step $s_\lambda^\gamma(\mathbf{A}^T \mathbf{A})$ by any $\mathbf{y} \in \mathbb{R}^d$ using ridge regression. The trick is to first approximately compute $\mathbf{A}^T \mathbf{A}(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{y} = r_\lambda(\mathbf{A}^T \mathbf{A}) \mathbf{y}$ where $r_\lambda(x) \stackrel{\text{def}}{=} \frac{x}{x + \lambda}$. Then, note that $s_{1/2}(r_\lambda(x)) = s_\lambda(x)$. Additionally, the symmetric step function $s_{1/2}$ can be well approximated with a low degree polynomial. Specifically, there exists a polynomial of degree $O(\gamma^{-1} \log(1/(\gamma\epsilon)))$ that is within additive ϵ of a true γ -soft step at $1/2$ and can be applied stably such that any error in computing $r_\lambda(\mathbf{A}^T \mathbf{A})$ remains bounded. The upshot, following from Theorem 7.4 of [AZL16] is:

Lemma 3 (Step Function via Ridge Regression). *Let $\mathcal{A}(\mathbf{A}, \mathbf{y}, \lambda, \epsilon)$ be an algorithm that on input $\mathbf{A} \in \mathbb{R}^{n \times d}$, $\mathbf{y} \in \mathbb{R}^d$, $\lambda, \epsilon > 0$ returns $\mathbf{x} \in \mathbb{R}^d$ such that $\|\mathbf{x} - (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{y}\|_2 \leq \epsilon \|\mathbf{y}\|_2$ with high probability. Then there is an algorithm $\mathcal{B}(\mathbf{A}, \mathbf{y}, \lambda, \gamma, \epsilon)$ which on input $\mathbf{A} \in \mathbb{R}^{n \times d}$ with $\|\mathbf{A}\|_2 \leq 1$, $\mathbf{y} \in \mathbb{R}^d$, $\lambda \in (0, 1)$, and $\gamma, \epsilon > 0$, returns $\mathbf{x} \in \mathbb{R}^d$ with*

$$\|\mathbf{x} - s_\lambda^\gamma(\mathbf{A}^T \mathbf{A}) \mathbf{y}\|_2 \leq \epsilon \|\mathbf{y}\|_2$$

where s_λ^γ is a γ -soft step at λ (i.e. satisfies Definition 2). $\mathcal{B}(\mathbf{A}, \mathbf{y}, \lambda, \gamma, \epsilon)$ requires $O(\gamma^{-1} \log(1/(\gamma\epsilon)))$ calls to $\mathcal{A}(\mathbf{A}, \mathbf{y}, \lambda, \epsilon')$ along with $O(\text{nnz}(\mathbf{A}))$ additional runtime, where $\epsilon' = \text{poly}(1/(\gamma\epsilon))$.

3.2 Ridge Regression

Given Lemma 3, to efficiently compute $s_\lambda^\gamma(\mathbf{A}^T \mathbf{A}) \mathbf{y}$ for $s_\lambda^\gamma(\cdot)$ satisfying Definition 2, it suffices to quickly approximate $(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{y}$ (i.e. to provide the algorithm $\mathcal{A}(\mathbf{A}, \mathbf{y}, \lambda, \epsilon)$ used in the lemma). In this section we provide two theorems which give the state-of-the-art ridge regression running times achievable in our parameter regime, using sampling, acceleration, and singular value deflation.

Naively, computing $(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{y}$ using an iterative system solver involves a dependence on the condition number $\sigma_1^2(\mathbf{A})/\lambda$. In our theorems, this condition number is replaced by a deflated condition number depending on $\sigma_k^2(\mathbf{A})$ for some input parameter $k \in [d]$. We achieve this improved dependence following the techniques presented in [GOSS16]. We first approximate the top k singular vectors of $\mathbf{A}$ and then construct a preconditioner based on this approximation, which significantly flattens the spectrum of the matrix. By using this preconditioner in conjunction with a stochastic gradient based linear system solver, we further enjoy an average condition number dependence. The following theorem summarizes the results.

Theorem 4 (Ridge Regression – Accelerated Preconditioned SVRG). *For any $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $\lambda > 0$, let $\mathbf{M}_\lambda \stackrel{\text{def}}{=} \mathbf{A}^T \mathbf{A} + \lambda \mathbf{I}$. Let $\bar{\kappa} \stackrel{\text{def}}{=} \frac{k\sigma_k^2(\mathbf{A}) + \sum_{i=k+1}^d \sigma_i^2(\mathbf{A})}{d\lambda}$ where $k \in [d]$ is an input parameter. There is an algorithm that builds a preconditioner for $\mathbf{M}_\lambda$ using precomputation time $\tilde{O}(\text{nnz}(\mathbf{A})k + dk^{\omega-1})$ for sparse $\mathbf{A}$ or $\tilde{O}(nd^{\omega(\log_d k)-1})$ time for dense $\mathbf{A}$, and for any input $\mathbf{y} \in \mathbb{R}^d$, returns $\mathbf{x}$ such that with high probability $\|\mathbf{x} - \mathbf{M}_\lambda^{-1} \mathbf{y}\|_{\mathbf{M}_\lambda} \leq \epsilon \|\mathbf{y}\|_{\mathbf{M}_\lambda^{-1}}$ in*

$$\tilde{O}\left(\text{nnz}(\mathbf{A}) + \sqrt{\text{nnz}(\mathbf{A})[d \cdot d_s(\mathbf{A}) + dk]\bar{\kappa}}\right)$$

time for sparse $\mathbf{A}$ or $\tilde{O}(nd + n^{1/2}d^{3/2}\sqrt{\bar{\kappa}})$ time for dense $\mathbf{A}$.

Proof. We give a proof in Appendix A. Note that the ϵ dependence in the runtime is $\log(1/\epsilon)$ and so is hidden by the $\tilde{O}(\cdot)$ notation. $\square$

When $\mathbf{A}$ is dense, the runtime of Theorem 4 is essentially the best known. Due to its average condition number dependence, the method always outperforms traditional iterative methods, like conjugate gradient, up to log factors. However, in the sparse case, traditional approaches can give faster runtimes if the rows of $\mathbf{A}$ are not uniformly sparse and $d_s(\mathbf{A})$ is large. We have the following, also proved in Appendix A using the same deflation-based preconditioner as in Theorem 4:

Theorem 5 (Ridge Regression – Preconditioned Iterative Method). *For any $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $\lambda > 0$, let $\mathbf{M}_\lambda \stackrel{\text{def}}{=} \mathbf{A}^T \mathbf{A} + \lambda \mathbf{I}$ and $\hat{\kappa} \stackrel{\text{def}}{=} \frac{\sigma_{k+1}^2(\mathbf{A})}{\lambda}$ where $k \in [d]$ is an input parameter. There is an algorithm that builds a preconditioner for $\mathbf{M}_\lambda$ using precomputation time $\tilde{O}(\text{nnz}(\mathbf{A})k + dk^{\omega-1})$, and for any input $\mathbf{y} \in \mathbb{R}^d$, returns $\mathbf{x}$ such that with high probability $\|\mathbf{x} - \mathbf{M}_\lambda^{-1} \mathbf{y}\|_{\mathbf{M}_\lambda} \leq \epsilon \|\mathbf{y}\|_{\mathbf{M}_\lambda^{-1}}$ in $\tilde{O}\left((\text{nnz}(\mathbf{A}) + dk) \lceil \sqrt{\hat{\kappa}} \rceil\right)$ time.*

3.3 Overall Runtimes For Spectral Windows

Combined with Lemma 3, the ridge regression routines above let us efficiently compute soft step functions of $\mathbf{A}$'s spectrum. Composing step functions then gives our key computational primitive: the ability to approximate soft window functions that restrict $\mathbf{A}$'s spectrum to a specified range. We first define our notion of soft window functions and then discuss runtimes. The corresponding Theorem 7 is our main tool for spectrum approximation in the remainder of the paper.

Definition 6 (Soft Window Function). *Given $a, b > 0$ with $a < b$, and $\gamma \in [0, 1]$, $h_{[a,b]}^\gamma : [0, 1] \rightarrow [0, 1]$ is a γ -soft window for the range $[a, b]$ if:*

$$h_{[a,b]}^\gamma(x) = \begin{cases} 1 & \text{for } x \in [a, b] \\ 0 & \text{for } x \in [0, (1-\gamma)a] \cup [(1+\gamma)b, 1] \end{cases}$$

and $h_{[a,b]}^\gamma(x) \in [0, 1]$ for $x \in [(1-\gamma)a, a] \cup [b, (1+\gamma)b]$.

Theorem 7 (Spectral Windowing). *For any $\mathbf{A} \in \mathbb{R}^{n \times d}$ with $\|\mathbf{A}\|_2 \leq 1$, $\mathbf{y} \in \mathbb{R}^d$, and $a, b, \gamma, \epsilon \in (0, 1]$, with $a < b$, there is an algorithm $\mathcal{W}(\mathbf{A}, \mathbf{y}, a, b, \gamma, \epsilon)$ that returns $\mathbf{x}$ satisfying w.h.p.:*

$$\left\| \mathbf{x} - h_{[a,b]}^\gamma(\mathbf{A}^T \mathbf{A}) \mathbf{y} \right\|_2 \leq \epsilon \|\mathbf{y}\|_2$$

where $h_{[a,b]}^\gamma$ is a soft window function satisfying Def. 6. Let $\bar{\kappa} \stackrel{\text{def}}{=} \frac{k\sigma_k^2(\mathbf{A}) + \sum_{i=k+1}^d \sigma_i^2(\mathbf{A})}{d \cdot a}$ and $\hat{\kappa} \stackrel{\text{def}}{=} \frac{\sigma_{k+1}^2(\mathbf{A})}{a}$ where $k \in [d]$ is an input parameter. The algorithm uses precomputation time $\tilde{O}(\text{nnz}(\mathbf{A})k + dk^{\omega-1})$ for sparse $\mathbf{A}$ or $\tilde{O}(nd^{\omega(\log_d k)-1})$ for dense $\mathbf{A}$ after which given any $\mathbf{y}$ it returns $\mathbf{x}$ in time:

$$\tilde{O}\left(\frac{\text{nnz}(\mathbf{A}) + \sqrt{\text{nnz}(\mathbf{A})[d \cdot d_s(\mathbf{A}) + dk]\bar{\kappa}}}{\gamma}\right) \quad \text{or} \quad \tilde{O}\left(\frac{(\text{nnz}(\mathbf{A}) + dk) \lceil \sqrt{\hat{\kappa}} \rceil}{\gamma}\right)$$

for sparse $\mathbf{A}$ or $\tilde{O}\left(\frac{nd + n^{1/2} d^{3/2} \sqrt{\bar{\kappa}}}{\gamma}\right)$ for dense $\mathbf{A}$.

Proof. If $b \geq 1/(1+\gamma)$ then we can simply define $h_{[a,b]}^\gamma(x) = s_a^\gamma(x)$ for any s_a^γ satisfying Definition 2. Otherwise, given soft steps s_a^γ and $s_{(1+\gamma)b}^{\gamma/2}$ satisfying Definition 2, we can define $h_{[a,b]}^\gamma(x) = s_a^\gamma(x) \cdot (1 - s_{(1+\gamma)b}^{\gamma/2}(x))$. Since $\frac{\gamma}{2} \leq \frac{\gamma}{1+\gamma}$ we can verify that this will be a valid soft window function for $[a, b]$ (i.e. satisfy Definition 6). Further, we have for any $\mathbf{y} \in \mathbb{R}^d$:

$$h_{[a,b]}^\gamma(\mathbf{A}^T \mathbf{A}) \mathbf{y} = s_a^\gamma(\mathbf{A}^T \mathbf{A}) (\mathbf{I} - s_{(1+\gamma)b}^{\gamma/2}(\mathbf{A}^T \mathbf{A})) \mathbf{y}. \quad (4)$$

We can compute $s_a^\gamma(\mathbf{A}^T \mathbf{A})\mathbf{y}$ and $s_{(1+\gamma)b}^{\gamma/2}(\mathbf{A}^T \mathbf{A})\mathbf{y}$ each up to error $\epsilon \|\mathbf{y}\|_2$ via Lemma 3. This gives the error bound in the theorem, since we have both $\|s_a^\gamma(\mathbf{A}^T \mathbf{A})\|_2 \leq 1$ and $\|\mathbf{I} - s_{(1+\gamma)b}^{\gamma/2}(\mathbf{A}^T \mathbf{A})\|_2 \leq 1$ so the computation in (4) does not amplify error. Our runtime follows from combining Theorems 4 and 5 with $\lambda = a, b$ with Lemma 3. The errors in these theorems are measured with respect to $\|\cdot\|_{\mathbf{M}_\lambda}$. To obtain the error in $\|\cdot\|_2$ as used by Lemma 3, we simply apply the theorems with $\epsilon' = \epsilon \kappa(\mathbf{M}_\lambda)$ which incurs an additional $\log(\kappa(\mathbf{M}_\lambda))$ cost. Since $a < b$ the runtime is dominated by the computation of $s_a^\gamma(\mathbf{A}^T \mathbf{A})\mathbf{y}$, which depends on the condition number $\bar{\kappa} \stackrel{\text{def}}{=} \frac{k\sigma_k^2(\mathbf{A}) + \sum_{i=k+1}^d \sigma_i^2(\mathbf{A})}{d \cdot a}$ when using SVRG (Theorem 4) or $\hat{\kappa} \stackrel{\text{def}}{=} \frac{\sigma_{k+1}^2(\mathbf{A})}{a}$ for a traditional iterative solver (Theorem 5). $\square$

4 Approximating Spectral Sums via Spectral Windows

We now use the window functions discussed in Section 3 to compute an approximate spectral histogram of $\mathbf{A}$. We give our main histogram algorithm and approximation guarantee in Section 4.1. In Section 4.2 we show how this guarantee translates to accurate spectral sum approximation for any smooth and sufficiently quickly growing function $f(x)$. In Section 4.3 we apply this general result to approximating the Schatten p -norms for all real $p > 0$, bounded Orlicz norms, and the Ky Fan norms.

4.1 Approximate Spectral Histogram

Our main histogram approximation method is given as Algorithm 1. The algorithm is reasonably simple. Assuming $\|\mathbf{A}\|_2 \leq 1$ (this is w.l.o.g. as we can just scale the matrix), and given cutoff λ , below which we will not evaluate $\mathbf{A}$'s spectrum, we split the range $[\lambda, 1]$ into successive windows $R_0, \dots, R_T$ where $R_t = [a_1(1 - \alpha)^t, a_1(1 - \alpha)^{t-1}]$. Here α determines the width of our windows. In our final spectral approximation algorithms, we will set $\alpha = \Theta(\epsilon)$. a_1 is a random shift, which insures that, in expectation, the boundaries of our soft windows do not overlap too many singular values. This argument requires that most of the range $[\lambda, 1]$ is not covered by boundary regions. Thus, we set the steepness parameter $\gamma = \Theta(\epsilon_2 \alpha)$ where ϵ_2 will control the error introduced by the boundaries. Finally, we iterate over each window, applying trace estimation to approximate the singular value count in each window.

In our final algorithms, the number of windows and samples required for trace estimation will be $\tilde{O}(\text{poly}(1/\epsilon))$. The dominant runtime cost will come from computing the window for the lowest range R_T , which will incur a dependence on the condition number of $\mathbf{A}^T \mathbf{A} + a_T \mathbf{I}$ with $a_T = \Theta(\lambda)$.

Theorem 8 (Histogram Approximation). *Let $a_1, \tilde{b}_0, \dots, \tilde{b}_T$ be output by Algorithm 1. Let $R_0 = [a_1, 1]$, $R_t = [a_1(1 - \alpha)^t, a_1(1 - \alpha)^{t-1}]$ for $t \geq 1$, and $b_t = |\{i : \sigma_i^2(\mathbf{A}) \in R_t\}|$ be the number of squared singular values of $\mathbf{A}$ on the range R_t . Then, for sufficiently small constants c_1, c_2, c_3 , with probability 99/100, for all $t \in \{0, \dots, \lceil \log_{(1-\alpha)} \lambda \rceil\}$, $\tilde{b}_t$ output by Algorithm 1 satisfies:*

$$(1 - \epsilon_1)b_t \leq \tilde{b}_t \leq (1 + \epsilon_1)b_t + \lceil \log_{(1-\alpha)} \lambda \rceil \cdot \epsilon_2(b_{t-1} + b_t + b_{t+1}).$$

That is, $\tilde{b}_t$ approximates the number of singular values of the range R_t up to multiplicative $(1 \pm \epsilon_1)$ error and additive error $\lceil \log_{(1-\alpha)} \lambda \rceil \cdot \epsilon_2(b_{t-1} + b_t + b_{t+1})$. Note that by setting $\epsilon_2 \leq \frac{\epsilon_1}{\lceil \log_{(1-\alpha)} \lambda \rceil}$, the error on each bucket is just multiplicative on its size plus the size of the two adjacent buckets, which contain singular values in nearby ranges. For simplicity we assume $\mathbf{A}$ passed to the algorithm has $\|\mathbf{A}\|_2 \leq 1$. This is without loss of generality: we can estimate $\|\mathbf{A}\|_2$ in $\tilde{O}(\text{nnz}(\mathbf{A}))$ time via the power or Lanczos methods [KW92, MM15], and scale down the matrix appropriately.

Algorithm 1 Approximate Spectral Histogram

Input: $\mathbf{A} \in \mathbb{R}^{n \times d}$ with $\|\mathbf{A}\|_2 \leq 1$, accuracy parameters $\epsilon_1, \epsilon_2 \in (0, 1)$, width parameter $\alpha \in (0, 1)$, and minimum singular value parameter $\lambda \in (0, 1)$.

Output: Set of range boundaries $a_{T+1} < a_T < \dots < a_1 < a_0$ and counts $\{\tilde{b}_0, \dots, \tilde{b}_T\}$ where $\tilde{b}_t$ approximates the number of squared singular values of $\mathbf{A}$ on $[a_{t+1}, a_t]$.

Set $\gamma = c_1 \epsilon_2 \alpha$, $T = \lceil \log_{(1-\alpha)} \lambda \rceil$, and $S = \frac{\log n}{c_2 \epsilon_1^2}$.

Set $a_0 = 1$ and choose a_1 uniformly at random in $[1 - \alpha/4, 1]$.

Set $a_t = a_1(1 - \alpha)^{t-1}$ for $2 \leq t \leq T + 1$.

for $t = 0 : T$ **do**

Set $\tilde{b}_t = 0$.

for $s = 1 : S$ **do**

Choose $\mathbf{y} \in \{-1, 1\}^d$ uniformly at random.

Set $\tilde{b}_t = \tilde{b}_t + \frac{1}{S} \cdot \mathbf{y}^T \mathcal{W}(\mathbf{A}^T \mathbf{A}, \mathbf{y}, a_{t+1}, a_t, \gamma, c_3 \epsilon_1^2/n)$.

If $\tilde{b}_t \leq 1/2$ set $\tilde{b}_t = 0$.

end for

end for

return a_1 and $\tilde{b}_t$ for $t = 0 : T$.

▷ Iterate over histogram buckets.

▷ Initialize bucket size estimate.

▷ Estimate bucket size via trace estimation.

▷ Apply soft window via Thm 7.

▷ Round small estimates to ensure relative error.

▷ Output histogram representation.

Ignoring logarithmic and ϵ dependencies, the runtime of Algorithm 1 is dominated by the calls to $\mathcal{W}$ for the bucket corresponding to the smallest singular values, with $a_T = \Theta(\lambda)$. This runtime is given by Theorem 7. Since balancing the deflation parameter k in that theorem with the minimum squared singular value λ considered can be complex, we wait to instantiate full runtimes until employing Algorithm 1 for specific spectral sum computations.

Proof. We use the notation of Algorithm 1, where $a_0 = 1$, a_1 is chosen uniformly in $[1 - \alpha/4, 1]$ and $a_t = a_1(1 - \alpha)^{t-1}$. With this notation, we have $R_t = [a_{t+1}, a_t]$. Let $h_{R_t}^\gamma$ be a γ -soft window for R_t (Definition 6) and let $\bar{R}_t = [(1 - \gamma)a_{t+1}, (1 + \gamma)a_t]$ be the interval on which $h_{R_t}^\gamma$ is nonzero. $\tilde{b}_t$ is an estimation of the trace of such a window applied to $\mathbf{A}^T \mathbf{A}$. We first show that, if these traces are computed exactly, they accurately estimate the singular value counts in each range R_t . We have:

$$\begin{aligned} \text{tr}(h_{R_t}^\gamma(\mathbf{A}^T \mathbf{A})) &= \sum_{\sigma_i^2(\mathbf{A}) \in R_t} h_{R_t}^\gamma(\sigma_i^2(\mathbf{A})) + \sum_{\sigma_i^2(\mathbf{A}) \in \bar{R}_t \setminus R_t} h_{R_t}^\gamma(\sigma_i^2(\mathbf{A})) \\ &= b_t + \sum_{\sigma_i^2(\mathbf{A}) \in \bar{R}_t \setminus R_t} h_{R_t}^\gamma(\sigma_i^2(\mathbf{A})). \end{aligned} \quad (5)$$

We can bound the second term using the random shift a_1 . Since $\gamma = c_1 \epsilon_2 \alpha < \alpha$, each singular value falls within at most two extended ranges: $\bar{R}_t$ and $\bar{R}_{t \pm 1}$ for some t . Let I be the set of indices whose singular values fall within two ranges. We have $i \in I$ only if $a_1(1 - \alpha)^t \in (1 \pm \gamma)\sigma_i^2(\mathbf{A})$ for some t . Letting $d = \lceil \log_{\sigma_i^2(\mathbf{A})} (1 - \alpha) \rceil$, this holds only if $a_1 \in (1 \pm \gamma) \left(\frac{\sigma_i^2(\mathbf{A})}{(1 - \alpha)^d} \right)$, which occurs with probability at most $\frac{8\gamma}{\alpha}$ since a_1 is chosen uniformly in the range $[1 - \alpha/4, 1]$. Thus we have $\mathbb{P}[i \in I] \leq \frac{8\gamma}{\alpha}$. Further, $a_t \stackrel{\text{def}}{=} a_1(1 - \alpha)^{t-1}$ is distributed uniformly in the range $[(1 - \alpha/4)(1 - \alpha)^{t-1}, (1 - \alpha)^{t-1}]$, so we know for certain that if the constant c_1 on γ is set small enough:

$$\begin{aligned} \bar{R}_t &\subset [(1 - \gamma)(1 - \alpha/4)(1 - \alpha)^t, (1 + \gamma)(1 - \alpha)^{t-1}] \\ &\subset [(1 - \alpha/3)(1 - \alpha)^t, (1 + \alpha/3)(1 - \alpha)^{t-1}] \\ &\subset [(1 - \alpha)^{t+1}, (1 - \alpha/4)(1 - \alpha)^{t-2}]. \end{aligned} \quad (6)$$

Let $M_t \stackrel{\text{def}}{=} [(1 - \alpha)^{t+1}, (1 - \alpha/4)(1 - \alpha)^{t-2}]$. Note that M_t is fixed (i.e. not a random variable). Regardless of the random shift a_1 , by (6), we always have $\bar{R}_t \subset M_t$. We also have $M_t \subset R_{t-1} \cup R_t \cup R_{t+1}$.

Let $\mathbb{I}[i \in I]$ be 1 if $i \in I$ and 0 otherwise. We can upper bound the second term of (5) by:

$$\begin{aligned} \sum_{\sigma_i^2(\mathbf{A}) \in \bar{R}_t \setminus R_t} h_{R_t}^\gamma(\sigma_i^2(\mathbf{A})) &\leq \sum_{\sigma_i^2(\mathbf{A}) \in \bar{R}_t \setminus R_t} \mathbb{I}[i \in I] \\ &\leq \sum_{\sigma_i^2(\mathbf{A}) \in M_t \setminus R_t} \mathbb{I}[i \in I] \\ &\leq \sum_{\sigma_i^2(\mathbf{A}) \in M_t} \mathbb{I}[i \in I]. \end{aligned} \quad (7)$$

The first bound follows from the fact that for $\sigma_i(\mathbf{A})^2 \in \bar{R}_t \setminus R_t$, we have $\sigma_i(\mathbf{A})^2 \in I$ and that $h_{R_t}^\gamma(\sigma_i^2(\mathbf{A})) \leq 1$. The second bound follows from (6), which shows that $\bar{R}_t \subset M_t$. Let $m_t = |\{i : \sigma_i^2(\mathbf{A}) \in M_t\}|$ be the number of squared singular values falling in M_t . Note that like M_t , m_t is fixed (i.e., not a random variable.) Thus, by linearity of expectation, we have:

$$\mathbb{E} \left[\sum_{\sigma_i^2(\mathbf{A}) \in M_t} \mathbb{I}[i \in I] \right] = \sum_{\sigma_i^2(\mathbf{A}) \in M_t} \mathbb{P}[i \in I] = \frac{8\gamma}{\alpha} \cdot m_t.$$

Letting $T = \lceil \log_{(1-\alpha)} \lambda \rceil$ as in Algorithm 1, by a Markov bound, with probability $1 - \frac{1}{200(T+1)}$:

$$\sum_{\sigma_i^2(\mathbf{A}) \in M_t} \mathbb{I}[i \in I] \leq \frac{1600(T+1)\gamma}{\alpha} \cdot m_t \leq T\epsilon_2 \cdot m_t$$

if c_1 is set small enough. By a union bound this holds for all $t \in \{0, \dots, T\}$ simultaneously with probability $\geq 199/200$. Plugging back into (7), we have, with probability $\geq 199/200$, simultaneously for all t :

$$\begin{aligned} \sum_{\sigma_i^2(\mathbf{A}) \in \bar{R}_t \setminus R_t} h_{R_t}^\gamma(\sigma_i^2(\mathbf{A})) &\leq T\epsilon_2 \cdot m_t \\ &\leq T\epsilon_2 \cdot (b_{t-1} + b_t + b_{t+1}) \end{aligned} \quad (8)$$

where the second bound follows from the fact that, regardless of the setting of the shift a_1 , $M_t \subset R_{t-1} \cup R_t \cup R_{t+1}$ so $m_t \leq (b_{t-1} + b_t + b_{t+1})$. Plugging (8) into (5) we have:

$$b_t \leq \text{tr}(h_{R_t}^\gamma(\mathbf{A}^T \mathbf{A})) \leq b_t + T\epsilon_2(b_{t-1} + b_t + b_{t+1}). \quad (9)$$

We conclude by showing that, before the final rounding step of Algorithm 1, $\tilde{b}_t \in (1 \pm \epsilon_1) \text{tr}(h_{R_t}^\gamma(\mathbf{A}^T \mathbf{A})) + \sqrt{c_3}\epsilon_1$ with high probability. In the final rounding step, if $\text{tr}(h_{R_t}^\gamma(\mathbf{A}^T \mathbf{A})) \leq 1/4$ (which can only occur if $b_t = 0$) and c_1, c_3 are sufficiently small, then we will have $\tilde{b}_t \leq 1/2$ and so will round down to $\tilde{b}_t = 0 = b_t$. Otherwise, the $\sqrt{c_3}\epsilon_1$ term will be absorbed into the relative error on $\text{tr}(h_{R_t}^\gamma(\mathbf{A}^T \mathbf{A}))$.

Overall, combined with (9) we will have, with probability $\geq 99/100$ for all t :

$$(1 - \epsilon)b_t \leq \tilde{b}_t \leq (1 + 2\epsilon_1)b_t + (1 + 2\epsilon_1)T\epsilon_2(b_{t-1} + b_t + b_{t+1})$$

which gives the theorem if we adjust ϵ_1, ϵ_2 by making c_1, c_2, c_3 sufficiently small. Thus we conclude by showing that in fact $\tilde{b}_t \in (1 \pm \epsilon_1) \text{tr}(h_{R_t}^\gamma(\mathbf{A}^T \mathbf{A})) + \sqrt{c_3}\epsilon_1$ with high probability. Setting $S = \frac{\log n}{c_2 \epsilon_1^2}$ as in

Algorithm 1, for $\mathbf{y}_1, \dots, \mathbf{y}_S$ chosen from $\{-1, 1\}^d$, with high probability $\frac{1}{S} \sum_{i=1}^S \mathbf{y}_i^T h_{R_t}^\gamma(\mathbf{A}^T \mathbf{A}) \mathbf{y}_i \in (1 \pm \epsilon_1) \text{tr}(h_{R_t}^\gamma(\mathbf{A}^T \mathbf{A}))$ by a standard trace estimation result [AT11].

Further, let $\mathbf{x}_i = \mathcal{W}(\mathbf{A}^T \mathbf{A}, \mathbf{y}, a_{t+1}, a_t, \gamma, c_3 \epsilon_1^2/n) \mathbf{y}_i$. By Theorem 7: $\|\mathbf{x}_i - h_{R_t}^\gamma(\mathbf{A}^T \mathbf{A}) \mathbf{y}_i\|_2 \leq \frac{c_3 \epsilon_1^2}{n^2} \|\mathbf{y}_i\|_2$ which by Cauchy-Schwarz gives $|\mathbf{y}_i^T \mathbf{x}_i - \mathbf{y}_i^T h_{R_t}^\gamma(\mathbf{A}^T \mathbf{A}) \mathbf{y}_i| \leq \sqrt{c_3/n} \epsilon_1 \|\mathbf{y}_i\|_2 = \sqrt{c_3} \epsilon_1$ since $\|\mathbf{y}_i\|_2 = \sqrt{n}$. Thus, overall we have, before the rounding step in which $\tilde{b}_t$ is set to 0 if $\tilde{b}_t < 1/2$:

$$\begin{aligned} \tilde{b}_t &= \frac{1}{S} \sum_{i=1}^S \mathbf{y}_i^T \mathcal{W}(\mathbf{A}^T \mathbf{A}, \mathbf{y}, a_{t+1}, a_t, \gamma, c_3 \epsilon_1^2/n) \mathbf{y}_i \\ &\in (1 \pm \epsilon_1) \text{tr}(h_{R_t}^\gamma(\mathbf{A}^T \mathbf{A})) + \sqrt{c_3} \epsilon_1. \end{aligned}$$

□

4.2 Application to General Spectral Sums

While Theorem 8 is useful in its own right, we now apply it to approximate a broad class of spectral sums. We need two assumptions on the sums that we approximate. First, for the histogram discretization to be relatively accurate, we need our function to be relatively smooth. Second, it is expensive to compute the histogram over very small singular values of $\mathbf{A}$ (i.e. with λ very small in Algorithm 1) as this makes the condition number in Theorem 7 large. So it is important that small singular values cannot contribute significantly to our sum. We start with the following definition:

Definition 9 (Multiplicative Smoothness). $f : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ is δ_f -multiplicatively smooth if for some $\delta_f \geq 1$, for all x , $|f'(x)| \leq \delta_f \frac{f(x)}{x}$.

We have the following claim, proven in Appendix D.

Claim 10. Let $f : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ be a δ_f -multiplicatively smooth function. For all $x, y \in \mathbb{R}^+$ and $c \in (0, \frac{1}{3\delta_f})$

$$y \in [(1-c)x, (1+c)x] \Rightarrow f(y) \in [(1-3\delta_f c)f(x), (1+3\delta_f c)f(x)].$$

For the example of the Schatten p -norm, for $f(x) = x^p$, $f'(x) = p \cdot x^{p-1}$ and so $f(x)$ is p -multiplicatively smooth.

We now give our general approximation theorem, showing that any spectral sum depending on sufficiently smooth and rapidly growing f can be computed using Algorithm 1:

Theorem 11 (Spectral Sums Via Approximate Histogram). Consider any $\mathbf{A} \in \mathbb{R}^{n \times d}$ and any function $f : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ satisfying:

- *Multiplicative Smoothness:* For some $\delta_f \geq 1$, f is δ_f -multiplicatively smooth (Defn. 9).
- *Small Tail:* For any $\epsilon > 0$ there exists $\lambda_f(\epsilon)$ such that for $x \in [0, \lambda_f(\epsilon)]$, $f(x) \leq \frac{\epsilon}{n} \mathcal{S}_f(\mathbf{A})$

Given error parameter $\epsilon \in (0, 1)$ and spectral norm estimate $M \in [\|\mathbf{A}\|_2, 2\|\mathbf{A}\|_2]$, for sufficiently small constant c , if we run Algorithm 1 on $\frac{1}{M} \mathbf{A}$ with input parameters $\epsilon_1, \epsilon_2 = c\epsilon$, $\alpha = c\epsilon/\delta_f$ and $\lambda = \lambda_f(c\epsilon)^2/M^2$ then with probability 99/100, letting $a_1, \tilde{b}_0, \dots, \tilde{b}_T$ be the outputs of the algorithm and $g(x) = f(x^{1/2})$:

$$(1 - \epsilon) \mathcal{S}_f(\mathbf{A}) \leq \sum_{t=0}^T g(M^2 \cdot a_1 (1 - \alpha)^t) \cdot \tilde{b}_t \leq (1 + \epsilon) \mathcal{S}_f(\mathbf{A}).$$

For parameter $k \in [d]$, letting $\bar{\kappa} \stackrel{\text{def}}{=} \frac{k\sigma_k^2(\mathbf{A}) + \sum_{i=k+1}^d \sigma_i^2(\mathbf{A})}{d \cdot \lambda}$ and $\hat{\kappa} \stackrel{\text{def}}{=} \frac{\sigma_{k+1}^2(\mathbf{A})}{\lambda}$, the algorithm runs in

$$\tilde{O} \left(\text{nnz}(\mathbf{A})k + dk^{\omega-1} + \frac{\text{nnz}(\mathbf{A}) + \sqrt{\text{nnz}(\mathbf{A})[d \cdot d_s(\mathbf{A}) + dk]\bar{\kappa}}}{\epsilon^5/(\delta_f^2 \log(1/\lambda))} \right)$$

$$\text{or } \tilde{O} \left(\text{nnz}(\mathbf{A})k + dk^{\omega-1} + \frac{(\text{nnz}(\mathbf{A}) + dk) \lceil \sqrt{\hat{\kappa}} \rceil}{\epsilon^5/(\delta_f^2 \log(1/\lambda))} \right)$$

time for sparse $\mathbf{A}$ or $\tilde{O} \left(nd^{\omega(\log_d k)-1} + \frac{nd + n^{1/2}d^{3/2}\sqrt{\bar{\kappa}}}{\epsilon^5/(\delta_f^2 \log(1/\lambda))} \right)$ for dense $\mathbf{A}$.

That is, we accurately approximate $\mathcal{S}_f(\mathbf{A})$ by discretizing over the histogram output by Algorithm 1. Note that we can boost our probability of success to $1 - \delta$ by repeating the algorithm $\Theta(\log(1/\delta))$ times and taking the median of the outputs.

Proof. Let $\bar{\mathbf{A}} \stackrel{\text{def}}{=} \frac{1}{M} \mathbf{A}$. Note that $\|\bar{\mathbf{A}}\|_2 \leq 1$ and so it is a valid matrix on which to apply Algorithm 1 and Theorem 8. Recall that we use the notation: $R_0 = [a_1, 1]$, $R_t = [a_1(1 - \alpha)^t, a_1(1 - \alpha)^{t-1}]$ for $t \geq 1$, $T = \lceil \log_{(1-\alpha)} \lambda \rceil$, and $b_t = |\{i : \sigma_i^2(\bar{\mathbf{A}}) \in R_t\}|$.

Since $g(x) = f(x^{1/2})$, by chain rule:

$$g'(x) = \frac{f'(x^{1/2})}{2x^{1/2}} \leq \delta_f \frac{f(x^{1/2})}{2x} = \frac{\delta_f}{2} \frac{g(x)}{x}.$$

So g is $\delta_f/2$ multiplicatively smooth. By this smoothness, and Claim 10, for any i with $\sigma_i^2(\bar{\mathbf{A}}) \in R_t$:

$$g(M^2 \cdot a_1(1 - \alpha)^t) \in \left(1 \pm \frac{3\delta_f \alpha}{2}\right) g(\sigma_i^2(\mathbf{A}))$$

$$\in (1 \pm \epsilon/4) \cdot f(\sigma_i(\mathbf{A}))$$

if the constant c on α is set small enough. Small enough c also ensures $g(M^2 a_1(1 - \alpha)^{t-1}) \in (1 \pm \epsilon/4)f(\sigma_i(\mathbf{A}))$ and $g(M^2 a_1(1 - \alpha)^{t+1}) \in (1 \pm \epsilon/4)f(\sigma_i(\mathbf{A}))$. So, both the multiplicative and additive error terms in Theorem 8 do not hurt us significantly. For now, assume that $\epsilon_2 = c\epsilon/T$. We will first prove the result with this assumption and then show that it can be relaxed to $\epsilon_2 = c\epsilon$ as given in the theorem statement. Let I be the set of indices with $\sigma_i^2(\bar{\mathbf{A}}) \in R_t$ for some $t \in \{0, \dots, T\}$. That is, the set of singular values covered by our histogram. Let I_{T+1} be the set of indices with $\sigma_i^2(\bar{\mathbf{A}}) \in R_{T+1}$ – that is, singular values which are not included in the histogram, but may be included in the additive error for the last bucket covering R_T . Applying Theorem 8:

$$\begin{aligned} \sum_{t=0}^T g(M^2 a_1(1 - \alpha)^t) \cdot \tilde{b}_t &\leq \sum_{t=0}^T g(M^2 a_1(1 - \alpha)^t) \cdot [(1 + \epsilon_1)b_t + \epsilon_2 T(b_{t-1} + b_t + b_{t+1})] \\ &\leq (1 + \epsilon/4) \sum_{t=0}^{T+1} \left[(1 + c\epsilon) \sum_{\sigma_i^2(\bar{\mathbf{A}}) \in R_t} f(\sigma_i(\mathbf{A})) + 3c\epsilon \sum_{\sigma_i^2(\bar{\mathbf{A}}) \in R_t} f(\sigma_i(\mathbf{A})) \right] \\ &\leq (1 + \epsilon) \sum_{i \in I \cup I_{T+1}} f(\sigma_i(\mathbf{A})) \\ &\leq (1 + \epsilon) \mathcal{S}_f(\mathbf{A}) \end{aligned} \tag{10}$$

if we set c small enough. The second inequality arises because each $\sigma_i(\mathbf{A})$ contributing to b_t appears at most three times as an additive error term for $\tilde{b}_{t-1}$, $\tilde{b}_t$, and $\tilde{b}_{t+1}$. In this inequality we include bucket R_{T+1} in the histogram, which only increases the right hand side.

On the lower bound side, we use our small tail assumption, that for i with $\sigma_i(\mathbf{A}) < \lambda_f(c\epsilon)$, $f(\sigma_i(\mathbf{A})) \leq \frac{c\epsilon}{n} \mathcal{S}_f(\mathbf{A})$. Using the notation of Algorithm 1, we have $a_{T+1} = a_1(1 - \alpha)^T \leq (1 - \alpha)^{\lceil \log_{(1-\alpha)} \lambda \rceil} \leq \lambda \stackrel{\text{def}}{=} \lambda_f(c\epsilon)^2/M^2$. So for any i with $\sigma_i(\mathbf{A}) \geq \lambda_f(c\epsilon)$, $\sigma_i^2(\bar{\mathbf{A}}) > a_{T+1}$ and thus falls in some bucket of our histogram so $i \in I$. We thus have:

$$\begin{aligned} \sum_{i \in I} f(\sigma_i(\mathbf{A})) &\geq \mathcal{S}_f(\mathbf{A}) - \sum_{i: \sigma_i(\mathbf{A}) < \lambda_f(c\epsilon)} f(\sigma_i(\mathbf{A})) \\ &\geq \mathcal{S}_f(\mathbf{A}) - n \cdot \frac{c\epsilon}{n} \mathcal{S}_f(\mathbf{A}) = (1 - c\epsilon) \mathcal{S}_f(\mathbf{A}). \end{aligned}$$

Applying Theorem 8 again, we have for sufficiently small c :

$$\begin{aligned} \sum_{t=0}^T g(a_1(1 - \alpha)^t) \cdot \tilde{b}_t &\geq (1 - \epsilon/4)(1 - c\epsilon) \sum_{t=0}^T \sum_{\sigma_i^2(\bar{\mathbf{A}}) \in R_t} f(\sigma_i(\mathbf{A})) \\ &= (1 - \epsilon/4)(1 - c\epsilon) \sum_{i \in I} f(\sigma_i(\mathbf{A})) \\ &\geq (1 - \epsilon) \mathcal{S}_f(\mathbf{A}). \end{aligned}$$

We conclude the theorem by noting that we can actually set $\epsilon_2 = c\epsilon$ instead of $\epsilon_2 = c\epsilon/T$ as used above. The additive error term on each bucket was bounded using a Markov bound, and to union bound over T buckets, we lost a factor of T . However, in expectation, the total contribution of additive error to our spectral sum estimation is just $O\left(\epsilon \sum_{i \in I \cup I_{T+1}} f(\sigma_i(\mathbf{A}))\right) = O(\mathcal{S}_f(\mathbf{A}))$ and so by a Markov bound is $\leq \epsilon/2 \mathcal{S}_f(\mathbf{A})$ with probability 99/100 if we set our constants small enough.

It just remains to discuss runtime, i.e. to calculate the runtime of Algorithm 1 with inputs $\epsilon_1, \epsilon_2 = c\epsilon$, $\alpha = c\epsilon/\delta_f$ and $\lambda = \lambda_f(c\epsilon)^2/M^2$. The number of outer loops of the algorithm is $T = \lceil \log_{(1-\alpha)} \lambda \rceil = \Theta\left(\frac{\delta_f}{\epsilon} \log(1/\lambda)\right)$. The number of inner loops is $S = \tilde{O}(1/\epsilon^2)$. And finally, within each loop the most expensive operation is computing $\mathcal{W}(\mathbf{A}^T \mathbf{A}, \mathbf{y}, a_{t+1}, a_t, \gamma, c_3 \epsilon_1^2/n)$. Our final runtime follows from plugging this into Theorem 7 noting that $\gamma = \Theta(\epsilon\alpha) = \Theta(\epsilon^2/\delta_f)$ and $a_{t+1} = \Omega(\lambda)$ for all $t \in \{0, \dots, T\}$. Note that we perform the precomputation step to construct a preconditioner for $\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I}$ just once, incurring cost $\tilde{O}(\text{nnz}(\mathbf{A}) + dk^{\omega-1})$ or $\tilde{O}(nd^{\omega(\log_d k)-1})$. $\square$

4.3 Application to Schatten p , Orlicz, and Ky Fan Norms

Theorem 11 is very general, allowing us to approximate any function satisfying a simple smoothness condition as long as the smaller singular values of $\mathbf{A}$ cannot contribute significantly to $\mathcal{S}_f(\mathbf{A})$. We now give some specific applications of the result.

Schatten p -Norms

Theorem 11 already gives the fastest known algorithms for Schatten p -norm estimation. We will not go into all runtime tradeoffs now as our best runtimes will be worked out in detail in Sections 6 and 7, but as an example:

Corollary 12 (Schatten p -norms via Histogram Approximation). *For any $\mathbf{A} \in \mathbb{R}^{n \times n}$ with uniformly sparse rows (i.e. $d_s(\mathbf{A}) = O(\text{nnz}(\mathbf{A})/n)$), given error parameter $\epsilon \in (0, 1)$ and $M \in$*

$[\|\mathbf{A}\|_2, 2\|\mathbf{A}\|_2]$, if we run Algorithm 1 on $\frac{1}{M}\mathbf{A}$ with $\epsilon_1, \epsilon_2 = c\epsilon$, $\alpha = c\epsilon/\max\{1, p\}$ and $\lambda = \frac{1}{M^2} \left(\frac{c\epsilon}{n} \|\mathbf{A}\|_p^p\right)^{2/p}$ for sufficiently small constant c then with probability 99/100, letting $a_1, \tilde{b}_0, \dots, \tilde{b}_T$ be the outputs of the algorithm we have:

$$(1 - \epsilon) \|\mathbf{A}\|_p^p \leq \sum_{t=0}^T [M^2 a_1 (1 - \alpha)^t]^{p/2} \cdot \tilde{b}_t \leq (1 + \epsilon) \|\mathbf{A}\|_p^p.$$

Further the algorithm runs in time:

$$\tilde{O}\left(\frac{\text{nnz}(\mathbf{A})p^2}{\epsilon^{5+1/p}}\right) \text{ for } p \geq 2 \quad \text{and} \quad \tilde{O}\left(\frac{\text{nnz}(\mathbf{A})n^{\frac{1/p-1/2}{1/p+1/2}} + n^{\frac{5/p-1/2}{2/p+1}} \sqrt{d_s(\mathbf{A})}}{p \cdot \epsilon^{5+1/p}}\right) \text{ for } p \leq 2.$$

For dense inputs this can be sped up to $\tilde{O}\left(\frac{n^{\frac{2.3729-1171p}{1+0346p}}}{p \cdot \epsilon^{5+1/p}}\right)$ using fast matrix multiplication or

$$\tilde{O}\left(\frac{n^{\frac{3+p/2}{1+p/2}}}{p \cdot \epsilon^{5+1/p}}\right) \text{ without fast matrix multiplication.}$$

For constant $\epsilon, p > 2$ the first runtime is $\tilde{O}(\text{nnz}(\mathbf{A}))$, and for the nuclear norm ($p = 1$), for constant ϵ the second runtime gives $\tilde{O}(\text{nnz}(\mathbf{A})n^{1/3} + n^{3/2}\sqrt{d_s(\mathbf{A})})$ which is at worst $\tilde{O}(\text{nnz}(\mathbf{A})n^{1/3} + n^2)$. For dense matrices, the nuclear norm estimation time is $\tilde{O}(n^{2.18})$ using fast matrix multiplication. It is already $\tilde{O}(n^{2.33}) = o(n^\omega)$, for the current ω , without using fast matrix multiplication.

Note that we can compute the spectral norm approximation used to scale $\mathbf{A}$ via the Lanczos or power method in $\tilde{O}(\text{nnz}(\mathbf{A}))$ time. λ depends on $\|\mathbf{A}\|_p$ which we are estimating. However, as we will discuss in the proof, we can use a rough estimate for $\|\mathbf{A}\|_p$ which suffices. As a more general strategy, λ can be identified via binary search. We can start with $\lambda = \sigma_1(\mathbf{A})^2/M^2$ and successively decrease λ running Algorithm 1 up to the stated runtime bounds. If it does not finish in the allotted time, we know that we have set λ too small. Thus, we can output the result with the smallest λ such that the algorithm completes within in the stated bounds.

Proof. We invoke Theorem 11 with $f(x) = x^p$. We have $f'(x) = p\frac{f(x)}{x}$ so $\delta_f = \max\{1, p\}$ and our setting of $\alpha = c\epsilon/\max\{1, p\}$ suffices. Additionally, for any c , we can set $\lambda_f(c\epsilon) = \left(\frac{c\epsilon}{n} \|\mathbf{A}\|_p^p\right)^{1/p} = \frac{c^{1/p}\epsilon^{1/p}}{n^{1/p}} \|\mathbf{A}\|_p$ and so our setting of λ suffices. Thus the accuracy bound follows from Theorem 11.

We now consider runtime. For $p \geq 2$:

$$\bar{\kappa} = \frac{k\sigma_k^2(\mathbf{A}) + \sum_{i=k+1}^n \sigma_i^2(\mathbf{A})}{n\lambda} \leq \frac{n^{2/p-1}}{\epsilon^{2/p}} \cdot \frac{\|\mathbf{A}\|_F^2}{\|\mathbf{A}\|_p^2}.$$

We can bound $\|\mathbf{A}\|_F \leq n^{1/2-1/p} \|\mathbf{A}\|_p$ and so have $\bar{\kappa} \leq \frac{1}{\epsilon^{2/p}}$. For $p < 2$ we have:

$$\bar{\kappa} = \frac{n^{2/p-1}}{\epsilon^{2/p}} \cdot \frac{k\sigma_k^2(\mathbf{A}) + \sum_{i=k+1}^n \sigma_i^2(\mathbf{A})}{\|\mathbf{A}\|_p^2} \leq \frac{n^{2/p-1}}{\epsilon^{2/p}} \cdot \frac{\sigma_k^{2-p}(\mathbf{A}) \sum_{i=1}^n \sigma_i^p(\mathbf{A})}{\|\mathbf{A}\|_p^2} = \frac{n^{2/p-1}}{\epsilon^{2/p}} \cdot \frac{\sigma_k^{2-p}(\mathbf{A})}{\|\mathbf{A}\|_p^{2-p}}.$$

Using the fact that $\sigma_k^p(\mathbf{A}) \leq \frac{1}{k} \|\mathbf{A}\|_p^p$ we have the tradeoff between k and $\bar{\kappa}$:

$$\bar{\kappa} \leq \frac{1}{\epsilon^{2/p}} \left(\frac{n}{k}\right)^{2/p-1}. \quad (11)$$

As mentioned, λ depends on the value of $\|\mathbf{A}\|_p$. We can simply lower bound $\|\mathbf{A}\|_p^p$ by $k\sigma_k^p(\mathbf{A})$, which we estimate up to multiplicative error when performing deflation. We can use this lower bound to set λ . Our estimated λ will only be smaller than the true value, giving a better approximation guarantee and the above condition number bound will still hold.

We now analyze our runtime. Recall that for $f(x) = x^p$, $\delta_f = \max\{1, p\}$. Correspondingly, $\log(1/\lambda) = \tilde{O}(\max\{1, 1/p\})$ and so $\delta_f^2 \log(1/\lambda) = \max\{p^2, 1/p\}$. Plugging into the first runtime of Theorem 11, using the uniform sparsity assumption and the fact that $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ we have:

$$\tilde{O} \left(\text{nnz}(\mathbf{A})k + nk^{\omega-1} + \frac{\text{nnz}(\mathbf{A})\sqrt{\bar{\kappa}} + \sqrt{\text{nnz}(\mathbf{A})nk\bar{\kappa}}}{\epsilon^5/(\max\{p^2, 1/p\})} \right).$$

For $p \geq 2$ we just set $k = 0$ and have $\tilde{O}(\text{nnz}(\mathbf{A})p^2/\epsilon^{5+1/p})$ runtime by our bound $\bar{\kappa} \leq \frac{1}{\epsilon^{2/p}}$. For $p \leq 2$, not trying to optimize $\text{poly}(1/\epsilon)$ terms, we write the runtime as

$$\tilde{O} \left(nd_s(\mathbf{A})k + nk^{\omega-1} + \frac{nd_s(\mathbf{A})\sqrt{\bar{\kappa}} + n\sqrt{d_s(\mathbf{A})k\bar{\kappa}}}{\epsilon^5 p} \right)$$

and balancing the first two coefficients on n , set $k = n^{\frac{1/p-1/2}{1/p+1/2}}$ which gives $\sqrt{\bar{\kappa}} = n^{\frac{1/p-1/2}{1/p+1/2}}$ by (11) and so $nd_s(\mathbf{A})k = nd_s(\mathbf{A})\sqrt{\bar{\kappa}}$. We then have $n\sqrt{d_s(\mathbf{A})k\bar{\kappa}} = n\sqrt{d_s(\mathbf{A})}k^{3/2} = n^{\frac{5/p-1/2}{2/p+1}}\sqrt{d_s(\mathbf{A})}$. Finally, the $nk^{\omega-1}$ is dominated by the $n\sqrt{d_s(\mathbf{A})}k^{3/2}$ term so we drop it.

Finally, for dense $\mathbf{A}$ we apply the third runtime which gives

$$\begin{aligned} \tilde{O} \left(n^{\omega(\log_n k)} + \frac{n^2\sqrt{\bar{\kappa}}}{\epsilon^5 p} \right) &= \tilde{O} \left(n^{\omega(\log_n k)} + \frac{n^{3/2+1/p}}{k^{1/p-1/2} \cdot \epsilon^{5+1/p} \cdot p} \right) \\ &= \tilde{O} \left(n^{\omega(\log_n k)} + \frac{n^{3/2+1/p-(\log_n k)(1/p-1/2)}}{\epsilon^{5+1/p} \cdot p} \right). \end{aligned}$$

We now balance the terms, again not optimizing ϵ dependence. Writing $\gamma = \log_n k$, $\omega(\gamma) = 2$ for $\gamma < \alpha$ where $\alpha > 0.31389$ and $2 + (\omega - 2)\frac{\gamma-\alpha}{1-\alpha}$ for $\gamma \geq \alpha$ [GU17]. Assuming $\gamma > \alpha$ we set:

$$2 + (\omega - 2)\frac{\gamma - \alpha}{1 - \alpha} = \frac{3}{2} + \frac{1}{p} - \frac{\gamma}{p} + \frac{\gamma}{2}$$

which gives $\gamma \approx \frac{1/p-.3294}{1/p+.0435} > \alpha$ for all $p < 2$ (so our assumption that $\gamma \geq \alpha$ was valid.) This yields total runtime $\tilde{O} \left(n^{\frac{2.3729-.0994p}{1+.0435p}} \right)$. Without using fast matrix multiplication, the first term in the runtime becomes n^2k and so we balance costs by setting: $n^{2+\gamma} = n^{3/2+1/p-\gamma/p+\gamma/2}$ which gives $\gamma = \frac{1/p-1/2}{1/p+1/2}$ and total runtime $\tilde{O} \left(n^{\frac{3+p/2}{1+p/2}} \right)$. □

Bounded Orlicz Norms

An Orlicz norm of a vector $\mathbf{x} \in \mathbb{R}^n$ is given by $G(\mathbf{x}) := \sum_{i=1}^n g(|\mathbf{x}_i|)$, where $g(\cdot)$ is convex, nonnegative, and has $g(0) = 0$. It is easy to see that applied to the vector of singular values, an Orlicz matrix norm is a special case of a spectral sum and can be approximated with Theorem 11 under sufficient conditions. A simple example shows that Orlicz norms for $g(\cdot)$ bounded by an envelope of x^{p_1} and x^{p_2} can be approximated in a similar time as $\|\mathbf{A}\|_{p_2}^{p_2}$ in Corollary 12.

Corollary 13 (Bounded Orlicz norms via Histogram Approximation). *For any convex function $g : \mathbb{R}^+ \rightarrow \mathbb{R}^+$, let p_1 be the minimal real for which for all arguments a, b to g with $a \leq b$, $(\frac{b}{a})^{p_1} \geq \frac{g(b)}{g(a)}$. Let p_2 be the maximum positive real for which for all $a \leq b$ we have $(\frac{b}{a})^{p_2} \leq \frac{g(b)}{g(a)}$.*

Then for any $\mathbf{A} \in \mathbb{R}^{n \times n}$ with uniformly sparse rows (i.e. $d_s(\mathbf{A}) = O(\text{nnz}(\mathbf{A})/n)$), given error parameter $\epsilon \in (0, 1)$ and $M \in [\|\mathbf{A}\|_2, 2\|\mathbf{A}\|_2]$, if we run Algorithm 1 on $\frac{1}{M}\mathbf{A}$ with $\epsilon_1, \epsilon_2 = c\epsilon$, $\alpha = c\epsilon/\max\{1, p_1\}$ and $\lambda = \frac{1}{M^2} \left(\frac{c\epsilon/2}{n} \|\mathbf{A}\|_{p_2}^{p_2} \right)^{2/p_2}$ for sufficiently small constant c then with probability $99/100$, letting $a_1, \tilde{b}_0, \dots, \tilde{b}_T$ be the outputs of the algorithm we have:

$$(1 - \epsilon)\mathcal{S}_g(\mathbf{A}) \leq \sum_{t=0}^T g\left(Ma_1^{1/2}(1 - \alpha)^{t/2}\right) \cdot \tilde{b}_t \leq (1 + \epsilon)\mathcal{S}_g(\mathbf{A}).$$

Further the algorithm runs in time:

$$\tilde{O}\left(\frac{\text{nnz}(\mathbf{A})(p_1 + 1)^2}{\epsilon^{5+1/p_2}}\right) \text{ for } p_2 \geq 2 \quad \text{and} \quad \tilde{O}\left(\frac{(p_1 + 1)^2 \cdot \text{nnz}(\mathbf{A})n^{\frac{1/p_2 - 1/2}{1/p_2 + 1/2}} + n^{\frac{5/p_2 - 1/2}{2/p_2 + 1}} \sqrt{d_s(\mathbf{A})}}{p_2 \cdot \epsilon^{5+1/p_2}}\right) \text{ for } p_2 \leq 2.$$

For dense inputs this can be sped up to $\tilde{O}\left((p_1 + 1)^2 \cdot \frac{n^{\frac{2.3729 - .0994p_2}{1 + .0435p_2}}}{p_2 \cdot \epsilon^{5+1/p_2}}\right)$ via fast matrix multiplication.

Note that λ depends on $\|\mathbf{A}\|_{p_2}^{p_2}$ which can be estimated for example using Corollary 12, or via binary search as described in the proof of Corollary 12.

Proof. We invoke Theorem 11 with $f(x) = g(x)$. While we do not show that $g(\cdot)$ exactly satisfies the multiplicative smoothness condition, the upper bound on $g(\cdot)$ directly implies the result of Claim 10 with $\delta_f = p_1$. So our setting of $\alpha = c\epsilon/\max\{1, p_1\}$ suffices as the bound in this Claim is the only smoothness condition used to prove Theorem 11.

Additionally, by the lower bound on $g(\cdot)$, for any c , we can set $\lambda_f(c\epsilon) = \left(\frac{c\epsilon/2}{n} \|\mathbf{A}\|_{p_2}^{p_2}\right)^{1/p_2}$. For any $\sigma_i(\mathbf{A}) \geq \lambda_f(c\epsilon)$:

$$\begin{aligned} g(\sigma_i(\mathbf{A})) &\geq \left(\frac{\sigma_i(\mathbf{A})}{\lambda_f(c\epsilon)}\right)^{p_2} g(\lambda_f(c\epsilon)) \\ &\geq \frac{2n}{c\epsilon} \cdot \frac{\sigma_i^{p_2}(\mathbf{A})}{\|\mathbf{A}\|_{p_2}^{p_2}} \cdot g(\lambda_f(c\epsilon)). \end{aligned}$$

which gives:

$$\begin{aligned} \sum_{\sigma_i(\mathbf{A}) \geq \lambda_f(c\epsilon)} g(\sigma_i(\mathbf{A})) &\geq \frac{2n \cdot g(\lambda_f(c\epsilon))}{c\epsilon \cdot \|\mathbf{A}\|_{p_2}^{p_2}} \cdot \sum_{\sigma_i(\mathbf{A}) \geq \lambda_f(c\epsilon)} \sigma_i^{p_2}(\mathbf{A}) \\ &\geq \frac{2n \cdot g(\lambda_f(c\epsilon))}{c\epsilon} \cdot (1 - c\epsilon/2) \end{aligned}$$

where the last bound follows from the fact that by the setting of $\lambda_f(c\epsilon)$, $\frac{\sum_{\sigma_i(\mathbf{A}) \geq \lambda_f(c\epsilon)} \sigma_i^{p_2}(\mathbf{A})}{\|\mathbf{A}\|_{p_2}^{p_2}} \geq 1 - c\epsilon/2$. Finally for any $\sigma_i(\mathbf{A}) \leq \lambda_f(c\epsilon)$, $g(\sigma_i(\mathbf{A})) \leq g(\lambda_f(c\epsilon))$. So overall we have:

$$\sum_{\sigma_i(\mathbf{A}) < \lambda_f(c\epsilon)} g(\sigma_i(\mathbf{A})) \leq n \cdot g(\lambda_f(c\epsilon)) \leq \frac{c\epsilon/2}{1 - c\epsilon/2} \cdot \sum_{\sigma_i(\mathbf{A}) \geq \lambda_f(c\epsilon)} g(\sigma_i(\mathbf{A})) \leq \frac{c\epsilon/2}{1 - c\epsilon} \mathcal{S}_g(\mathbf{A}) \leq c\epsilon \mathcal{S}_g(\mathbf{A}).$$

Thus our setting of λ suffices and the accuracy bound follows from Theorem 11.

It remains to discuss runtime. We have $\log(1/\lambda) = \tilde{O}(\max\{1, 1/p_2\})$ and $\delta_f^2 = \max\{1, p_1^2\} \leq (p_1 + 1)^2$. The runtimes follow from Theorem 11 via the same arguments used in Corollary 12. $\square$

Ky Fan Norms

The Ky Fan w -norm of a matrix is the sum of its top w singular values: $\|\mathbf{A}\|_{KF(w)} \stackrel{\text{def}}{=} \sum_{i=1}^w \sigma_i(\mathbf{A})$ (note that these norms are typically called the ‘Ky Fan k -norms’, however we use w to avoid overloading notation on k). Such a norm is not strictly speaking a spectral sum. However it can still be approximated using our histogram method. We have the following corollary of Theorem 8:

Corollary 14 (Ky Fan norms via Histogram Approximation). *For any $\mathbf{A} \in \mathbb{R}^{n \times n}$, given rank w , error parameter $\epsilon \in (0, 1)$ and $M \in [\|\mathbf{A}\|_2, 2\|\mathbf{A}\|_2]$, if we run Algorithm 1 on $\frac{1}{M}\mathbf{A}$ with $\epsilon_1 = c\epsilon$, $\epsilon_2 = \frac{c_2\epsilon^2}{\log(1/\lambda)}$, $\alpha = c\epsilon$ and $\lambda = \left(\frac{c\epsilon}{Mw} \|\mathbf{A}\|_{KF(w)}\right)^2$ for sufficiently small constants c, c_2 then with probability 99/100, letting $a_1, \tilde{b}_0, \dots, \tilde{b}_T$ be the outputs of the algorithm we have:*

$$(1 - \epsilon) \|\mathbf{A}\|_{KF(w)} \leq \sum_{i=1}^w M a_1^{1/2} (1 - \alpha)^{\tilde{t}(i)/2} \leq (1 + \epsilon) \|\mathbf{A}\|_{KF(w)}$$

where $\tilde{t}(i)$ is the smallest integer with $\sum_{t=0}^{\tilde{t}(i)} (1 + 2c\epsilon) \tilde{b}_t \geq i$ or $\tilde{t}(i) = \infty$ if no such integer exists (and hence the i^{th} term of the sum is just 0). Further the algorithm runs in time:

$$\tilde{O}\left(\frac{\text{nnz}(\mathbf{A})\sqrt{w} + nw}{\epsilon^7}\right).$$

Note that to compute the top w singular values explicitly would require $\tilde{O}(\text{nnz}(\mathbf{A})w + nw^{\omega-1})$ time using for example a block Krylov method [MM15]. Also note that λ depends on the norm we are attempting to compute. As discussed in Corollary 12, we can approximate λ via binary search, successively decreasing it until our call to Algorithm 1 exceeds the stated runtime.

Proof. We can assume without loss of generality that $\|\mathbf{A}\|_2 \leq 1$ and $M = 1$ since rescaling will not effect our approximation factor. By Theorem 8, since we set $\epsilon_1 = c\epsilon$, $\epsilon_2 = \frac{c_2\epsilon^2}{\log(1/\lambda)}$, and have $T = \lceil \log_{1-\alpha} \lambda \rceil = \Theta\left(\frac{\log(1/\lambda)}{\epsilon}\right)$:

$$(1 - c\epsilon)b_t \leq \tilde{b}_t \leq b_t + 2c\epsilon(b_{t-1} + b_t + b_{t+1}) \quad (12)$$

if we set c_2 small enough compared to c . This give $b_t \leq \frac{1}{1-c\epsilon}\tilde{b}_t \leq (1 + 2c\epsilon)\tilde{b}_t$ if c is set small enough. Thus, since these scaled bucket sizes strictly overestimate the true bucket sizes we have $\tilde{t}(i) \leq t(i)$, where $t(i)$ is the smallest integer with $\sum_{t=0}^{t(i)} b_t \geq i$. This gives, since $\alpha = c\epsilon$ and since by our setting of λ , at most an $c\epsilon$ fraction of $\|\mathbf{A}\|_{KF(w)}$ falls outside of the ranges $R_0, \dots, R_T$:

$$\|\mathbf{A}\|_{KF(w)} = \sum_{i=1}^w \sigma_i(\mathbf{A}) \leq (1 + 4c\epsilon) \sum_{i=1}^w M a_1^{1/2} (1 - \alpha)^{t(i)/2} \leq (1 + 4c\epsilon) \sum_{i=1}^w M a_1^{1/2} (1 - \alpha)^{\tilde{t}(i)/2}.$$

On the other side, let $\tilde{\mathbf{v}}$ be a vector that, for each $t \in \{0, \dots, T\}$ has $(1 + 4c\epsilon)\tilde{b}_t$ entries each set to $M a_1^{1/2} (1 - \alpha)^{t/2}$. Define the top- w norm of $\tilde{\mathbf{v}}$ to be the sum of its largest w entries, denoted by

$\|\tilde{\mathbf{v}}\|_{T(w)}$. Our estimate of $\|\mathbf{A}\|_{KF(w)}$ is equal to $\|\tilde{\mathbf{v}}\|_{T(w)}$. Note that we can also add arbitrary zeros as padding entries to $\tilde{\mathbf{v}}$ and not change this norm.

Similarly, let $\mathbf{v}'$ be a vector with $(1 + 10c\epsilon)b_t$ values each set to $(1 + 2c\epsilon) \cdot a_1^{1/2}(1 - \alpha)^{t/2}$ for $t \in \{0, \dots, T\}$ and $2\epsilon c \cdot b_{T+1}$ entries set to $(1 + 2c\epsilon) \cdot a_1^{1/2}(1 - \alpha)^{T/2} \leq \lambda$.

By (12), $(1 + 2c\epsilon)\tilde{b}_t \leq (1 + 6c\epsilon)b_t + 2c\epsilon b_{t-1} + 2c\epsilon b_{t+1}$. This fact combined with the fact that the entries in $\mathbf{v}'$ are scaled up by a $(1 + 2c\epsilon)$ factor ensure that the entries of $\mathbf{v}'$ dominate those of $\tilde{\mathbf{v}}$ and so $\|\mathbf{v}'\|_{T(w)} \geq \|\tilde{\mathbf{v}}\|_{T(w)}$. Further, $\|\mathbf{v}'\|_{T(w)} \leq (1 + 13c\epsilon) \|\mathbf{A}\|_{KF(w)}$ since we have scaled up each entry by at most $(1 + 2c\epsilon)$ factor and scaled up each bucket size b_t by a $(1 + 10c\epsilon)$ factor. This bound gives our final multiplicative approximation after adjusting the constant c on ϵ .

It remains to discuss runtime. We invoke Theorem 7, setting $k = w^{1/2}$. This gives:

$$\hat{\kappa} = \frac{\sigma_{k+1}^2(\mathbf{A})}{\lambda} = \frac{\sigma_{k+1}^2(\mathbf{A})w^2}{\|\mathbf{A}\|_{KF(w)}^2 (\epsilon c)^2} \leq \frac{w}{(c\epsilon)^2}$$

where we use the fact that for $k = w^{1/2}$, $\sigma_k(\mathbf{A}) \leq \frac{1}{w^{1/2}} \|\mathbf{A}\|_{KF(w)}$. With our settings of λ , α and ϵ_1, ϵ_2 , Algorithm 1 performs $\tilde{\Theta}(1/\epsilon_1^2 \cdot 1/\alpha) = \tilde{\Theta}(1/\epsilon^3)$ calls to $\mathcal{W}$, each which requires $\tilde{\Theta}(1/\gamma) = \tilde{\Theta}(\epsilon_2\alpha) = \tilde{\Theta}(1/\epsilon^3)$ regression calls due to our setting of α and ϵ_2 . Plugging $\hat{\kappa}$ into the second runtime of Theorem 7 gives the corollary. $\square$

5 Lower Bounds

In this section we give hardness results for high accuracy spectrum approximation. Specifically, we show how to detect if an undirected unweighted graph contains a triangle using accurate approximation algorithms for various important spectral sums such as the Schatten p -norms, log-determinant, the SVD entropy, and the trace inverse. Our spectral sum bounds further imply hardness for the important primitives of computing effective resistances in a graph or leverage scores in a matrix.

In the seminal work of [WW10] it was shown that any truly subcubic time algorithm for triangle detection yields a truly subcubic time algorithm for Boolean matrix multiplication (BMM). Consequently, these results show that computing any of these quantities too precisely is in a sense as difficult as BMM. Furthermore, as it is a longstanding open question whether or not there is any subcubic time combinatorial algorithm for BMM, i.e. an algorithm which avoids the powerful algebraic manipulations inherent in the fastest algorithms for BMM, these results can be viewed as showing that we do not expect simple iterative methods to yield precise algorithms for these problems without an immense breakthrough in linear algebra.

In Section 5.1 we give a general result on reducing spectral sums to triangle detection. Then in Section 5.2 we use this to show hardness for computing various well studied spectral sums. In Section 5.3 we conclude by showing hardness for computing graph and numerical linear algebra primitives, i.e. effective resistances and leverage scores.

5.1 Reductions From Triangle Detection

Here we provide our main technical tool for reducing spectral sum computation to triangle detection. As discussed in Section 1.4, our reduction leverages the well known fact that the number of triangles in any unweighted graph G is equal to $\text{tr}(\mathbf{A}^3)/6$ where $\mathbf{A}$ is the adjacency matrix for G . Consequently, given any function $f : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ whose power series is reasonably behaved, we can show that for suitably small δ the quantity $\text{tr}(f(\mathbf{I} + \delta\mathbf{A}))$ is dominated by the contribution of $\text{tr}(\mathbf{A}^k)$ for $k \in (0, 3)$. Therefore computing $\text{tr}(f(\mathbf{I} + \delta\mathbf{A}))$ approximately lets us distinguish between whether or not $\text{tr}(\mathbf{A}^3) = 0$ or $\text{tr}(\mathbf{A}^3) \geq 6$.

We formalize this in the following theorem. As it simplifies the result, we focus on the case where f is defined on the interval $(0, 2)$, however, this suffices for our purposes and can be generalized via shifting and scaling of x .

Theorem 15 (Many Spectral Sums are as Hard as Triangle Detection). *Let $f : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ be an arbitrary function, such that for $x \in (0, 2)$ we can express it as*

$$f(x) = \sum_{k=0}^{\infty} c_k (x-1)^k \text{ where } \left| \frac{c_k}{c_3} \right| \leq h^{k-3} \text{ for all } k > 3 \quad (13)$$

Then given the adjacency matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ of any simple graph G that has no self-loops and spectral sum estimate $X \in (1 \pm \epsilon_1/9) \sum_{i=1}^n f(\sigma_i(\mathbf{I} - \delta \mathbf{A}))$ for scaling δ and accuracy ϵ satisfying

$$\delta = \min \left\{ \frac{1}{n}, \frac{1}{10n^4 h} \right\} \text{ and } \epsilon_1 = \min \left\{ 1, \left| \frac{c_3 \delta^3}{c_0 n} \right|, \left| \frac{c_3 \delta}{c_2 n^2} \right| \right\}$$

we can detect if G has a triangle in $O(\text{nnz}(\mathbf{A}))$ time.

Consequently, given an algorithm which on input $\mathbf{B} \in \mathbb{R}^{n \times n}$ outputs $Y \in (1 \pm \epsilon) \sum_{i=1}^n f(\sigma_i(\mathbf{B}))$ in $O(n^\gamma \epsilon^{-c})$ time we can produce an $O(n^2 + n^\gamma \epsilon_1^{-c})$ time triangle detection algorithm.

Proof. Let $\mathbf{A}$, G , δ , ϵ_1 , and X be as in the theorem statement and let $\mathbf{B} \stackrel{\text{def}}{=} \mathbf{I} - \delta \mathbf{A}$. By Gershgorin's circle theorem, $\|\mathbf{A}\|_2 \leq n-1$ and since $\delta \leq 1/n$, $\|\delta \mathbf{A}\|_2 < 1$. Consequently $\mathbf{B}$ is symmetric PSD, $\sigma_i(\mathbf{B}) = \lambda_i(\mathbf{B}) \in (0, 2)$ for all $i \in [n]$, and therefore applying (13) yields:

$$\sum_{i=1}^n f(\sigma_i(\mathbf{B})) = \sum_{i=1}^n f(1 - \delta \lambda_i(\mathbf{A})) = \sum_{i=1}^n \sum_{k=0}^{\infty} c_k (\delta \lambda_i(\mathbf{A}))^k = \sum_{k=0}^{\infty} c_k \delta^k \text{tr}(\mathbf{A}^k).$$

$\delta \leq \frac{1}{10n^4 h}$ is enough to insure that the first three terms of this power series dominate. Specifically:

$$\left| \sum_{k=4}^{\infty} c_k \delta^k \text{tr}(\mathbf{A}^k) \right| = \left| c_3 \delta^3 \sum_{k=4}^{\infty} \frac{c_k}{c_3} \delta^{k-3} \text{tr}(\mathbf{A}^k) \right| \leq |c_3| \delta^3 \sum_{k=4}^{\infty} \frac{1}{10^{k-3}} \left(\frac{1}{n^4} \right)^{k-3} \text{tr}(\mathbf{A}^k) \leq \frac{|c_3| \delta^3}{9}$$

where the last inequality uses the fact that $\text{tr}(\mathbf{A}^k) \leq \|\mathbf{A}\|_2^{k-2} \|\mathbf{A}\|_F^2 \leq n^k \leq n^{4(k-3)}$ for all $k > 3$. Further, since $\text{tr}(\mathbf{A}^0) = n$, $\text{tr}(\mathbf{A}) = 0$, and $\text{tr}(\mathbf{A}^2) = \|\mathbf{A}\|_F^2 \leq n^2$ we have:

$$0 \leq c_0 \text{tr}(\mathbf{A}^0) + c_1 \delta \text{tr}(\mathbf{A}) + c_2 \delta^2 (\text{tr}(\mathbf{A}^2)) \leq |c_3| \delta^3 \cdot \left(\left| \frac{c_0 n}{c_3 \delta^3} \right| + \left| \frac{c_2 n^2}{c_3 \delta} \right| \right) \leq \frac{|c_3| \delta^3}{\epsilon_1}.$$

Now, clearly in $O(\text{nnz}(\mathbf{A}))$ time we can compute $\text{tr}(\mathbf{A}^2) = \|\mathbf{A}\|_F^2$ in $O(\text{nnz}(\mathbf{A}))$ as well as:

$$\begin{aligned} X - c_0 n - c_2 \delta^2 \text{tr}(\mathbf{A}^2) &= c_3 \delta^3 \text{tr}(\mathbf{A}^3) \pm \frac{|c_3| \delta^3}{9} \pm \frac{\epsilon_1}{9} \left(\frac{|c_3| \delta^3}{9} + c_3 \delta^3 \text{tr}(\mathbf{A}^3) + \frac{|c_3| \delta^3}{\epsilon_1} \right) \\ &= c_3 \delta^3 \left[\text{tr}(\mathbf{A}^3) \left(1 \pm \frac{1}{20} \right) \pm \frac{1}{3} \right] \end{aligned}$$

So we can detect if $\text{tr}(\mathbf{A}^3) = 0$ or if $\text{tr}(\mathbf{A}^3) \geq 6$ and hence whether or not G has a triangle. $\square$

Note that in this reduction, so long as δ is small (i.e. $\leq 2n$) then $\mathbf{B} = \mathbf{I} - \delta \mathbf{A}$ is a very well conditioned matrix (its condition number is at most a constant). Consequently, our lower bounds apply even when approximately applying for example $\mathbf{B}^{-1}$ or $\mathbf{B}^{1/2}$ to high precision is inexpensive.

The theorem (and the results in Section 5.3) suggests that the difficulty in computing spectral sums arises more from the need to measure the contribution from multiple terms precisely, than from the difficulty in manipulating $\mathbf{B}$ for the purposes of applying it to a single vector.

Also, note that the matrix $\mathbf{B}$ in this reduction is symmetrically diagonally dominant (SDD). So, even for these highly structured matrices which admit near linear time application of $\mathbf{B}^{-1}$ [ST04] as well as approximate factorization [KS16], accurate spectral sums are difficult. We leverage this in Section 5.3.

5.2 Hardness for Computing Spectral Sums

Here we use Theorem 15 to show hardness for various spectral sum problems. To simplify our presentation, we focus on the case of dense matrices, showing bounds of the form $\Omega(n^\gamma \epsilon^{-c})$. However, note that Theorem 15 and the approach we use also yields lower bounds on the running time for sparse matrices and can be stated in terms of $\text{nnz}(\mathbf{A})$.

Schatten p -norm for all $p \neq 2$

For $x \in (0, 2)$, using the Taylor Series about 1 we can write

$$x^p = \sum_{k=0}^{\infty} c_k (1-x)^k \text{ where } c_k = \frac{\prod_{i=0}^{k-1} (p-i)}{k!} \quad (14)$$

This series converges since $|c_k| \leq 1$ for $k > p$ and for $x \in (0, 2)$, $(1-x) < 1$. Note that when p is a non-negative integer, only the first p terms of the expansion are nonzero. When p is non-integral, the sum is infinite. We will apply Theorem 15 slightly differently for different values of p . We first give our strongest result:

Corollary 16 (Schatten 3-Norm Hardness). *Given an algorithm which on input $\mathbf{B} \in \mathbb{R}^{n \times n}$ returns $X \in (1 \pm \epsilon) \|\mathbf{B}\|_3^3$ in $O(n^\gamma \epsilon^{-c})$ we can produce an algorithm that detects if an n -node graph contains a triangle in $O(n^{\gamma+4c})$ time.*

Proof. For $p = 3$, $c_k = 0$ for $k > 3$. So we apply Theorem 15 with $h = 0$ and hence $\delta = 1/n$ and $\epsilon_1 = \frac{c_3 \delta^3}{c_0 n} = \frac{1}{n^4}$. $\square$

Note that for p very close to 3 a similar bound holds as $h \approx 0$. If $p = 3$ Theorem 31 gives an algorithm running in $\tilde{O}(n^2/\epsilon^3)$ time. Significant improvement to the ϵ dependence in this algorithm therefore either requires loss in the n dependence or would lead to $o(n^w)$ time triangle detection for the current value of ω . We next extend to all $p \neq 1, 2$.

Corollary 17 (Schatten p -Norm Hardness, $p \neq 1, 2$). *For any $p > 0$, $p \neq 1, 2$, given algorithm $\mathcal{A}$ which for any $\mathbf{B} \in \mathbb{R}^{n \times n}$ returns $X \in (1 \pm \epsilon) \|\mathbf{B}\|_p^p$ in $O(n^\gamma \epsilon^{-c})$ time, we can detect if an n -node graph contains a triangle in $O\left(n^{\gamma+13c} \cdot \frac{p^{3c}}{|\min\{p, (p-1), (p-2)\}|^c}\right)$ time.*

Proof. We have $\frac{c_k}{c_3} \leq p^{k-3}$ for all $k > 3$ as well as $\left|\frac{c_0}{c_3}\right| = \left|\frac{1}{p(p-1)(p-2)}\right| \leq \left|\frac{1}{2 \min\{p, (p-1), (p-2)\}}\right|$ and similarly $\left|\frac{c_2}{c_3}\right| \leq \left|\frac{1}{2 \min\{p, (p-1)\}}\right|$. We apply Theorem 15 with $\delta = \Theta\left(\frac{1}{n^4 p}\right)$ and $\epsilon_1 = \frac{c_3 \delta^3}{c_0 n} = \Theta\left(\frac{|\min\{p, (p-1), (p-2)\}|}{n^{13p^3}}\right)$, which gives the result. $\square$

In the typical case when $p \ll n$, the p^{3c} term above is negligible. The $\frac{1}{|\min\{p, (p-1), (p-2)\}|^c}$ term is meaningful however. Our bound becomes weak as p approaches 2 (and meaningless when $p = 2$). This is unsurprising, as for p very close to 2, $\|\mathbf{B}\|_p^p \approx \|\mathbf{B}\|_F^2$, which can be computed exactly in $\text{nnz}(\mathbf{B})$ time. The bound also becomes weak for $p \approx 1$, which is natural as our reduction only uses PSD B , for which $\|\mathbf{B}\|_1 = \text{tr}(\mathbf{B})$ which can be computed in n time. However, we can remedy this issue by working with a (non-PSD) square root of $\mathbf{B}$ which is easy to compute:

Corollary 18 (Schatten p -Norm Hardness, $p \approx 1$). *For any p , given an algorithm which for any $\mathbf{B} \in \mathbb{R}^{p \times n}$ returns $X \in (1 \pm \epsilon) \|\mathbf{B}\|_p^p$ in $O(f(\text{nnz}(\mathbf{B}), n) \cdot \frac{1}{\epsilon^c})$ time, we can detect if an n -node graph with m edges contains a triangle in $O\left(f(m, n) \cdot n^{13c} \cdot \frac{p^{3c/2}}{|\min\{p/2, (p/2-1), (p/2-2)\}|^c} + m\right)$ time.*

Note that for $p \approx 1$, $\frac{p^{3c/2}}{|\min\{p/2, (p/2-1), (p/2-2)\}|^c}$ is just a constant. Again, the bound is naturally weak when $p \approx 2$ as $(p/2 - 1)$ goes to 0

Proof. For $\mathbf{B} = \mathbf{I} - \delta \mathbf{A}$ as in Theorem 15. Let $\mathbf{L} = \mathbf{D} - \mathbf{A}$ be the Laplacian of G where $\mathbf{D}$ is the diagonal degree matrix. We can write $\mathbf{B} = \delta \mathbf{L} + \widehat{\mathbf{D}}$ where $\widehat{\mathbf{D}} = \mathbf{I} - \delta \mathbf{D}$ is PSD since $\delta \leq 1/n$. Letting $\mathbf{M} \in \mathbb{R}^{\binom{n}{2} \times n}$ be the vertex edge incidence matrix of $\mathbf{A}$, and $\mathbf{N} = [\delta^{1/2} \mathbf{M}^T, \widehat{\mathbf{D}}^{1/2}]$, we have $\mathbf{N} \mathbf{N}^T = \mathbf{B}$. Thus, $\|\mathbf{N}\|_p = \|\mathbf{B}\|_p^{p/2}$ and so approximating this norm gives triangle detection by Corollary 17. Note that $\text{nnz}(\mathbf{N}) = O(\text{nnz}(\mathbf{A}))$ and further $\mathbf{N}$ matrix can be formed in this amount of time, giving our final runtime claim. $\square$

Note that for $p = 1$, since $\mathbf{N}$ has maximum row sparsity 2, we obtain a runtime via Theorem 33 of $\tilde{O}(\epsilon^{-3} (mn^{1/3} + n^{3/2})) = o(n^\omega)$ for the current value of ω , even when $m = n^2$, implying that significantly improving this ϵ dependence would either improve matrix multiplication or come at a cost in the polynomials of the other parameters.

SVD Entropy:

Corollary 19 (SVD Entropy Hardness). *Given algorithm $\mathcal{A}$ which for any $\mathbf{B} \in \mathbb{R}^{n \times n}$ returns $X \in (1 \pm \epsilon) \sum_{i=1}^n f(\sigma_i(\mathbf{B}))$ for $f(x) = x \log x$ in $O(n^\gamma \epsilon^{-c})$ time, we can detect if an n -node graph contains a triangle in $O(n^{\gamma+6c})$ time.*

Proof. For $x \in (0, 2)$, using the Taylor Series about 1 we can write $x \log x = \sum_{k=0}^{\infty} c_k (x-1)^k$ where $c_0 = 1 \log(1) = 0$, $c_1 = \log(1) + 1 = 1$, and $|c_k| = \frac{(k-2)!}{k!} \leq 1$ for $k \geq 2$. $c_k < c_3$ for all $k > 3$ and $\frac{c_0}{c_3} = 0$ while $\frac{c_2}{c_3} = \frac{1}{3}$. Applying Theorem 15 with $\delta = \frac{1}{10n^4}$ and $\epsilon_1 = \frac{\delta}{3n^2} = \frac{1}{30n^6}$ gives the result. $\square$

Log Determinant:

Corollary 20 (Log Determinant Hardness). *Given algorithm $\mathcal{A}$ which for any $\mathbf{B} \in \mathbb{R}^{n \times n}$ returns $X \in (1 \pm \epsilon) \log(\det(\mathbf{B}))$ in $O(n^\gamma \epsilon^{-c})$ time, we can detect if an n -node graph contains a triangle in $O(n^{\gamma+6c})$ time.*

Proof. Using the Taylor Series about 1 we can write $\log x = \sum_{k=0}^{\infty} c_k (x-1)^k$ where $c_0 = 0$, $|c_i| = 1/i$ for $i \geq 1$. Therefore $c_k < c_3$ for all $k > 3$, $\frac{c_0}{c_3} = 0$, and $\frac{c_2}{c_3} = \frac{3}{2}$. Applying Theorem 15 with $\delta = \frac{1}{10n^4}$ and $\epsilon_1 = \frac{\delta}{2n^2} = \frac{1}{20n^6}$ gives the result. $\square$

In Appendix B, Lemma 52 we show that a similar result holds for computing $\det(\mathbf{B}) = \prod_{i=1}^n \lambda_i(\mathbf{B})$. In [BS83] it is shown that, given an arithmetic circuit for computing $\det(\mathbf{B})$, one can generate a circuit of the same size (up to a constant) that computes $\mathbf{B}^{-1}$. This also yields a

circuit for matrix multiplication by a classic reduction.³ Our results, combined with the reduction of [WW10] of Boolean matrix multiplication to triangle detection, show that a sub-cubic time algorithm for the approximating $\log(\det(\mathbf{B}))$ or $\det(\mathbf{B})$ up to sufficient accuracy, yields a sub-cubic time matrix multiplication algorithm, providing a reduction based connection between determinant and matrix multiplication analogous to the circuit based result of [BS83].

Trace of Exponential:

Corollary 21 (Trace of Exponential Hardness). *Given algorithm $\mathcal{A}$ which for any $\mathbf{B} \in \mathbb{R}^{n \times n}$ returns $X \in (1 \pm \epsilon) \text{tr}(\exp(\mathbf{B}))$ in $O(n^\gamma \epsilon^{-c})$ time, we can detect if an n -node graph contains a triangle in $O(n^{\gamma+13c})$ time.*

Proof. Using the Taylor Series about 1 we can write $e^x = \sum_{k=0}^{\infty} \frac{e(x-1)^k}{k!}$. We have $\frac{c_0}{c_3} = 6$, $\frac{c_2}{c_3} = 3$, and for all $k \geq 3$, $c_k < c_3$. Applying Theorem 15 with $\delta = \frac{1}{10n^4}$ and $\epsilon_1 = \frac{c_3 \delta^3}{c_0 n} = \frac{1}{6000n^{13}}$ gives the result. $\square$

5.3 Leverage Score and Effective Resistance Hardness

Here we show hardness for precisely computing all effective resistances and leverage scores of a graph. Our main tool is the following result (which in turn is an easy corollary of Theorem 15) for an algorithm that precisely computes the trace inverse of a strictly symmetric diagonally dominant (SDD) $\mathbf{B}$, i.e. $\mathbf{B} = \mathbf{B}^\top$ and $\mathbf{B}_{ii} > \sum_{j \neq i} \mathbf{B}_{ij}$.

Corollary 22 (Trace of Inverse Hardness). *Given an algorithm which for any strictly SDD $B \in \mathbb{R}^{n \times n}$ with non-positive off-diagonal entries returns $X \in (1 \pm \epsilon) \text{tr}(B^{-1})$ in $O(n^\gamma \epsilon^{-c})$ time for $\gamma \geq 2$, we can produce an algorithm which detects if an n -node graph contains a triangle in $O(n^{\gamma+13c})$ time.*

Proof. For $x \in (0, 2)$ we can write $\frac{1}{x} = \sum_{k=0}^{\infty} (1-x)^k$, and then apply Theorem 15 with $\delta = \frac{1}{10n^4}$ and $\epsilon_1 = \frac{\delta^3}{n} = \frac{1}{1000n^{13}}$. Checking that the $\mathbf{B}$ in Theorem 15 is strictly SDD with non-positive off-diagonal entries yields the result. $\square$

Using this we prove hardness for precisely computing effective resistances in a graph. Recall that for a weighted undirected graph $G = (V, E, w)$ its Laplacian, $\mathbf{L} \in \mathbb{R}^{V \times V}$ is given by $\mathbf{L}_{ij} = -w_{ij}$ if there is an edge between i and j and 0 otherwise and $\mathbf{L}_{ii} = -\sum_{i \neq j} \mathbf{L}_{ij}$ or equivalently $\mathbf{L} = \mathbf{D} - \mathbf{A}$ where $\mathbf{D}$ is the diagonal degree matrix and $\mathbf{A}$ is the weighted adjacency matrix associated with G . Note that this describes a clear bijection between a Laplacian and its associated graph and we therefore use them fairly interchangeably in the remainder of this section.

The effective resistance between vertices i and j is given by $(\mathbf{1}_i - \mathbf{1}_j)^\top \mathbf{L}^\dagger (\mathbf{1}_i - \mathbf{1}_j)$ where $\dagger$ denotes the Moore-Penrose pseudoinverse. In the following lemma we prove that compute all the effective resistances between a vertex and its neighbors in the graph can be used to compute the trace of the inverse of any strictly SDD matrix with non-positive off-diagonals and therefore doing this precisely is as hard as triangle detection via Corollary 22. Our proof is based off a fairly standard reduction between solving strictly SDD matrices with negative off-diagonals and solving Laplacian systems.

Lemma 23 (Effective Resistance Yields Trace). *Suppose we have an algorithm which given Laplacian $\mathbf{L} \in \mathbb{R}^{n \times n}$ with m -non-zero entries, entry $i \in [n]$, and error $\epsilon \in (0, 1)$ computes a $1 \pm \epsilon$*

³Matrix multiplication reduces to inversion by the fact that $\begin{bmatrix} \mathbf{I} & \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{B} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{I} & -\mathbf{A} & \mathbf{AB} \\ \mathbf{0} & \mathbf{I} & -\mathbf{B} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{bmatrix}$. See [Isa08].

approximation to the total effective resistance between i and the neighbors of i in the graph associated with $\mathbf{L}$, that is

$$X \in (1 \pm \epsilon) \sum_{j \in [n]: L_{ij} \neq 0} (\mathbf{1}_i - \mathbf{1}_j)^\top \mathbf{L}^\dagger (\mathbf{1}_i - \mathbf{1}_j)$$

in time $O(m^\gamma \epsilon^{-c})$. Then there is an algorithm that computes the trace of the inverse of $n \times n$ strict SDD matrix with m non-zero entries and non-positive off-diagonals in $O(m^\gamma \epsilon^{-c})$ time.

Proof. Let $\mathbf{M} \in \mathbb{R}^{n \times n}$ be an arbitrary strictly SDD matrix with non-positive off-diagonals, i.e. $\mathbf{M} = \mathbf{M}^\top$, $\mathbf{M}_{ii} > \sum_{j \neq i} |\mathbf{M}_{ij}|$, and $\mathbf{M}_{ij} \leq 0$ for all $i \neq j$. Let $\mathbf{v} \stackrel{\text{def}}{=} \mathbf{M}\mathbf{1}$, $\alpha \stackrel{\text{def}}{=} \mathbf{1}^\top \mathbf{M}\mathbf{1}$, and

$$\mathbf{L} \stackrel{\text{def}}{=} \begin{pmatrix} \mathbf{M} & -\mathbf{v} \\ -\mathbf{v}^\top & \alpha \end{pmatrix}.$$

Now, clearly by our assumptions on $\mathbf{M}$ we have that $\mathbf{v} > \mathbf{0}$ entrywise and therefore $\alpha > 0$. Therefore, the off-diagonal entries of $\mathbf{L}$ are non-positive and by construction $\mathbf{L}\mathbf{1} = \mathbf{0}$. Consequently, $\mathbf{L}$ is a $(n+1) \times (n+1)$ symmetric Laplacian matrix with $\text{nnz}(\mathbf{M}) + 2n + 1$ non-zero entries.

Now, consider any $x \in \mathbb{R}^n$ and $y \in \mathbb{R}$ that satisfy the following for some $i \in [n]$

$$\begin{pmatrix} \mathbf{M} & -\mathbf{v} \\ -\mathbf{v}^\top & \alpha \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ y \end{pmatrix} = \begin{pmatrix} \mathbf{1}_i \\ -1 \end{pmatrix}.$$

Since $\mathbf{L}$ is a symmetric Laplacian and the associated graph is connected by construction we know that $\ker(\mathbf{L}) = \text{span}(\{\mathbf{1}\})$ and there must exist such x and y . Furthermore, since $\mathbf{M}$ is strictly SDD it is invertible and since $\mathbf{M}\mathbf{1} = \mathbf{v}$ we have that

$$\mathbf{x} = \mathbf{M}^{-1}(y \cdot \mathbf{v} + \mathbf{1}_i) = y \cdot \mathbf{1} + \mathbf{M}^{-1}\mathbf{1}_i$$

and consequently

$$(\mathbf{1}_i - \mathbf{1}_{n+1})^\top \mathbf{L}^\dagger (\mathbf{1}_i - \mathbf{1}_{n+1}) = \mathbf{1}_i^\top \mathbf{x} - y = \mathbf{1}_i^\top \mathbf{M}^{-1} \mathbf{1}_i.$$

Consequently, if we used the algorithm to get a multiplicative approximation X as stated then $X \in (1 \pm \epsilon) \text{tr}(\mathbf{M}^{-1})$ and the result follows. $\square$

Using this, we also show that computing leverage scores of matrix, a powerful and prevalent notion in convex analysis and numerical linear algebra, is also difficult to compute. This follows from the well known fact that effective resistances in graphs and leverage scores of matrices are the same up to scaling by known quantities.

Corollary 24 (Leverage Score Hardness). *Suppose we have an algorithm which given $\mathbf{A} \in \mathbb{R}^{n \times d}$ can compute $\tilde{\sigma}$ that is a $1 \pm \epsilon$ multiplicative approximation to the leverage scores of $\mathbf{A}$, i.e.*

$$\tilde{\sigma}_i \in (1 \pm \epsilon) \mathbf{1}_i^\top \mathbf{A} (\mathbf{A}^\top \mathbf{A})^\dagger \mathbf{A}^\top \mathbf{1}_i \text{ for all } i \in [n]$$

in time $O(\text{nnz}(\mathbf{A})^\gamma \epsilon^{-c})$. Then there is a $O(n^{2\gamma+13c})$ time algorithm for detecting if an n -node graph contains a triangle.

Proof. Let $\mathbf{L} \in \mathbb{R}^{n \times n}$ be a symmetric Laplacian. Let $E = \{\{i, j\} \subseteq [n] \times [n] : \mathbf{L}_{ij} \neq 0\}$, i.e. the set of edges in the graph associated with L . Let $m = |E|$ and $\mathbf{B} \in \mathbb{R}^{m \times n}$ be the incidence matrix associated with $\mathbf{L}$, i.e. for all $e = \{i, j\} \in E$ we have $\mathbf{B}_{e,i} = \sqrt{-\mathbf{L}_{ij}}$ and $\mathbf{B}_{e,j} = -\sqrt{-\mathbf{L}_{ij}}$ for some canonical choice of ordering of i and j and let all other entries of $\mathbf{B} = 0$. Clearly $\text{nnz}(\mathbf{B}) = \text{nnz}(\mathbf{L})$ and we can form $\mathbf{B}$ in $O(\text{nnz}(\mathbf{L}))$ time.

It is well known and easy to check that $\mathbf{L} = \mathbf{B}^\top \mathbf{b}$. Consequently, for all $e = \{i, j\} \in E$ we have

$$\mathbf{1}_e^\top \mathbf{B}(\mathbf{B}^\top \mathbf{B})^\dagger \mathbf{B}^\top \mathbf{1}_e = (-\mathbf{L}_{ij}) \cdot (\mathbf{1}_i - \mathbf{1}_j)^\top \mathbf{L}^\dagger (\mathbf{1}_i - \mathbf{1}_j)$$

Now if we compute $\tilde{\sigma}$ using the assumed algorithm in $O(\text{nnz}(\mathbf{L})^\gamma \epsilon^{-c}) = O(n^{2\gamma} \epsilon^{-c})$ time, then since $-\mathbf{L}_{ij}$ is non-negative in an additional $O(\text{nnz}(\mathbf{L}) = O(n^2))$ time this yields a $1 \pm \epsilon$ multiplicative approximation to the total effective resistance between any i and all its neighbors in the graph associated with $\mathbf{L}$. Consequently, by Lemma 23 and Corollary 22 the result follows. $\square$

6 Improved Error Dependence via Polynomial Approximation

For constant ϵ , Theorem 11 and the resulting Corollary 12 matches our fastest runtimes for the Schatten p -norms. However, it is possible to significantly improve the ϵ dependence by generalizing our approach. Instead of splitting our spectrum into many small spectral windows, we split into windows of constant multiplicative width and approximate x^p via a low degree polynomial over each window. The degree of this polynomial only varies with $\log(1/\epsilon)$.

In Theorem 11, fixing δ_f to be constant for illustration, each window has width $\alpha = \Theta(\epsilon)$ and so there are $T = \lceil \log_{1-\alpha} \lambda \rceil = \Theta(\log \lambda / \epsilon)$ windows. Additionally, we must set the steepness parameter to be $\gamma = \Theta(\epsilon \alpha) = \Theta(\epsilon^2)$. This loses us a total $1/\epsilon^2$ factor in our runtime as compared to our improved algorithm which will set $\alpha = \Theta(1)$ and so $\gamma = \Theta(\epsilon)$.

We begin by showing that if we well approximate f on each window, then we well approximate $\mathcal{S}_f(\mathbf{A})$ overall. In the following theorems for simplicity we work with PSD matrices, as we will eventually apply these lemmas to $\mathbf{A}^T \mathbf{A}$.

Lemma 25. *Consider $f : \mathbb{R} \rightarrow \mathbb{R}^+$, parameters $\alpha, \epsilon \in (0, 1)$, and gap parameter $\gamma \in (0, \alpha)$. Set $a_0 = 1$, pick a_1 uniformly at random from $[1 - \alpha, 1]$, and set $a_t = a_1(1 - \alpha)^{t-1}$ for $t \geq 2$. Let $R_t = [a_{t+1}, a_t]$ for all $t \geq 0$. Let $h_{R_t}^\gamma$ be a γ -soft window for R_t (Definition 6) and let $\bar{R}_t = [(1 - \gamma)a_{t+1}, (1 + \gamma)a_t]$ be the interval on which $h_{R_t}^\gamma$ is nonzero. Furthermore, for $t \geq 0$, let q_t be a ‘well-behaved’ approximation to f on $\bar{R}_t$ in the following sense:*

- *Multiplicative Approximation:* $|q_t(x) - f(x)| \leq \epsilon f(x)$ for $x \in \bar{R}_t$.
- *Approximate Monotonicity:* $q_t(y) \leq c_1 q_t(x)$ for all $y \leq x$ for some $c_1 \geq 1$.
- *Range Preserving:* $q_t(0) = 0$.

Then, for any PSD $\mathbf{A} \in \mathbb{R}^{d \times d}$ with $\|\mathbf{A}\|_2 \leq 1$, with probability $9/10$:

$$\left(1 - \epsilon - \frac{40c_1\gamma}{\alpha}\right) \mathcal{S}_f(\mathbf{A}) \leq \sum_{t=0}^{\infty} \mathcal{S}_{q_t}(h_{R_t}^\gamma(\mathbf{A})\mathbf{A}) \leq \left(1 + \epsilon + \frac{40c_1\gamma}{\alpha}\right) \mathcal{S}_f(\mathbf{A}).$$

Proof. Note that the restriction that $q_t(0) = 0$ along with the approximate monotonicity property implies that $q(x)$ is nonnegative, so $\mathcal{S}_{q_t}$ is a valid spectral sum and $\mathcal{S}_{q_t}(h_{R_t}^\gamma(\mathbf{A})\mathbf{A}) = \text{tr}(q_t(h_{R_t}^\gamma(\mathbf{A})\mathbf{A}))$, so we will be able to estimate this sum via stochastic trace estimation.

Let t_i denote the unique index such that $\sigma_i(\mathbf{A}) \in R_{t_i}$. Since $\gamma < \alpha$, each $\sigma_i(\mathbf{A})$ lies in the support of at most two overlapping soft windows (in at most two ranges $\bar{R}_{t_i}$ and $\bar{R}_{t_i \pm 1}$). Let T be the set of indices whose singular values fall in the support of exactly one soft window and $\bar{T}$ be its

complement. We first bound the error introduced on singular values with indices in T .

$$\begin{aligned}
\sum_{t=0}^{\infty} \mathcal{S}_{q_t}(h_{R_t}^{\gamma}(\mathbf{A})\mathbf{A}) &= \sum_{i=1}^d \sum_{t=0}^{\infty} q_t(\sigma_i(\mathbf{A}) \cdot h_{R_t}^{\gamma}(\sigma_i(\mathbf{A}))) \\
&= \sum_{i \in T} q_{t_i}(\sigma_i(\mathbf{A})) + \sum_{i \in \bar{T}} \sum_{t=0}^{\infty} q_t(\sigma_i(\mathbf{A}) \cdot h_{R_t}^{\gamma}(\sigma_i(\mathbf{A}))) \\
&\in (1 \pm \epsilon) \sum_{i \in T} f(\sigma_i(\mathbf{A})) + \sum_{i \in \bar{T}} \sum_{t=0}^{\infty} q_t(\sigma_i(\mathbf{A}) \cdot h_{R_t}^{\gamma}(\sigma_i(\mathbf{A}))). \tag{15}
\end{aligned}$$

The last inequality follows from the multiplicative approximation assumption on q_t that $|q_t(x) - f(x)| \leq \epsilon f(x)$ for $x \in \bar{R}_t$. Let us now consider a particular $i \in \bar{T}$ and bound $\sum_{t=0}^{\infty} q_t(\sigma_i(\mathbf{A}) \cdot h_{R_t}^{\gamma}(\sigma_i(\mathbf{A})))$. Note that there are precisely two non-zero terms in this summation – one corresponding to t_i and the other, to $t_i \pm 1$, which we denote by $\bar{t}_i$. Using the above hypothesis on q_t again, we see that

$$q_{t_i}(\sigma_i(\mathbf{A}) \cdot h_{R_{t_i}}^{\gamma}(\sigma_i(\mathbf{A}))) = q_{t_i}(\sigma_i(\mathbf{A})) \in (1 \pm \epsilon) f(\sigma_i(\mathbf{A})). \tag{16}$$

For the term involving $\bar{t}_i$, we have $\sigma_i(\mathbf{A}) \in \bar{R}_{\bar{t}_i}$ and by the approximate monotonicity requirement that $y \leq x \Rightarrow q_t(y) \leq c_1 q_t(x)$ have:

$$q_{\bar{t}_i}(\sigma_i(\mathbf{A}) \cdot h_{R_{\bar{t}_i}}^{\gamma}(\sigma_i(\mathbf{A}))) \leq c_1 q_{\bar{t}_i}(\sigma_i(\mathbf{A})) \in c_1 (1 \pm \epsilon) f(\sigma_i(\mathbf{A})). \tag{17}$$

Due to the random positioning of the windows, in expectation, the total contribution of the singular values lying at the intersection of two windows is small. Specifically,

$$\mathbb{E}_{\alpha_1} \sum_{i \in \bar{T}} f(\sigma_i(\mathbf{A})) = \sum_{i=1}^d \mathbb{P}[i \in \bar{T}] \cdot f(\sigma_i(\mathbf{A})) \leq \frac{2\gamma}{\alpha} \sum_{i=1}^d f(\sigma_i(\mathbf{A})) = \frac{2\gamma}{\alpha} \mathcal{S}_f(\mathbf{A}) \tag{18}$$

where the bound on $\mathbb{P}[i \in \bar{T}]$ follows from the fact that $i \in \bar{T}$ only if $a_1(1 - \alpha)^t \in (1 \pm \gamma)\sigma_i(\mathbf{A})$ for some t . This holds only if $a_1 \in (1 \pm \gamma) \left(\frac{\sigma_i(\mathbf{A})}{(1 - \alpha)^{\lceil \log_{\sigma_i(\mathbf{A})} 1 - \alpha \rceil}} \right)$, which occurs with probability $\frac{2\gamma}{\alpha}$ since a_1 is chosen uniformly in the range $[1 - \alpha, 1]$. By a Markov bound applied to (18), with probability 9/10 we have $\sum_{i \in \bar{T}} f(\sigma_i(\mathbf{A})) \leq \frac{20\gamma}{\alpha} \mathcal{S}_f(\mathbf{A})$. Combining with (15), (16), and (17), we obtain, with probability 9/10:

$$\left(1 - \epsilon - \frac{40c_1\gamma}{\alpha}\right) \mathcal{S}_f(\mathbf{A}) \leq \sum_{t=0}^{\infty} \mathcal{S}_{q_t}(h_{R_t}^{\gamma}(\mathbf{A})\mathbf{A}) \leq \left(1 + \epsilon + \frac{40c_1\gamma}{\alpha}\right) \mathcal{S}_f(\mathbf{A}).$$

□

We now show that, as long as the contribution of the smaller singular values of $\mathbf{A}$ to $\mathcal{S}_f(\mathbf{A})$ is not too large, we can truncate this sum and still obtain an accurate approximation. Specifically:

Corollary 26. *For any PSD $\mathbf{A} \in \mathbb{R}^{d \times d}$ with $\|\mathbf{A}\|_2 \leq 1$, let $f : \mathbb{R} \rightarrow \mathbb{R}^+$ be a function such that, for any $\epsilon > 0$ there exists $\lambda_f(\epsilon)$ such that for $x \in [0, \lambda_f(\epsilon)]$, $f(x) \leq \frac{\epsilon}{n} \mathcal{S}_f(\mathbf{A})$. Given parameters $\alpha, \epsilon \in (0, 1)$, and gap parameter $\gamma < \alpha$, for $t \geq 0$ define R_t , $h_{R_t}^{\gamma}$ and, $\bar{R}_t$ as in Lemma 25. Let q_t be a well-behaved approximation to f on $\bar{R}_t$ as in Lemma 25. With probability 9/10:*

$$\left(1 - \frac{40c_1\gamma}{\alpha} - 5c_1\epsilon\right) \mathcal{S}_f(\mathbf{A}) \leq \sum_{t=0}^{\lceil \log_{1-\alpha} \lambda_f(\epsilon) \rceil} \mathcal{S}_{q_t}(h_{R_t}^{\gamma}(\mathbf{A})\mathbf{A}) \leq \left(1 + \frac{40c_1\gamma}{\alpha} + 5c_1\epsilon\right) \mathcal{S}_f(\mathbf{A}). \tag{19}$$

Proof. This follows from Lemma 25 along with the small tail assumption. Specifically, by (17):

$$\sum_{\lceil \log_{1-\alpha} \lambda_f(\epsilon) \rceil + 1}^{\infty} \mathcal{S}_{q_t}(h_{R_t}^{\gamma}(\mathbf{A})\mathbf{A}) \leq \sum_{\{i|\sigma_i(\mathbf{A}) \leq \lambda_f(\epsilon)\}} 4c_1 f(\sigma_i(\mathbf{A})) \leq 4c_1 n \cdot \frac{\epsilon}{n} \mathcal{S}_f(\mathbf{A}) = 4c_1 \epsilon \mathcal{S}_f(\mathbf{A}).$$

□

We now show that well behaved polynomial approximations exist for the function x^p for general real p , whose spectral sums give the Schatten p -norms.

Lemma 27 (Polynomial Approximation of Power Function). *For all $p \in [-1, 0)$ and $k \geq 0$ let*

$$a_k \stackrel{\text{def}}{=} \prod_{j=1}^k \left(1 - \frac{p+1}{j}\right) \quad \text{and} \quad q_k(x) = \sum_{j=0}^k a_j (1-x)^j.$$

Then for all $x \in (0, 1]$ and $k \geq 0$ we have $0 \leq a_k \leq 1$ and

$$0 \leq x^p - q_k(x) \leq \frac{\exp(-kx)}{x}.$$

Proof. Induction shows that the k -th derivative of $f(x) = x^p$ at $x \in \mathbb{R}$, is given by

$$f^{(k)}(x) = \left(\prod_{j=1}^k (p+1-j) \right) x^{p-k}.$$

Furthermore, direct calculation reveals that for all $x, t \in \mathbb{R}$ and $k \geq 1$

$$\frac{f^{(k)}(t)}{k!} (x-t)^k = \left(\prod_{j=1}^k \frac{p+1-j}{j} \right) \cdot t^{p-k} \cdot (t-x)^k \cdot (-1)^k = a_k t^p \left(1 - \frac{x}{t}\right)^k. \quad (20)$$

Consequently, $q_k(x)$ is the degree k Taylor expansion of x^p about 1 evaluated at x . Furthermore, since $f^{(k+1)}(t) = \frac{p-k}{t} \cdot f^{(k)}(t)$, the integral form of the remainder of a Taylor series expansion shows

$$|x^p - q_k(x)| = \left| \int_1^x \frac{f^{(k+1)}(t)}{k!} (x-t)^k dt \right| = |a_k| \cdot |p-k| \cdot \left| \int_1^x t^{p-1} \left(1 - \frac{x}{t}\right)^k dt \right| \leq \frac{k-p}{x} \exp(-kx).$$

In the last step we took the worst case of $t = 1$ in the integral and used that since $p \in [-1, 0)$ it is the case that $1 - (p+1)/j \in (0, 1]$ and therefore $0 \leq a_k \leq 1$. Consequently, $\lim_{i \rightarrow \infty} q_k(x) = x^p$ and

$$0 \leq x^p - q_k(x) = \sum_{j=k+1}^{\infty} a_j (1-x)^j \leq (1-x)^{k+1} \sum_{j=0}^{\infty} (1-x)^j \leq \frac{\exp(-kx)}{x}.$$

□

Corollary 28. *For any $p, \epsilon > 0$, $a, b \in (0, 1]$ with $a < b$, there is a polynomial q of degree $O\left(\log\left(\frac{b}{a\epsilon}\right) \cdot \frac{b}{a} + p\right)$ such that:*

- *Multiplicative Approximation:* $|q(x) - x^p| \leq \epsilon x^p$ for $x \in [a, b]$.
- *Monotonicity:* $q(y) < q(x)$ for all $y < x$.

- *Range Preserving*: $q(0) = 0$.

Proof. Set $p' = p - \lceil p \rceil$ and $i = c_1 \log\left(\frac{b}{a\epsilon}\right) \cdot \frac{b}{a}$ for large enough constant c_1 . Instantiate Lemma 27 with p' and i to obtain $q_i(x)$. Set $q(x) = b^p \cdot (x/b)^{\lceil p \rceil} \cdot q_i(x/b)$. It is clear that $q(0) = 0$. Further,

$$\begin{aligned} |x^p - q(x)| &= b^p (x/b)^{\lceil p \rceil} \cdot |(x/b)^{p'} - q_i(x/b)| \\ &\leq x^p (b/x)^{1+p'} \cdot \exp\left(-c_1 \log\left(\frac{b}{a\epsilon}\right) \frac{b}{a} \cdot \frac{x}{b}\right) \end{aligned}$$

For $x \in [a, b]$, $x/b \geq a/b$. Further, $(b/x)^{1+p'} \leq b/a$ and so if we choose c_1 large enough we have $|x^p - q(x)| < \epsilon x^p$. We finally show monotonicity. We have $x^{p'} - q_i(x) = \sum_{j=i+1}^{\infty} a_j (1-x)^j$. All a_j are positive, so this difference is monotonically decreasing in x . We thus have, for $y < x$:

$$\frac{q(y)}{q(x)} = \frac{(y/b)^{\lceil p \rceil} q_i(y/b)}{(x/b)^{\lceil p \rceil} q_i(x/b)} \leq \frac{y}{x} \cdot \frac{q_i(y/b)}{q_i(x/b)} \leq \frac{y}{x} \cdot \frac{(y/b)^{p'}}{(x/b)^{p'}} \leq 1$$

since $p' \in [0, 1]$. This gives us monotonicity. $\square$

We now combine the approximations of Corollary 28 with Corollary 26 to give our improved Schatten norm estimation algorithm.

Lemma 29 (Schatten Norms via Polynomial Approximation). *For any $\mathbf{A} \in \mathbb{R}^{n \times d}$ with $\|\mathbf{A}\|_2 \leq 1$, $p > 0$ and error parameter $\epsilon \in (0, 1)$, for sufficiently small c_1, c_2 , let $\alpha = 1/2$, $\lambda = \left(\frac{c_1^{2/p} \epsilon^{2/p}}{d^{2/p}}\right) \|\mathbf{A}\|_p^2$, and $\gamma = c_2 \epsilon$. For $t \geq 0$ define R_t , $h_{R_t}^\gamma$, and $\bar{R}_t$ as in Lemma 25. Let $q_t(x)$ be as defined in Corollary 28 with $p' = p/2$, $\epsilon' = c_3 \epsilon$ for sufficiently small c_3 and $[a, b] = R_t$. With probability 9/10:*

$$\sum_{t=0}^{\lceil \log 1/\lambda \rceil} \mathcal{S}_{q_t}(h_{R_t}^\gamma(\mathbf{A}^T \mathbf{A}) \mathbf{A}^T \mathbf{A}) \in (1 \pm \epsilon) \|\mathbf{A}\|_p^p. \quad (21)$$

Proof. We apply Corollary 26 to $\mathbf{A}^T \mathbf{A}$, with $f(x) = x^{p/2}$, $\epsilon = c_1 \epsilon$, $\gamma = c_2 \epsilon$ and $\alpha = 1/2$. q_t satisfies the necessary conditions by Corollary 28. Further, we have $\lambda_f(c_1 \epsilon) = \left(\frac{c_1^{2/p} \epsilon^{2/p}}{d^{2/p}}\right) \|\mathbf{A}\|_p^2 = \lambda$. Plugging in $\alpha = 1/2$, $\gamma = c_2 \epsilon$ into (19) we have:

$$\sum_{t=0}^{\lceil \log 1/\lambda \rceil} \mathcal{S}_{q_t}(h_{R_t}^\gamma(\mathbf{A}^T \mathbf{A}) \mathbf{A}^T \mathbf{A}) \in \left(1 \pm c_3 \epsilon \pm \frac{40c_2 \epsilon}{1/2} \pm c_1 \epsilon\right) \mathcal{S}_f(\mathbf{A} \mathbf{A}^T) = \|\mathbf{A}\|_p^p$$

which gives $(1 \pm \epsilon)$ approximation if we set c_1, c_2, c_3 small enough. $\square$

Theorem 30 (Schatten Norm Polynomial Approximation Algorithm). *For any $\mathbf{A} \in \mathbb{R}^{n \times d}$, $p > 0$ and $k \in [d]$ there is an algorithm returning $X \in (1 \pm \epsilon) \|\mathbf{A}\|_p^p$ that runs in time:*

$$\begin{aligned} &\tilde{O}\left(\text{nnz}(\mathbf{A})k + dk^{\omega-1} + \frac{\max\{p, 1/p^3\}}{\epsilon^{\max\{3, 1+1/p\}}} \cdot \left[\text{nnz}(\mathbf{A}) + \sqrt{\text{nnz}(\mathbf{A})[d \cdot d_s(\mathbf{A}) + dk]} \cdot (d/k)^{\max\{0, 1/p-1/2\}}\right]\right) \\ &\text{or } \tilde{O}\left(\text{nnz}(\mathbf{A})k + dk^{\omega-1} + \frac{\max\{p, 1/p^3\}}{\epsilon^{\max\{3, 1+1/p\}}} \cdot \left[(\text{nnz}(\mathbf{A}) + dk)(d/k)^{1/p}\right]\right) \end{aligned}$$

time for sparse $\mathbf{A}$ or

$$\tilde{O}\left(nd^{\omega(\log_d k)-1} + \frac{\max\{p, 1/p^3\}}{\epsilon^{\max\{3, 1+1/p\}}} \cdot \left[nd + \frac{n^{1/2}d^{1+1/p}}{k^{\max\{0, 1/p-1/2\}}}\right]\right)$$

for dense $\mathbf{A}$.

Proof. We apply Lemma 29, first scaling $\mathbf{A}$ so that $\|\mathbf{A}\|_2 \leq 1$. We apply a trace estimator to approximate $\mathcal{S}_{q_t}(h_{R_t}^\gamma(\mathbf{A}^T \mathbf{A}) \mathbf{A}^T \mathbf{A}) = \text{tr}(q_t(h_{R_t}^\gamma(\mathbf{A}^T \mathbf{A}) \mathbf{A}^T \mathbf{A}))$ up to $(1 \pm \epsilon)$ multiplicative error plus very small additive error for each of our $O(\log 1/\lambda) = \tilde{O}(\max\{1, 1/p\})$ windows. The trace estimation requires $\tilde{O}\left(\frac{\lceil p \rceil}{\epsilon^2}\right)$ applications of $h_{R_t}^\gamma(\mathbf{A}^T \mathbf{A}) \mathbf{A}^T \mathbf{A}$, since the degree of q_t as given in Corollary 28 is $\tilde{O}(1) + p$.

The cost to apply $h_{R_t}^\gamma(\mathbf{A}^T \mathbf{A}) \mathbf{A}^T \mathbf{A}$ to a vector is given by Theorem 7. Following the argument in Corollary 12, if we write $a_t = \frac{\delta_t^{2/p}}{d^{2/p}} \cdot \|\mathbf{A}\|_p^2$ for some $\delta_t \geq \epsilon$,

$$\bar{\kappa}_t = \frac{k\sigma_k^2(\mathbf{A}) + \sum_{k=1}^d \sigma_i^2(\mathbf{A})}{da_t} \leq \frac{1}{\delta_t^{2/p}} \left(\frac{d}{k}\right)^{2/p-1}.$$

Similarly, $\hat{\kappa}_t = \frac{\sigma_{k+1}^2(\mathbf{A})}{a_t} \leq \frac{1}{\delta_t^{2/p}} \left(\frac{d}{k}\right)^{2/p}$. We can further optimize our ϵ dependence by noting that it suffices to approximate $\text{tr}(q_t(h_{R_t}^\gamma(\mathbf{A}^T \mathbf{A}) \mathbf{A}^T \mathbf{A}))$ up to multiplicative error $1 \pm \tilde{O}\left(\frac{\epsilon}{\min\{1, \delta_t\} \log(1/\lambda)}\right)$ since even if there are d singular values below a_{t-1} , they will contribute at most a δ_t fraction of $\|\mathbf{A}\|_p^2$. So our total cost of approximating the trace for each of our windows using the first runtime of Theorem 7 is:

$$\tilde{O}\left(\sum_{t=0}^{\lceil \log 1/\lambda \rceil} \frac{\min\{1, \delta_t\}^2 \cdot \lceil p \rceil (\log 1/\lambda)^2 \cdot \text{nnz}(\mathbf{A}) + \sqrt{\text{nnz}(\mathbf{A})[d \cdot d_s(\mathbf{A}) + dk]} \cdot (d/k)^{1/p-1/2}}{\epsilon^2 \cdot \delta_t^{1/p}}\right)$$

Factoring out the $\text{nnz}(\mathbf{A}) + \sqrt{\text{nnz}(\mathbf{A})[d \cdot d_s(\mathbf{A}) + dk]} \cdot (d/k)^{1/p-1/2}$ term, we have:

$$\begin{aligned} \tilde{O}\left(\sum_{t=0}^{\lceil \log 1/\lambda \rceil} \frac{\min\{1, \delta_t\}^2 \cdot \max\{p, 1/p^2\}}{\epsilon^3 \delta_t^{1/p}}\right) &= \tilde{O}\left(\frac{\max\{p, 1/p^2\}}{\epsilon^3} \cdot \left[\sum_{\{t: \delta_t < 1\}} \frac{1}{\delta_t^{1/p-2}} + \sum_{\{t: \delta_t \geq 1\}} \frac{1}{\delta_t^{1/p}}\right]\right) \\ &= \tilde{O}\left(\frac{\max\{p, 1/p^2\}}{\epsilon^3} \cdot \left[\frac{1}{\epsilon^{1/p-2}} \left(1 + \frac{1}{2^{1/p-2}} + \frac{1}{4^{1/p-2}} + \dots + \epsilon^{1/p-2}\right) + \log(1/\lambda)\right]\right). \end{aligned}$$

Note that if $p > 2$, $1/p - 2 < 0$ the terms in the geometric sum are increasing so the largest term is $\epsilon^{1/p-2}$ and so the whole thing is dominated by $\tilde{O}(\frac{p \log(1/\lambda)}{\epsilon^3}) = \tilde{O}(\frac{p}{\epsilon^3})$. If $p < 2$, then the largest term in the geometric sum is 1 as so similarly the whole thing is $\tilde{O}(\frac{1/p^2 \log(1/\lambda)}{\epsilon^{\max\{3, 1+1/p\}}}) = \tilde{O}(\frac{1/p^3}{\epsilon^{\max\{3, 1+1/p\}}})$.

A similar argument gives the runtimes for standard iterative and dense methods. Adding in the cost for deflation-based preconditioning gives the final runtimes. Note that while λ depends on $\|\mathbf{A}\|_p^p$ which is unknown, it again suffices to lower bound the truncation point using $\sigma_k^p(\mathbf{A}) \leq \frac{1}{k} \|\mathbf{A}\|_p^p$. This lower bound gives only a better approximation and our bounds on $\bar{\kappa}_t$ and $\hat{\kappa}_t$ still hold. $\square$

7 Optimized Runtime Results

The runtimes given in Theorem 30 are quite complex, with many tradeoffs depending on the properties of $\mathbf{A}$. In this section, we instantiate the theorem showing how to optimize k and achieve our final runtime bounds. For simplicity, we consider the case when $n = d$.

7.1 Schatten p -Norms

We first tackle the relatively simply case when $\mathbf{A}$ is dense.

Theorem 31 (Schatten p -Norms for Dense Matrices). *For any $p \geq 0$ and $\mathbf{A} \in \mathbb{R}^{n \times n}$ there is an algorithm returning $X \in (1 \pm \epsilon) \|\mathbf{A}\|_p^p$ which runs in $\tilde{O}(p \cdot n^2/\epsilon^3)$ time for $p \geq 2$ and*

$$\tilde{O}\left(\frac{1}{p^3 \cdot \epsilon^{\max\{3, 1+1/p\}}} \cdot n^{\frac{2.3729 - .0994p}{1 + .0435p}}\right)$$

time for $p < 2$. If we do not use fast matrix multiplication the runtime is $\tilde{O}\left(\frac{1}{p^3 \cdot \epsilon^{\max\{3, 1+1/p\}}} \cdot n^{\frac{3+p/2}{1+p/2}}\right)$.

Note that for the important case of $p = 1$ our runtime is $\tilde{O}(n^{2.18}/\epsilon^3)$ or $\tilde{O}(n^{2.33}/\epsilon^3)$ if we do not use fast matrix multiplication. As p approaches 0, our runtime approaches $\tilde{O}(n^\omega)$ which is natural, as $p = 0$ gives the matrix rank, which seems difficult to determine. As p approaches 2 the runtime smoothly approaches $\tilde{O}(n^2/\epsilon^3)$, which is then required for all $p \geq 2$.

Proof. The bound for $p \geq 2$ follows immediately from Theorem 30 with k set to 0. For $p < 2$ we have runtime:

$$\tilde{O}\left(n^{\omega(\log_n k)} + \frac{1}{p^3 \cdot \epsilon^{\max\{3, 1+1/p\}}} \cdot \left[\frac{n^{3/2+1/p}}{k^{1/p-1/2}}\right]\right)$$

We can partially optimize this term, ignoring p and ϵ factors for simplicity and setting k to equalize the coefficients on n . Our optimization is identical to the argument in Corollary 12 (set $k = n^{\frac{1/p-1/2}{1/p+1/2}}$ when not using fast matrix multiplication), giving the stated runtimes. $\square$

We now tackle the more complex case when $\mathbf{A}$ is sparse. We first consider $p \geq 2$.

Theorem 32 (Schatten p -norms, $p \geq 2$, for Sparse Matrices). *For any $p \geq 2$, and $\mathbf{A} \in \mathbb{R}^{n \times n}$ there is an algorithm returning $X \in (1 \pm \epsilon) \|\mathbf{A}\|_p^p$ with high probability in time:*

$$\tilde{O}\left(\frac{p}{\epsilon^3} \cdot \sqrt{\text{nnz}(\mathbf{A})nd_s(\mathbf{A})}\right) \text{ or } \tilde{O}\left(\frac{p}{\epsilon^3} \left[\text{nnz}(\mathbf{A})n^{\frac{1}{1+p}} + n^{1+\frac{2}{1+p}}\right]\right)$$

Note that if our matrix is uniformly sparse, $d_s(\mathbf{A}) = O(\text{nnz}(\mathbf{A})/n)$ and so the first runtime becomes $\tilde{O}(p \cdot \text{nnz}(\mathbf{A})/\epsilon^3)$. The second runtime can be better when $\mathbf{A}$ is not uniformly sparse.

Proof. The first runtime follows by setting $k = 0$ in Theorem 30. Note that $\text{nnz}(\mathbf{A}) \leq \sqrt{\text{nnz}(\mathbf{A})nd_s(\mathbf{A})}$. For the second, we use the second runtime of Theorem 30 which gives

$$\tilde{O}\left(\text{nnz}(\mathbf{A})k + nk^{\omega-1} + \frac{p}{\epsilon^3} \cdot \left[(\text{nnz}(\mathbf{A}) + nk)(n/k)^{1/p}\right]\right)$$

Setting $k = \left(\frac{n}{p}\right)^{1/p}$ to balance the coefficients on the $\text{nnz}(\mathbf{A})$ terms gives $k = n^{\frac{1}{1+p}}$ which yields the result. Note that the $nk^{\omega-1}$ term is dominated by the cost of the regressions, even if we do not use fast matrix multiplication this term will be nk^2 . $\square$

Finally, we consider the most complex cost, $p \leq 2$ for sparse matrices. We have:

Theorem 33 (Schatten p -norms, $p \leq 2$, for Sparse Matrices). *For any $p \in (0, 2]$, and $\mathbf{A} \in \mathbb{R}^{n \times n}$ there is an algorithm returning $X \in (1 \pm \epsilon) \|\mathbf{A}\|_p^p$ with high probability in time:*

$$\tilde{O}\left(\frac{1}{p^3 \cdot \epsilon^{\max\{3, 1/p\}}} \cdot \left[\text{nnz}(\mathbf{A})n^{\frac{1/p-1/2}{1/p+1/2}}\sqrt{\gamma_s} + \sqrt{\text{nnz}(\mathbf{A})} \cdot n^{\frac{4/p-1}{2/p+1}}\right]\right)$$

$$\text{or } \tilde{O}\left(\frac{1}{p^3 \cdot \epsilon^{\max\{3, 1/p\}}} \left[\text{nnz}(\mathbf{A})n^{\frac{1}{1+p}} + n^{1+\frac{2}{1+p}}\right]\right)$$

where $\gamma_s = \frac{d_s(\mathbf{A})n}{\text{nnz}(\mathbf{A})} \geq 1$.

Note that in the special case of $p = 1$ the first runtime gives $\tilde{O}\left(\frac{1}{\epsilon^3} \left[\text{nnz}(\mathbf{A}) n^{1/3} \sqrt{\gamma_s} + \sqrt{\text{nnz}(\mathbf{A}) n} \right]\right)$. The second term here is at worst n^2 , and could be much smaller for sparse $\mathbf{A}$.

Proof. For the second runtime, we use the second runtime of Theorem 30, balancing costs exactly as in Theorem 32 (setting $k = n^{\frac{1}{1+p}}$). For the first runtime, we consider the first runtime of Theorem 30 which gives:

$$\begin{aligned} & \tilde{O}\left(\text{nnz}(\mathbf{A})k + nk^{\omega-1} + \frac{1}{p^3 \cdot \epsilon^{\max\{3, 1/p\}}} \cdot \left[\text{nnz}(\mathbf{A}) + \sqrt{\text{nnz}(\mathbf{A})[n \cdot d_s(\mathbf{A}) + nk]} \cdot (n/k)^{1/p-1/2} \right]\right) \\ & \tilde{O}\left(\frac{nd_s(\mathbf{A})k}{\gamma_s} + nk^{\omega-1} + \frac{1}{p^3 \cdot \epsilon^{\max\{3, 1/p\}}} \cdot \left[\frac{nd_s(\mathbf{A}) + n\sqrt{d_s(\mathbf{A})}\sqrt{k}}{\sqrt{\gamma_s}} \right] \cdot \left(\frac{n}{k}\right)^{1/p-1/2}\right) \end{aligned}$$

Ignoring ϵ , p , and γ_s dependence, one way we can balance the costs is by setting:

$$nd_s(\mathbf{A})k = nd_s(\mathbf{A}) \left(\frac{n}{k}\right)^{1/p-1/2}$$

which gives $k = n^{\frac{1/p-1/2}{1/p+1/2}}$ and final runtime:

$$\begin{aligned} & \tilde{O}\left(\frac{1}{p^3 \cdot \epsilon^{\max\{3, 1/p\}}} \cdot \left[\text{nnz}(\mathbf{A}) n^{\frac{1/p-1/2}{1/p+1/2}} \sqrt{\gamma_s} + n\sqrt{d_s(\mathbf{A})/\gamma_s} \cdot k^{3/2} \right]\right) = \\ & \tilde{O}\left(\frac{1}{p^3 \cdot \epsilon^{\max\{3, 1/p\}}} \cdot \left[\text{nnz}(\mathbf{A}) n^{\frac{1/p-1/2}{1/p+1/2}} \sqrt{\gamma_s} + n\sqrt{d_s(\mathbf{A})/\gamma_s} \cdot n^{\frac{3/p-3/2}{2/p+1}} \right]\right). \end{aligned}$$

Note that we drop the $nk^{\omega-1}$ term as it is dominated by the last term with $k^{3/2}$ in it. This gives our final result by noting that $d_s(\mathbf{A})/\gamma_s = \text{nnz}(\mathbf{A})/n$, the average row sparsity. $\square$

7.2 Constant Factor Approximation without Row-Sparsity Dependence

Theorem 34 (Removing the Uniform Sparsity Constraint). *Let $\gamma \in (0, 1)$, $p \geq 1$, and $\mathbf{A} \in \mathbb{R}^{n \times n}$. There is an algorithm returning X with $\|\mathbf{A}\|_p \leq X = O(1/\gamma)\|\mathbf{A}\|_p$ with high probability in time $\tilde{O}(p \text{nnz}(\mathbf{A})n^\gamma)$ for $p \geq 2$, and in time $\tilde{O}\left(\frac{1}{p^3} \left[\text{nnz}(\mathbf{A}) n^{\frac{1/p-1/2}{1/p+1/2} + \gamma/2} + \sqrt{\text{nnz}(\mathbf{A})} \cdot n^{\frac{4/p-1}{2/p+1}} \right]\right)$ for $p < 2$.*

Proof. For $i \in \{1, 2, \dots, \lceil \frac{1}{\gamma} \rceil\}$, let $\mathbf{A}^i \in \mathbb{R}^{n \times n}$ be the matrix whose rows agree with that of $\mathbf{A}$ provided the number of non-zero entries in these rows are in the interval $[n^{\gamma i \cdot \frac{\text{nnz}(\mathbf{A})}{n}}, n^{\gamma(i+1) \cdot \frac{\text{nnz}(\mathbf{A})}{n}})$, while the remaining rows of $\mathbf{A}^i$ are set to 0^n . Let $\mathbf{A}^0 = \mathbf{A} - \sum_{i=1}^{\lceil \frac{1}{\gamma} \rceil} \mathbf{A}^i$. Since $p \geq 1$ we can apply the triangle inequality,

$$\|\mathbf{A}\|_p \leq \sum_{i=0}^{\lceil \frac{1}{\gamma} \rceil} \|\mathbf{A}^i\|_p \leq (\lceil \frac{1}{\gamma} \rceil + 1) \max_i \|\mathbf{A}^i\|_p. \quad (22)$$

For any matrix $\mathbf{B}$ obtained from a matrix $\mathbf{A}$ by replacing some rows of $\mathbf{A}$ with 0^n and preserving the remaining rows, $\|\mathbf{B}\|_p \leq \|\mathbf{A}\|_p$. This follows from the fact that $\|\mathbf{B}x\|_2 \leq \|\mathbf{A}x\|_2$ for all vectors x , together with the min-max theorem for singular values. Hence,

$$\|\mathbf{A}\|_p \geq \max_i \|\mathbf{A}^i\|_p. \quad (23)$$

Combining (22) and (23), an $O(1)$ -approximation to $\max_i \|\mathbf{A}^i\|_p$ is an $O(\frac{1}{\gamma})$ -approximation to $\|\mathbf{A}\|_p$.

We remove the zero rows from the $\mathbf{A}^i$, obtaining corresponding matrices $\mathbf{B}^i$. By definition of $\mathbf{A}^i$, the number of rows in $\mathbf{B}^i$ is at most $n^{1-\gamma^i}$. For each $\mathbf{B}^i$, which is an $s_i \times n$ matrix for $s_i \leq n^{1-\gamma^i}$, we can right-multiply it by an $n \times t_i$ OSNAP matrix $\mathbf{T}^i$, with $t_i = O(n^{1-\gamma^i} \log n)$ columns and $O(\log n)$ non-zero entries per column, so that $\|\mathbf{B}^i \mathbf{T}^i\|_p = (1 \pm 1/2) \|\mathbf{B}^i\|_p$ for all i with probability $1 - 1/\text{poly}(n)$, see [Coh16] (the fact that $\|\mathbf{B}^i \mathbf{T}^i\|_p = (1 \pm 1/2) \|\mathbf{B}^i\|_p$ follows from the fact that $\mathbf{T}^i$ is a subspace embedding of the row space of $\mathbf{B}^i$ together with Lemma C.2 of [LNW14]). The time to compute $\mathbf{C}^i = \mathbf{B}^i \mathbf{T}^i$ is $\tilde{O}(\text{nnz}(\mathbf{A}))$. Since $\mathbf{T}^i$ has $O(\log n)$ non-zero entries per row, each row of $\mathbf{C}^i$ has a number of non-zero entries within an $O(\log n)$ factor of the corresponding row of $\mathbf{B}^i$.

We compute $O(1)$ -approximations to the $\|\mathbf{C}^i\|_p$ for each i , and then take the maximum over i . To do so, for $p \geq 2$ we run the algorithm of Theorem 32 on each $\mathbf{C}^i$, while for $p < 2$ we run the algorithm of Theorem 33.

Let d_s^i denote the parameter d_s when run on $\mathbf{C}^i$. Then $d_s^i = O(n^{\gamma^i + \gamma} \frac{\text{nnz}(\mathbf{A})}{n} \log n)$. We can assume $\mathbf{C}^i$ is a square $t_i \times t_i$ matrix by padding with zeros, where $t_i = O(n^{1-\gamma^i} \log n)$. Hence, $t_i d_s^i = O(\text{nnz}(\mathbf{A}) n^\gamma \log n)$. Also, $\text{nnz}(\mathbf{C}^i) = O(\text{nnz}(\mathbf{B}^i) \log n) = O(\text{nnz}(\mathbf{A}) \log n)$.

The total time to apply Theorem 32 for $p \geq 2$ across all $\mathbf{C}^i$ is $\tilde{O}(p \text{nnz}(\mathbf{A}) n^\gamma)$. The total time to apply Theorem 33 for $p < 2$ across all $\mathbf{C}^i$ is $\tilde{O}\left(\frac{1}{p^3} \left[\text{nnz}(\mathbf{A}) n^{\frac{1/p-1/2}{1/p+1/2} + \gamma/2} + \sqrt{\text{nnz}(\mathbf{A})} \cdot n^{\frac{4/p-1}{2/p+1}} \right]\right)$, using that the γ_s of that theorem is $O(n^\gamma \log n)$. $\square$

7.3 SVD Entropy

In this section, we will show how to approximately estimate the SVD entropy of a matrix $\mathbf{A}$ assuming its condition number $K \stackrel{\text{def}}{=} \sigma_1(\mathbf{A})/\sigma_n(\mathbf{A})$ to be bounded by n^{ϵ_1} . Recall that the SVD entropy of a matrix $\mathbf{A}$ is given by $H(\mathbf{B}) \stackrel{\text{def}}{=} -\sum_i \sigma_i(\mathbf{B}) \log \sigma_i(\mathbf{B})$, where $\mathbf{B} \stackrel{\text{def}}{=} \frac{\mathbf{A}}{\|\mathbf{A}\|_1}$ is the matrix $\mathbf{A}$ scaled inversely by its Schatten 1-norm. Using the results in Section 7.1, we obtain a $(1 + \epsilon)$ approximation Z to $\|\mathbf{A}\|_1$ i.e., $\|\mathbf{A}\|_1 \leq Z \leq (1 + \epsilon) \|\mathbf{A}\|_1$. Define $\tilde{\mathbf{B}} \stackrel{\text{def}}{=} \frac{\mathbf{A}}{Z}$. This means that

$$\begin{aligned} |H(\tilde{\mathbf{B}}) - H(\mathbf{B})| &= \left| \sum_i \left(\frac{\sigma_i(\mathbf{A})}{\|\mathbf{A}\|_1} \log \frac{\sigma_i(\mathbf{A})}{\|\mathbf{A}\|_1} - \frac{\sigma_i(\mathbf{A})}{Z} \log \frac{\sigma_i(\mathbf{A})}{Z} \right) \right| \\ &\leq \left| \sum_i \sigma_i(\mathbf{A}) \left(\frac{1}{\|\mathbf{A}\|_1} - \frac{1}{Z} \right) \log \frac{\sigma_i(\mathbf{A})}{\|\mathbf{A}\|_1} \right| + \left| \sum_i \frac{\sigma_i(\mathbf{A})}{Z} \log \frac{\|\mathbf{A}\|_1}{Z} \right| \\ &\leq \frac{\epsilon}{1 + \epsilon} \left| \sum_i \frac{\sigma_i(\mathbf{A})}{\|\mathbf{A}\|_1} \log \frac{\sigma_i(\mathbf{A})}{\|\mathbf{A}\|_1} \right| + \epsilon \leq 2\epsilon \log n. \end{aligned}$$

The following theorem gives our main result for SVD entropy.

Theorem 35. *Given any PSD $\mathbf{A} \in \mathbb{R}^{n \times n}$ and error parameter $\epsilon \in (0, 1)$. Then, there exists an algorithm that computes the approximate SVD entropy of $\mathbf{A}$ and outputs $\hat{S}$, such that, with probability 9/10*

$$\hat{S} \in (1 \pm O(\epsilon)) H(\mathbf{B}), \quad (24)$$

where $\mathbf{B} = \frac{\mathbf{A}}{\|\mathbf{A}\|_1}$. The runtime of the algorithm will be

$$\tilde{O} \left(\frac{\text{nnz}(\mathbf{A}) n^{\frac{1}{3}} + n^{\frac{3}{2}} \sqrt{d_s}}{\tilde{\epsilon}^6} \log(1/\epsilon) \right),$$

Algorithm 2 SVD entropy estimation via multi-point interpolation.

Input: $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\epsilon \in (0, 1)$.

Output: Approximate SVD entropy $\hat{S}$.

Compute $\tilde{\mathbf{B}} = \frac{\mathbf{A}}{Z}$, where Z is a $(1 + \epsilon)$ approximation to $\|\mathbf{A}\|_1$.

Choose k_1 points $\alpha_1, \dots, \alpha_{k_1}$ as in (25), and set $\tilde{\epsilon} = \epsilon / (12c_1(k_1 + 1)^3 \log n)$.

for $i = 1 : k_1$ **do**

 Compute $Z_{1+\alpha_i}$, a $(1 + \tilde{\epsilon})$ -approximation for Schatten norm $\|\tilde{\mathbf{B}}\|_{1+\alpha_i}$.

 Compute $\tilde{T}_{1+\alpha_i}^{SVD}(\tilde{\mathbf{B}}) = (1 - Z_{1+\alpha_i}/Z^{1+\alpha_i})/\alpha_i$.

end for

Return $\hat{S}$ an estimate of $T(0)$ by interpolation using the points $\tilde{T}_{1+\alpha_i}^{SVD}(\tilde{\mathbf{B}})$.

where $d_s(\mathbf{A}) = O(\text{nnz}(\mathbf{A})/n)$, and $\tilde{\epsilon} = O(\frac{\epsilon}{\log n})$.

The algorithm we consider to estimate the SVD entropy of $\tilde{\mathbf{B}}$ follows the techniques of [HNO08]. The algorithm depends on the β^{th} SVD Tsallis entropy $T_\beta^{SVD}(\tilde{\mathbf{B}})$ of matrix $\tilde{\mathbf{B}}$, which is given by

$$T_\beta^{SVD}(\tilde{\mathbf{B}}) = \frac{1 - \|\tilde{\mathbf{B}}\|_\beta^\beta}{\beta - 1}.$$

We can describe the *multi-point interpolation* method for SVD entropy estimation with additive and multiplicative error approximations using [HNO08, Algorithm 1], see Algorithm 2.

Here, given an error parameter ϵ , we compute an approximate SVD Tsallis entropy $\tilde{T}_{1+\alpha_i}^{SVD}$ at k_1 different points $\{1 + \alpha_0, \dots, 1 + \alpha_{k_1}\}$, where $k_1 \stackrel{\text{def}}{=} \log \frac{1}{\epsilon} + \log c_1 + \log \log n$. $\alpha_1, \dots, \alpha_{k_1}$ are defined as follows. Let $\ell \stackrel{\text{def}}{=} 1/(2c_1(k_1 + 1) \log n)$ and let $g(\cdot)$ be defined as:

$$g(y) = \frac{k_1^2 \ell y - \ell(k_1^2 + 1)}{2k_1^2 + 1}, \text{ then, } \alpha_i \stackrel{\text{def}}{=} g(\cos(i\pi/k_1)). \quad (25)$$

Now define an error parameter $\tilde{\epsilon} \stackrel{\text{def}}{=} \epsilon / (12c_1(k_1 + 1)^3 \log n)$.

Algorithm 2 gives the stepwise algorithm. The runtime of the algorithm is dominated by the cost of computing $Z_{1+\alpha_i}$, a $(1 + \tilde{\epsilon})$ -approximation for Schatten norm $\|\tilde{\mathbf{B}}\|_{1+\alpha_i}$, particularly the smallest power Schatten norm. We get the smallest power when $\alpha_i = \frac{-1}{2c_1(k_1 + 1) \log n}$. We have $k_1 = \log(1/\epsilon)$ such Schatten norms to be estimated. Thus, we obtain the runtime in Theorem 35 by using $p = \alpha = O\left(1 - \frac{1}{\log(n/\epsilon)}\right)$ in the runtime for Schatten norm estimation given in Corollary 12, and observing that $n^{1/\log(n/\epsilon)}$ is between 1 and 2 (we set $p \approx 1$).

Additive approximation : Section 3.3.2 of [HNO08] tells us that if in Algorithm 2, the approximate $\tilde{T}_{1+\alpha_i}^{SVD}(\tilde{\mathbf{B}})$ is an additive $\tilde{\epsilon}$ approximation to $T_{1+\alpha_i}^{SVD}(\tilde{\mathbf{B}})$ for every $i \in [k_1]$, then we can use these to compute an additive- ϵ approximation to $H(\tilde{\mathbf{B}})$. Since $\frac{-1}{2c_1 k_1 \log n} < \alpha_i < \frac{-1}{16c_1 k_1^3 \log n}$, and since $\|\tilde{\mathbf{B}}\|_1 = 1$, obtaining $\frac{1}{16c_1 k_1^3 \log n}$ multiplicative approximation to $\|\tilde{\mathbf{B}}\|_{1+\alpha_{k_1}}$ suffices. This approximation can be obtained using the results of Section 7.1.

Multiplicative approximation : Article [HNO08] further extends the multi-point interpolation method to achieve multiplicative approximation for Shannon entropy (equivalently for SVD entropy) using the following modifications. We set the number of interpolation points $k_1 = \max\{5, \log(1/\epsilon)\}$. Then, section 3.4 in [HNO08] shows that if $\tilde{T}_{1+\alpha_i}^{SVD}(\tilde{\mathbf{B}})$ (in Algorithm 2) for every $i \in [k_1]$ computed are to be $(1+\tilde{\epsilon})$ -multiplicative approximation to $T_{1+\alpha_i}^{SVD}(\mathbf{B})$ (where $\tilde{\epsilon}$ is as defined above), then we can achieve multiplicative approximation for $H(\mathbf{B})$ using the multi-point interpolation method with these $\tilde{T}_{1+\alpha_i}^{SVD}(\tilde{\mathbf{B}})$. So, we need to obtain $(1+\tilde{\epsilon})$ -relative error approximations to the $T_{1+\alpha_i}^{SVD}(\mathbf{B})$ at each α_i .

If the matrix has reasonable (large) SVD entropy (meaning the singular values are uniform and there is no one singular value that is very large), then we can obtain $(1 \pm \tilde{\epsilon})$ approximation to the $T_{1+\alpha_i}^{SVD}(\mathbf{B})$ by simply computing $(1 \pm \tilde{\epsilon})$ approximation to $\|\tilde{\mathbf{B}}\|_{1+\alpha_i}$ using results from section 7.1. However, if the matrix has very small entropy, i.e., we have one singular value with very large magnitude and remaining singular values are very small, then achieving a multiplicative approximation will be difficult. This is because, we are approximating $T_{1+\alpha_i}^{SVD}$ of $\tilde{\mathbf{B}}$, which is matrix $\mathbf{A}$ scaled by a $(1 \pm \epsilon)$ approximation of its Schatten 1-norm. The $(1+\tilde{\epsilon})$ approximation obtained for $T_{1+\alpha_i}^{SVD}(\tilde{\mathbf{B}})$ might not be close to $T_{1+\alpha_i}^{SVD}(\mathbf{B})$ in this case.

This issue is equivalent to the ‘heavy element’ issue (one of the entries in the vector is very large) discussed in [HNO08]. Hence, to overcome the above issue, we can follow the workaround proposed in [HNO08]. Specifically, Lemma 5.5 in [HNO08] shows that a $(1+\tilde{\epsilon})$ -approximation to $1 - \hat{\sigma}_1$ together with a $(1+\tilde{\epsilon})$ -approximation to $\sum_{j>1} \hat{\sigma}_j^{1+\alpha_i}$, where $\hat{\sigma}_1 \geq \hat{\sigma}_2 \geq \dots \geq \hat{\sigma}_n$ are the singular values of $\mathbf{B}$ with $\sum_i \hat{\sigma}_i = 1$, suffices to get a $(1+\tilde{\epsilon})$ -approximation to $T_{1+\alpha_i}^{SVD}(\mathbf{B})$ at α_i . It follows that we just need to obtain $(1+\tilde{\epsilon})$ -approximations for these latter two quantities for each α_i .

It can be shown that when the Krylov method [MM15] (or power method) is used to deflate the top singular vector, we have that any unitarily invariant norm of the tail (remaining part of the spectrum) is preserved, see Appendix C for the proof. This means we can get $(1+\tilde{\epsilon})$ -approximations to both $\|\mathbf{A}_{-1}\|_1$ and $(\|\mathbf{A}_{-1}\|_{1+\alpha_i})^{1+\alpha_i}$, where $\mathbf{A}_{-1}$ is matrix $\mathbf{A}$ with the top singular vector exactly deflated off, by running the algorithm presented in this paper. That is, get the Schatten 1 and Schatten $(1+\alpha_i)$ norms for the matrix $\mathbf{A}$ with the top singular vector deflated off. We can also approximate $\|\mathbf{A}\|_1$ up to $(1+\tilde{\epsilon})$ relative error using our algorithm. Then, the two quantities above $(1 - \hat{\sigma}_1$ and $\sum_{j>1} \hat{\sigma}_j^{1+\alpha_i})$ are exactly $\|\mathbf{A}_{-1}\|_1 / \|\mathbf{A}\|_1$ and $(\|\mathbf{A}_{-1}\|_{1+\alpha_i})^{1+\alpha_i} / \|\mathbf{A}\|_1$, respectively. Since we have relative $(1+\tilde{\epsilon})$ -approximations to the numerators and denominators of both these quantities, we obtain then up to $(1+\tilde{\epsilon})$ -relative error the quantities $1 - \hat{\sigma}_1$ and $\sum_{j>1} \hat{\sigma}_j^{1+\alpha_i}$, as needed to achieve a multiplicative approximation to the SVD entropy. Note that we do not need to compute $\|\mathbf{A}\|_1$ exactly here even when the matrix has very low SVD entropy. A similar argument can be seen in section 6 of [HNO08].

Acknowledgements

We thank Vladimir Braverman for pointing out an error in our original proof of Theorem 8, which has been corrected.

References

- [ABB00] Orly Alter, Patrick O Brown, and David Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National*

Academy of Sciences, 97(18):10101–10106, 2000.

- [AKR15] Alexandr Andoni, Robert Krauthgamer, and Ilya P. Razenshteyn. Sketching and embedding are equivalent for norms. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing (STOC)*, pages 479–488, 2015.
- [AT11] Haim Avron and Sivan Toledo. Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix. *Journal of the ACM*, 58(2):8, 2011.
- [AZ17] Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *Proceedings of the 49th Annual ACM Symposium on Theory of Computing (STOC)*, 2017.
- [AZL16] Zeyuan Allen-Zhu and Yuanzhi Li. Faster principal component regression via optimal polynomial approximation to $\text{sgn}(x)$. [arXiv:1608.04773](#), 2016.
- [BCF09] Constantine Bekas, Alessandro Curioni, and I Fedulova. Low cost high performance uncertainty quantification. In *Proceedings of the 2nd Workshop on High Performance Computational Finance*, page 8. ACM, 2009.
- [BCKY16] Vladimir Braverman, Stephen R Chestnut, Robert Krauthgamer, and Lin F Yang. Sketches for matrix norms: Faster, smaller and more general. [arXiv:1609.05885](#), 2016.
- [BDKZ15] Christos Boutsidis, Petros Drineas, Prabhanjan Kambadur, and Anastasios Zouzias. A randomized algorithm for approximating the log determinant of a symmetric positive definite matrix. [arXiv:1503.00374](#), 2015.
- [BP14] Monami Banerjee and Nikhil R Pal. Feature selection with SVD entropy: some modification and extension. *Information Sciences*, 264:118–134, 2014.
- [BS83] Walter Baur and Volker Strassen. The complexity of partial derivatives. *Theoretical computer science*, 22(3):317–330, 1983.
- [Car14] Petre Caraiani. The predictive power of singular value decomposition entropy for stock market dynamics. *Physica A: Statistical Mechanics and its Applications*, 393:571–578, 2014.
- [CLM⁺15] Michael B. Cohen, Yin Tat Lee, Cameron Musco, Christopher Musco, Richard Peng, and Aaron Sidford. Uniform sampling for matrix approximation. In *Proceedings of the 6th Conference on Innovations in Theoretical Computer Science (ITCS)*, pages 181–190, 2015.
- [Coh16] Michael B. Cohen. Nearly tight oblivious subspace embeddings by trace inequalities. In *Proceedings of the 27th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 278–287, 2016.
- [CR12] Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119, 2012.
- [Dem13] James Demmel. An arithmetic complexity lower bound for computing rational functions, with applications to linear algebra. *submitted to SIMAX*, 2013.

- [DHJZ15] Haishun Du, Qingpu Hu, Manman Jiang, and Fan Zhang. Two-dimensional principal component analysis based on Schatten p-norm for image feature extraction. *Journal of Visual Communication and Image Representation*, 32:55–62, 2015.
- [DKJ⁺07] Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, pages 209–216, 2007.
- [DNPS16] Edoardo Di Napoli, Eric Polizzi, and Yousef Saad. Efficient estimation of eigenvalue counts in an interval. *Numerical Linear Algebra with Applications*, 2016.
- [DTV11] Amit Deshpande, Madhur Tulsiani, and Nisheeth K. Vishnoi. Algorithms and hardness for subspace approximation. In *Proceedings of the 22nd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 482–496, 2011.
- [FGKS15] Roy Frostig, Rong Ge, Sham M Kakade, and Aaron Sidford. Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.
- [FHT08] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.
- [FMMS16] Roy Frostig, Cameron Musco, Christopher Musco, and Aaron Sidford. Principal component projection without principal component analysis. 2016.
- [FORF16] JK Fitzsimons, MA Osborne, SJ Roberts, and JF Fitzsimons. Improved stochastic trace estimation using mutually unbiased bases. [arXiv:1608.00117](https://arxiv.org/abs/1608.00117), 2016.
- [GHJ⁺16] Dan Garber, Elad Hazan, Chi Jin, Sham M. Kakade, Cameron Musco, Praneeth Netrapalli, and Aaron Sidford. Faster eigenvector computation via shift-and-invert preconditioning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016. Full version at [arXiv:1605.08754](https://arxiv.org/abs/1605.08754).
- [GOSS16] Alon Gonen, Francesco Orabona, and Shai Shalev-Shwartz. Solving ridge regression using sketched preconditioned SVRG. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016.
- [GU17] François Le Gall and Florent Urrutia. Improved rectangular matrix multiplication using powers of the Coppersmith-Winograd tensor. [arXiv:1708.05622](https://arxiv.org/abs/1708.05622), 2017.
- [Gut92] Ivan Gutman. Total π -electron energy of benzenoid hydrocarbons. In *Advances in the Theory of Benzenoid Hydrocarbons II*, pages 29–63. Springer, 1992.
- [Gut01] Ivan Gutman. The energy of a graph: old and new results. In *Algebraic combinatorics and applications*, pages 196–211. Springer, 2001.
- [GVL12] Gene H. Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.
- [GXL15] Rongbao Gu, Wei Xiong, and Xinjie Li. Does the singular value decomposition entropy have predictive power for stock market? evidence from the shenzhen stock market. *Physica A: Statistical Mechanics and its Applications*, 439:103–113, 2015.

- [Hig08] Nicholas J Higham. *Functions of matrices: theory and computation*. SIAM, 2008.
- [HLM12] Moritz Hardt, Katrina Ligett, and Frank McSherry. A simple and practical algorithm for differentially private data release. In *Advances in Neural Information Processing Systems 25 (NIPS)*, pages 2348–2356. 2012.
- [HMAS17] Insu Han, Dmitry Malioutov, Haim Avron, and Jinwoo Shin. Approximating the spectral sums of large-scale matrices using Chebyshev approximations. *SIAM Journal on Scientific Computing*, 39(4), 2017.
- [HNO08] Nicholas JA Harvey, Jelani Nelson, and Krzysztof Onak. Sketching and streaming entropy via approximation theory. In *Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 489–498, 2008.
- [Hut90] Michael F Hutchinson. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 19(2):433–450, 1990.
- [Isa08] Garth Isaak. Fast matrix multiplication and inversion, math 242 notes. <http://www.lehigh.edu/~gi02/m242/08linstras.pdf>, 2008.
- [JNS13] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the 45th Annual ACM Symposium on Theory of Computing (STOC)*, pages 665–674, 2013.
- [Jol02] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.
- [JZ13] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems 26 (NIPS)*, pages 315–323, 2013.
- [KKS65] V. V. Klyuev and N. I. Kokovkin-Shcherbak. Minimization of the number of arithmetic operations in the solution of linear algebra systems of equations. 5(1):25-43, 1965. translated by H. F. Cleaves.
- [KM09] Jonathan A. Kelner and Aleksander Madry. Faster generation of random spanning trees. In *Proceedings of the 50th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 13–21, 2009.
- [KO17] Ashish Khetan and Sewoong Oh. Spectrum estimation from a few entries. [arXiv:1703.06327](https://arxiv.org/abs/1703.06327), 2017.
- [KS70] N. I. Kokovkin-Shcherbak. Minimization of numerical algorithms for solving arbitrary systems of linear equations. 22(4):494-502, 1970.
- [KS16] Rasmus Kyng and Sushant Sachdeva. Approximate gaussian elimination for Laplacians: Fast, sparse, and simple. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2016.
- [KV16] Weihao Kong and Gregory Valiant. Spectrum estimation from samples. [arXiv:1602.00061](https://arxiv.org/abs/1602.00061), 2016.

- [KW92] J Kuczyński and H Woźniakowski. Estimating the largest eigenvalue by the power and Lanczos algorithms with a random start. *SIAM Journal on Matrix analysis and Applications*, 13(4):1094–1122, 1992.
- [LG12] François Le Gall. Faster algorithms for rectangular matrix multiplication. In *Proceedings of the 53rd Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 514–523, 2012.
- [LM12] Chao Li and Gerome Miklau. Measuring the achievable error of query sets under differential privacy. [arXiv:1202.3399](https://arxiv.org/abs/1202.3399), 2012.
- [LMP13] Mu Li, Gary L. Miller, and Richard Peng. Iterative row sampling. In *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 127–136, 2013.
- [LN15] Yu Lu and Sahand N Negahban. Individualized rank aggregation using nuclear norm regularization. In *Communication, Control, and Computing (Allerton), 2015 53rd Annual Allerton Conference on*, pages 1473–1479. IEEE, 2015.
- [LNW14] Yi Li, Huy L. Nguyen, and David P. Woodruff. On sketching matrix norms and the top singular vector. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing (STOC)*, pages 1562–1581, 2014.
- [LS14] Yin Tat Lee and Aaron Sidford. Path finding methods for linear programming: Solving linear programs in $\tilde{o}(\text{vrnk})$ iterations and faster algorithms for maximum flow. In *Proceedings of the 55th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 424–433, 2014.
- [LSW15] Yin Tat Lee, Aaron Sidford, and Sam Chiu-wai Wong. A faster cutting plane method and its implications for combinatorial and convex optimization. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 1049–1065, 2015.
- [LSY16] Lin Lin, Yousef Saad, and Chao Yang. Approximating spectral densities of large matrices. *SIAM Review*, 58(1):34–65, 2016.
- [LW16a] Yi Li and David P. Woodruff. On approximating functions of the singular values in a stream. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing (STOC)*, 2016.
- [LW16b] Yi Li and David P. Woodruff. Tight bounds for sketching the operator norm, Schatten norms, and subspace embeddings. In *Proceedings of the 20th International Workshop on Randomization and Computation (RANDOM)*, 2016.
- [LW17] Yi Li and David P. Woodruff. Embeddings of Schatten norms with applications to data streams. [arXiv:1702.05626](https://arxiv.org/abs/1702.05626), 2017.
- [LYCG14] Lei Luo, Jian Yang, Jinhui Chen, and Yicheng Gao. Schatten p-norm based matrix regression model for image classification. In *Pattern Recognition*, pages 140–150. Springer, 2014.

- [MM15] Cameron Musco and Christopher Musco. Randomized block krylov methods for stronger and faster approximate singular value decomposition. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pages 1396–1404, 2015.
- [Mor73] Jacques Morgenstern. Note on a lower bound on the linear complexity of the fast Fourier transform. *Journal of the ACM (JACM)*, 20(2):305–306, 1973.
- [MST15] Aleksander Madry, Damian Straszak, and Jakub Tarnawski. Fast generation of random spanning trees and the effective resistance metric. In *Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2019–2036, 2015.
- [Nes13] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.
- [NHD12] Feiping Nie, Heng Huang, and Chris Ding. Low-rank matrix recovery via efficient Schatten p-norm minimization. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.
- [NNS⁺14] Praneeth Netrapalli, UN Niranjan, Sujay Sanghavi, Animashree Anandkumar, and Prateek Jain. Non-convex robust PCA. In *Advances in Neural Information Processing Systems 27 (NIPS)*, pages 1107–1115, 2014.
- [NWC⁺12] Feiping Nie, Hua Wang, Xiao Cai, Heng Huang, and Chibiao Ding. Robust matrix completion via joint Schatten p-norm and lp-norm minimization. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 566–574. IEEE, 2012.
- [Par98] Beresford N Parlett. *The symmetric eigenvalue problem*. SIAM, 1998.
- [Ras04] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pages 63–71. Springer, 2004.
- [Raz03] Ran Raz. On the complexity of matrix product. *SIAM Journal on Computing*, 32(5):1356–1369, 2003.
- [RKA15] Farbod Roosta-Khorasani and Uri Ascher. Improved bounds on sample size for implicit matrix trace estimators. *Foundations of Computational Mathematics*, 15(5):1187–1212, 2015.
- [RS03] Ran Raz and Amir Shpilka. Lower bounds for matrix product in bounded depth circuits with arbitrary gates. *SIAM Journal on Computing*, 32(2):488–513, 2003.
- [Saa03] Yousef Saad. *Iterative methods for sparse linear systems*. SIAM, 2003.
- [Saa11] Yousef Saad. *Numerical Methods for Large Eigenvalue Problems: Revised Edition*, volume 66. SIAM, 2011.
- [Shp03] Amir Shpilka. Lower bounds for matrix product. *SIAM Journal on Computing*, 32(5):1185–1200, 2003.
- [SLO13] Andreas Stathopoulos, Jesse Laeuchli, and Kostas Orginos. Hierarchical probing for estimating the trace of the matrix inverse on toroidal lattices. *SIAM Journal on Scientific Computing*, 35(5):S299–S322, 2013.

- [SS08] Daniel A. Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing (STOC)*, pages 563–568, 2008.
- [SSZ14] Shai Shalev-Shwartz and Tong Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 64–72, 2014.
- [ST04] Daniel A Spielman and Shang-Hua Teng. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing (STOC)*, pages 81–90, 2004.
- [TB97] Lloyd N. Trefethen and David Bau. *Numerical Linear Algebra*. SIAM, 1997.
- [UCS17] Shashanka Ubaru, Jie Chen, and Yousef Saad. Fast estimation of $\text{tr}(f(a))$ via stochastic Lanczos quadrature. 2017. manuscript.
- [US16] Shashanka Ubaru and Yousef Saad. Fast methods for estimating the numerical rank of large matrices. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 468–477, 2016.
- [USS17] Shashanka Ubaru, Yousef Saad, and Abd-Krim Seghouane. Fast estimation of approximate matrix ranks using spectral densities. *Neural Computation*, 2017.
- [VGLH06] Roy Varshavsky, Assaf Gottlieb, Michal Linial, and David Horn. Novel unsupervised feature filtering of biological data. *Bioinformatics*, 22(14):e507–e513, 2006.
- [Wil12] Virginia Vassilevska Williams. Multiplying matrices faster than Coppersmith-Winograd. In *Proceedings of the 44th Annual ACM Symposium on Theory of Computing (STOC)*, pages 887–898, 2012.
- [Wil15] Virginia Vassilevska Williams. Hardness of easy problems: Basing hardness on popular conjectures such as the strong exponential time hypothesis (invited talk). In *10th International Symposium on Parameterized and Exact Computation, IPEC 2015, September 16-18, 2015, Patras, Greece*, pages 17–29, 2015.
- [Win70] Shmuel Winograd. On the number of multiplications necessary to computer certain functions. 23:165-179, 1970.
- [Win87] Shmuel Winograd. Arithmetic complexity of computations. 1987.
- [WLK⁺16] Lingfei Wu, Jesse Laeuchli, Vassilis Kalantzis, Andreas Stathopoulos, and Efstratios Gallopoulos. Estimating the trace of the matrix inverse by interpolating from the diagonal of an approximate inverse. *Journal of Computational Physics*, 326:828–844, 2016.
- [Woo14] David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1-2):1–157, 2014.
- [WW10] Virginia Vassilevska Williams and Ryan Williams. Subcubic equivalences between path, matrix and triangle problems. In *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 645–654, 2010.

- [WWZ14] Karl Wimmer, Yi Wu, and Peng Zhang. Optimal query complexity for estimating the trace of a matrix. In *Proceedings of the 41st International Colloquium on Automata, Languages and Programming (ICALP)*, pages 1051–1062, 2014.
- [XGL⁺16] Yuan Xie, Shuhang Gu, Yan Liu, Wangmeng Zuo, Wensheng Zhang, and Lei Zhang. Weighted Schatten p -norm minimization for image denoising and background subtraction. *IEEE transactions on image processing*, 25(10):4842–4857, 2016.
- [XQT⁺16] Y. Xie, Y. Qu, D. Tao, W. Wu, Q. Yuan, and W. Zhang. Hyperspectral image restoration via iteratively regularized weighted Schatten p -norm minimization. *IEEE Transactions on Geoscience and Remote Sensing*, PP(99):1–18, 2016.
- [ZWJ15] Yuchen Zhang, Martin J Wainwright, and Michael I Jordan. Distributed estimation of generalized matrix rank: Efficient algorithms and lower bounds. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

A Linear System Solvers

In this section we give runtimes for solving ridge regression using both traditional iterative methods and stochastic gradient descent equipped with deflation-based preconditioning. We start with a few standard notions from convex optimization, which are necessary for our stochastic method bounds.

Definition 36 (Strong convexity). *A function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is μ -strongly convex if, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,*

$$f(\mathbf{x}) - f(\mathbf{y}) \geq \nabla f(\mathbf{y})^T (\mathbf{x} - \mathbf{y}) + \frac{\mu}{2} \|\mathbf{x} - \mathbf{y}\|_2^2.$$

Equivalently, if $\nabla^2 f \succeq \mu \mathbf{I}$.

Definition 37 (Smoothness). *A function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is β -smooth if for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,*

$$f(\mathbf{x}) - f(\mathbf{y}) \leq \nabla f(\mathbf{y})^T (\mathbf{x} - \mathbf{y}) + \frac{\beta}{2} \|\mathbf{x} - \mathbf{y}\|_2^2.$$

Equivalently, if $\nabla^2 f \preceq \beta \mathbf{I}$.

The rate of convergence that iterative methods achieve on f typically depends on the ratio β/μ , which corresponds to the condition number of a linear system. We next show how ridge regression can be recast as minimizing a convex function f , and show that our error from the true ridge solution is proportional to error in minimizing f .

Fact 38 (Ridge Regression as Convex Optimization). *For any $\mathbf{A} \in \mathbb{R}^{n \times d}$, $\mathbf{b} \in \mathbb{R}^d$, and $\lambda > 0$, let $\mathbf{M}_\lambda \stackrel{\text{def}}{=} \mathbf{A}^T \mathbf{A} + \lambda \mathbf{I}$ and $\mathbf{x}^* \stackrel{\text{def}}{=} \mathbf{M}_\lambda^{-1} \mathbf{b}$. $\mathbf{x}^*$ is the minimizer of the convex function:*

$$f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I}) \mathbf{x} - \mathbf{b}^T \mathbf{x}, \quad (26)$$

which has gradient $\nabla f(\mathbf{x}) = (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I}) \mathbf{x} - \mathbf{b}$. f is λ -strongly convex and $(\sigma_1(\mathbf{A})^2 + \lambda)$ -smooth.

Proof. To check that $\mathbf{x}^*$ is the minimizer, notice that:

$$\nabla f(\mathbf{x}^*) = (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{b} - \mathbf{b} = \mathbf{0}.$$

Since the function is quadratic, $\mathbf{x}^*$ is the unique minimizer. $\nabla^2 f = \mathbf{M}_\lambda$, so by Definitions 36 and 37, f is λ -strongly convex and $(\sigma_1(\mathbf{A})^2 + \lambda)$ -smooth. $\square$

Fact 39 (Function Error Equals Norm Error). *For any $\mathbf{x} \in \mathbb{R}^d$, letting $\mathbf{M}_\lambda$, $\mathbf{x}^*$ and f be defined as in Fact 38,*

$$\|\mathbf{x} - \mathbf{x}^*\|_{\mathbf{M}_\lambda}^2 = 2[f(\mathbf{x}) - f(\mathbf{x}^*)]. \quad (27)$$

Proof. Since $\mathbf{x}^* = \mathbf{M}_\lambda^{-1} \mathbf{b}$:

$$\begin{aligned} \|\mathbf{x} - \mathbf{x}^*\|_{\mathbf{M}_\lambda}^2 &\stackrel{\text{def}}{=} (\mathbf{x} - \mathbf{x}^*)^T \mathbf{M}_\lambda (\mathbf{x} - \mathbf{x}^*) \\ &= \mathbf{x}^T \mathbf{M}_\lambda \mathbf{x} - 2\mathbf{x}^T \mathbf{M}_\lambda \mathbf{x}^* + \mathbf{x}^{*T} \mathbf{M}_\lambda \mathbf{x}^* \\ &= \mathbf{x}^T \mathbf{M}_\lambda \mathbf{x} - 2\mathbf{x}^T \mathbf{b} + \mathbf{b}^T \mathbf{M}_\lambda^{-1} \mathbf{b} \\ &= 2[f(\mathbf{x}) - f(\mathbf{x}^*)]. \end{aligned}$$

$\square$

We will focus on making multiplicative progress in $[f(\mathbf{x}) - f(\mathbf{x}^*)]$ which will lead to multiplicative progress in $\|\mathbf{x} - \mathbf{x}^*\|_{\mathbf{M}}^2$ and a close approximation in $\log(1/\epsilon)$ iterations.

A.1 Unaccelerated SVRG

We first prove an unaccelerated and unpreconditioned runtime for the Stochastic Variance Reduced Gradient (SVRG) algorithm, introduced in [JZ13].

Theorem 40 (Standard SVRG Runtime). *For any $\mathbf{A} \in \mathbb{R}^{n \times d}$, $\mathbf{b} \in \mathbb{R}^d$, and $\lambda > 0$, let $\mathbf{M}_\lambda \stackrel{\text{def}}{=} \mathbf{A}^T \mathbf{A} + \lambda \mathbf{I}$ and $\mathbf{x}^* \stackrel{\text{def}}{=} \mathbf{M}_\lambda^{-1} \mathbf{y}$. There is an algorithm that returns $\mathbf{x}$ with: $\mathbb{E} \|\mathbf{x} - \mathbf{x}^*\|_{\mathbf{M}_\lambda} \leq \epsilon \|\mathbf{x}^*\|_{\mathbf{M}_\lambda}$ in*

$$O \left(\left[\text{nnz}(\mathbf{A}) + \frac{d_s(\mathbf{A}) \|\mathbf{A}\|_F^2}{\lambda} \right] \cdot \log(1/\epsilon) \right)$$

time.

SVRG proceeds in epochs. In each epoch we will make one full gradient computation – $\nabla f(\mathbf{x}_0) = \mathbf{M}_\lambda \mathbf{x}_0 - \mathbf{b}$ where $\mathbf{x}_0$ is the value of our iterate at the beginning of the epoch. We will then make a number of stochastic gradient updates each ‘variance reduced’ using $\nabla f(\mathbf{x}_0)$ and show that we make constant factor progress on our function value in expectation. Stringing together $\log(1/\epsilon)$ epochs yields Theorem 40. We write our function $f(\mathbf{x})$ given by (26) as a sum:

$$\begin{aligned} f(\mathbf{x}) &= \frac{1}{2} \mathbf{x}^T (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I}) \mathbf{x} - \mathbf{b}^T \mathbf{x} \\ &= \sum_{i=1}^n \frac{1}{2} \mathbf{x}^T \left(\mathbf{a}_i \mathbf{a}_i^T + \frac{\lambda \|\mathbf{a}_i\|_2^2}{\|\mathbf{A}\|_F^2} \mathbf{I} \right) \mathbf{x} - \frac{1}{n} \mathbf{b}^T \mathbf{x}. \end{aligned} \quad (28)$$

where $\mathbf{a}_i \in \mathbb{R}^d$ is the i^{th} row of $\mathbf{A}$ and $\mathbf{a}_i \mathbf{a}_i^T \in \mathbb{R}^{d \times d}$ is the rank-1 matrix which is its contribution to $\mathbf{A}^T \mathbf{A}$. We start with a well known lemma showing that it is possible to make constant progress on the value of $f(\mathbf{x})$ in each epoch of SVRG:

Lemma 41 (SVRG Epoch). *Consider a set of convex functions $\{\psi_1, \psi_2, \dots, \psi_n\}$ mapping $\mathbb{R}^d \rightarrow \mathbb{R}$. Let $f(\mathbf{x}) = \sum_{i=1}^n \psi_i(\mathbf{x})$ and $\mathbf{x}^* = \arg \min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$. Suppose we have a probability distribution p on $[1, 2, \dots, n]$ and that starting from some initial point $\mathbf{x}_0 \in \mathbb{R}^d$ in each iteration we select $i \in [1, \dots, n]$ with probability p_i and set:*

$$\mathbf{x}_{k+1} := \mathbf{x}_k - \frac{\eta}{p_i} (\nabla \psi_i(\mathbf{x}_k) - \nabla \psi_i(\mathbf{x}_0)) + \eta \nabla f(\mathbf{x}_0) \quad (29)$$

for some step size η . If f is μ -strongly convex and if for all $\mathbf{x} \in \mathbb{R}^d$ we have

$$\sum_{i=1}^n \frac{1}{p_i} \|\nabla \psi_i(\mathbf{x}) - \nabla \psi_i(\mathbf{x}^*)\|_2^2 \leq 2\bar{S} [f(\mathbf{x}) - f(\mathbf{x}^*)] \quad (30)$$

for some variance parameter $\bar{S}$ then for all $m \geq 1$ we have:

$$\mathbb{E} \left[\frac{1}{m} \sum_{k=1}^m f(\mathbf{x}_k) - f(\mathbf{x}^*) \right] \leq \frac{1}{1 - 2\eta\bar{S}} \left[\frac{1}{\mu\eta m} + 2\eta\bar{S} \right] \cdot [f(\mathbf{x}_0) - f(\mathbf{x}^*)]$$

Consequently, if we pick η to be a small multiple of $1/\bar{S}$, then for $m = O(\bar{S}/\mu)$ we will decrease the error by a constant multiplicative factor in expectation.

Proof. See for example Theorem 9 of [GHJ⁺16]. □

To apply the above Lemma we need the variance bound:

Lemma 42 (SVRG Variance Bound). *If $\psi_i(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T \left(\mathbf{a}_i \mathbf{a}_i^T + \frac{\lambda \|\mathbf{a}_i\|_2^2}{\|\mathbf{A}\|_F^2} \mathbf{I} \right) \mathbf{x} - \frac{1}{n} \mathbf{b}^T \mathbf{x}$ and $p_i = \frac{\|\mathbf{a}_i\|_2^2}{\|\mathbf{A}\|_F^2}$ then letting $\bar{S} = \|\mathbf{A}\|_F^2 + 2\lambda$ we have:*

$$\sum_{i=1}^n \frac{1}{p_i} \|\nabla \psi_i(\mathbf{x}) - \nabla \psi_i(\mathbf{x}^*)\|_2^2 \leq 2\bar{S} [f(\mathbf{x}) - f(\mathbf{x}^*)]$$

Proof. We have $\nabla \psi_i(\mathbf{x}) = \left(\mathbf{a}_i \mathbf{a}_i^T + \frac{\lambda \|\mathbf{a}_i\|_2^2}{\|\mathbf{A}\|_F^2} \mathbf{I} \right) \mathbf{x} - \frac{1}{n} \mathbf{b}$. Write $\mathbf{x} - \mathbf{x}^* = \mathbf{y}$ for simplicity of notation.

$$\begin{aligned} & \sum_{i=1}^n \frac{1}{p_i} \|\nabla \psi_i(\mathbf{x}) - \nabla \psi_i(\mathbf{x}^*)\|_2^2 \\ &= \sum_{i=1}^n \frac{\|\mathbf{A}\|_F^2}{\|\mathbf{a}_i\|_2^2} \cdot \left\| \left(\mathbf{a}_i \mathbf{a}_i^T + \frac{\lambda \|\mathbf{a}_i\|_2^2}{\|\mathbf{A}\|_F^2} \mathbf{I} \right) \mathbf{y} \right\|_2^2 \\ &= \|\mathbf{A}\|_F^2 \sum_{i=1}^n \left(\frac{1}{\|\mathbf{a}_i\|_2^2} \cdot \mathbf{y}^T \left((\mathbf{a}_i \mathbf{a}_i^T)^2 + 2 \frac{\lambda \|\mathbf{a}_i\|_2^2}{\|\mathbf{A}\|_F^2} \mathbf{a}_i \mathbf{a}_i^T + \left(\frac{\lambda \|\mathbf{a}_i\|_2^2}{\|\mathbf{A}\|_F^2} \right)^2 \mathbf{I} \right) \mathbf{y} \right) \\ &= \|\mathbf{A}\|_F^2 \cdot \left[\sum_{i=1}^n \left(\frac{\mathbf{y}^T (\mathbf{a}_i \mathbf{a}_i^T)^2 \mathbf{y}}{\|\mathbf{a}_i\|_2^2} \right) + 2 \sum_{i=1}^n \frac{\lambda \mathbf{y}^T \mathbf{a}_i \mathbf{a}_i^T \mathbf{y}}{\|\mathbf{A}\|_F^2} + \sum_{i=1}^n \left(\frac{\|\mathbf{y}\|_2^2 \lambda^2 \|\mathbf{a}_i\|_2^2}{\|\mathbf{A}\|_F^4} \right) \right] \\ &= \|\mathbf{A}\|_F^2 \cdot \left[\mathbf{y}^T \mathbf{A}^T \mathbf{A} \mathbf{y} + \frac{2\lambda \mathbf{y}^T \mathbf{A} \mathbf{A}^T \mathbf{y}}{\|\mathbf{A}\|_F^2} + \frac{\|\mathbf{y}\|_2^2 \lambda^2}{\|\mathbf{A}\|_F^2} \right] \\ &\leq \|\mathbf{A}\|_F^2 \left[\|\mathbf{y}\|_{\mathbf{M}_\lambda}^2 + \frac{2\lambda \|\mathbf{y}\|_{\mathbf{M}_\lambda}^2}{\|\mathbf{A}\|_F^2} \right] \\ &\leq 2(\|\mathbf{A}\|_F^2 + 2\lambda) [f(\mathbf{x}) - f(\mathbf{x}^*)] \end{aligned}$$

where the last step uses Fact 39. □

We can now plug this variance bound into Lemma 41 to prove Theorem 40

Proof of Theorem 40. Using Lemma 42 we can instantiate Lemma 41 with $\bar{S} = \|\mathbf{A}\|_F^2 + 2\lambda$. If we set

$$m = O\left(\frac{\bar{S}}{\mu}\right) = O\left(\frac{\|\mathbf{A}\|_F^2 + \lambda}{\lambda}\right) = O\left(\frac{\|\mathbf{A}\|_F^2}{\lambda}\right)$$

and $\eta = O\left(\frac{1}{\|\mathbf{A}\|_F^2 + \lambda}\right)$ (which we can compute explicitly) then after an m step SVRG epoch:

$$\mathbb{E} \left[\frac{1}{m} \sum_{k=1}^m f(\mathbf{x}_k) - f(\mathbf{x}^*) \right] \leq \frac{1}{2} [f(\mathbf{x}_0) - f(\mathbf{x}^*)].$$

If we choose k uniformly from $[1, \dots, m]$ this gives us $\mathbb{E}[f(\mathbf{x}_k) - f(\mathbf{x}^*)] \leq \frac{1}{2} [f(\mathbf{x}_0) - f(\mathbf{x}^*)]$. So stringing together $\log(1/\epsilon)$ of these epochs and letting $\mathbf{x}_0 = \mathbf{0}$ in the first epoch gives the theorem. Each epoch requires m stochastic gradient steps given by (29), taking $O(md_s) = O\left(\frac{\|d_s \mathbf{A}\|_F^2}{\lambda}\right)$ time plus $\text{nnz}(\mathbf{A})$ time to compute $\nabla f(\mathbf{x}_0)$, giving us the stated runtime.

Note that, naively, to produce $\mathbf{x}_{k+1}$ we need $O(d)$ time, not $O(d_s)$ time since our gradient step includes adding multiples of $\mathbf{b}$ and $\eta \nabla f(\mathbf{x}_0)$ both of which might be dense vectors in $\mathbb{R}^d$. However, in each epoch, we will just keep track of the coefficients of these two components in our iterate, allowing us to still compute (29) in $O(d_s)$ time. $\square$

A.2 Unaccelerated SVRG with Deflation-Based Preconditioning

If we assume that $d_s = O\left(\frac{\text{nnz}(\mathbf{A})}{n}\right)$ (i.e. our rows are uniformly sparse) we see that the runtime of Theorem 40 is dominated by $\frac{d_s(\mathbf{A})\|\mathbf{A}\|_F^2}{\lambda} = \text{nnz}(\mathbf{A}) \cdot \frac{\sum_{i=1}^d \sigma_i^2(\mathbf{A})}{n\lambda}$. $\bar{\kappa} \stackrel{\text{def}}{=} \frac{\sum_{i=1}^d \sigma_i^2(\mathbf{A})}{n\lambda}$ can be seen as an *average condition number*, which is always smaller than the condition number $\kappa \stackrel{\text{def}}{=} \frac{\sigma_1^2}{\lambda}$. This average condition number dependence means that SVRG can significantly outperform traditional iterative solvers that require a number of iterations depending on κ . However, this advantage is limited if $\|\mathbf{A}\|_F^2$ is very concentrated in a few large singular values, and hence $\bar{\kappa} \approx \kappa$. We can perform better in such situations by deflating off these large singular values and preconditioning with our deflated matrix, significantly ‘flattening’ the spectrum of $\mathbf{A}$. This method was used in [GOSS16], and we follow their approach closely, giving our own proof for completeness and so that we can express runtimes in terms of the necessary parameters for our downstream results. In particular, we show that the preconditioning methods of [GOSS16] can be implemented efficiently for sparse systems.

Theorem 43 (Preconditioned SVRG Runtime). *For any $\mathbf{A} \in \mathbb{R}^{n \times d}$, $\mathbf{b} \in \mathbb{R}^d$, and $\lambda > 0$, let $\mathbf{M}_\lambda \stackrel{\text{def}}{=} \mathbf{A}^T \mathbf{A} + \lambda \mathbf{I}$. Let $\bar{\kappa} \stackrel{\text{def}}{=} \frac{k\sigma_k^2(\mathbf{A}) + \sum_{i=k+1}^d \sigma_i^2(\mathbf{A})}{d\lambda}$ where $k \in [0, \dots, d]$ is an input parameter. There is an algorithm that uses $O(\text{nnz}(\mathbf{A})k \log d + dk^{\omega-1})$ precomputation time for sparse matrices or $O(nd^{\omega(1,1,\log_d k)-1} \log d)$ for dense matrices after which, given any input $\mathbf{y} \in \mathbb{R}^d$, letting $\mathbf{x}^* \stackrel{\text{def}}{=} \mathbf{M}_\lambda^{-1} \mathbf{y}$ the algorithm returns $\mathbf{x}$ with $\mathbb{E} \|\mathbf{x} - \mathbf{x}^*\|_{\mathbf{M}_\lambda} \leq \epsilon \|\mathbf{x}^*\|_{\mathbf{M}_\lambda}$ in*

$$O(\text{nnz}(\mathbf{A}) \log(1/\epsilon) + \log(1/\epsilon) [d \cdot d_s(\mathbf{A}) + dk] \bar{\kappa})$$

time for sparse $\mathbf{A}$ or

$$O(\log(1/\epsilon)(nd + d^2 \lceil \bar{\kappa} \rceil))$$

time for dense $\mathbf{A}$.

Our ideal algorithm would compute the top k singular vectors of $\mathbf{A}$ and deflate these off our matrix to flatten the spectrum. However, for efficiency we will instead compute approximate singular vectors, using an iterative method, like simultaneous iteration or a block Krylov iteration. These algorithms give the following guarantee [MM15]:

Lemma 44. *There is an algorithm that, with high probability in $O(\text{nnz}(\mathbf{A})k \log d + dk^{\omega-1})$ time returns $\mathbf{Z} \in \mathbb{R}^{d \times k}$ such that $\|\mathbf{A} - \mathbf{Z}\mathbf{Z}^T \mathbf{A}\|_2^2 \leq 2\sigma_{k+1}^2(\mathbf{A})$ and, for all $i \leq k$, letting $\tilde{\sigma}_i^2 \stackrel{\text{def}}{=} \mathbf{z}_i^T \mathbf{A}^T \mathbf{A} \mathbf{z}_i$, $|\tilde{\sigma}_i - \sigma_i^2(\mathbf{A})| \leq 2\sigma_{k+1}^2(\mathbf{A})$. For dense inputs the runtime can be sped up to $O(nd^{\omega(1,1,\log_d k)-1} \log d)$ by applying fast matrix multiplication at each iteration.⁴*

[GOSS16] shows that we can build a good preconditioner from such a $\mathbf{Z}$. Specifically:

⁴Each iteration requires $nd^{\omega(1,1,\log_d k)-1} \geq d^{\omega(1,1,\log_d k)}$ time since we assume $n \geq d$, which dominates the $O(dk^{\omega-1})$ time required to orthogonalize the approximate singular directions.

Lemma 45 (Theorem 5 of [GOSS16]). *For any $\mathbf{A} \in \mathbb{R}^{n \times d}$, $\mathbf{b} \in \mathbb{R}^d$, and $\lambda > 0$, given $\mathbf{Z} \in \mathbb{R}^{d \times k}$ satisfying the guarantees of Lemma 44, let*

$$\mathbf{P}^{-1/2} = \mathbf{Z} \tilde{\Sigma}^{-1/2} \mathbf{Z}^T + \frac{(\mathbf{I} - \mathbf{Z} \mathbf{Z}^T)}{\sqrt{\tilde{\sigma}_k^2 + \lambda}} \text{ where } \tilde{\Sigma}_{i,i}^{-1/2} = \frac{1}{\sqrt{\tilde{\sigma}_i^2 + \lambda}}.$$

Then we have:

$$\frac{\text{tr}(\mathbf{P}^{-1/2}(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I}) \mathbf{P}^{-1/2})}{\lambda_d(\mathbf{P}^{-1/2}(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I}) \mathbf{P}^{-1/2})} = O\left(\frac{k\sigma_k^2(\mathbf{A}) + \sum_{i=k+1}^d \sigma_i^2(\mathbf{A})}{\lambda} + d\right).$$

Intuitively, after applying the preconditioner, all top singular values are capped at $\sigma_k^2(\mathbf{A})$, giving a much flatter spectrum and better performance when optimizing with SVRG. To make use of the above bound, we first define a preconditioned ridge regression problem.

Definition 46 (Preconditioned Ridge Regression). *For any $\mathbf{A} \in \mathbb{R}^{n \times d}$, $\mathbf{b} \in \mathbb{R}^d$, and $\lambda > 0$, let $\mathbf{M}_\lambda \stackrel{\text{def}}{=} \mathbf{A}^T \mathbf{A} + \lambda \mathbf{I}$ and $\mathbf{x}^* \stackrel{\text{def}}{=} \mathbf{M}_\lambda^{-1} \mathbf{b}$. Letting $\mathbf{P}^{-1/2}$ be as described in Lemma 45, let $\hat{\mathbf{M}}_\lambda = \mathbf{P}^{-1/2} \mathbf{M}_\lambda \mathbf{P}^{-1/2}$ and define the preconditioned regression objective function by:*

$$\hat{f}(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T \hat{\mathbf{M}}_\lambda \mathbf{x} - \mathbf{b}^T \mathbf{P}^{-1/2} \mathbf{x}.$$

$\hat{f}$ is minimized at $\hat{\mathbf{x}}^* = \hat{\mathbf{M}}_\lambda^{-1} \mathbf{P}^{-1/2} \mathbf{b} = \mathbf{P}^{1/2}(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{b}$.

Fact 47 (Preconditioned Solution). *For any $\mathbf{x} \in \mathbb{R}^d$, if $\|\mathbf{x} - \hat{\mathbf{x}}^*\|_{\hat{\mathbf{M}}_\lambda} \leq \epsilon \|\hat{\mathbf{x}}^*\|_{\hat{\mathbf{M}}_\lambda}$ then $\|\mathbf{P}^{-1/2} \mathbf{x} - \mathbf{x}^*\|_{\mathbf{M}_\lambda} \leq \epsilon \|\mathbf{x}^*\|_{\mathbf{M}_\lambda}$ where $\mathbf{x}^*$ is the minimizer of $f(\mathbf{x})$.*

Proof. We have $\hat{\mathbf{x}}^* = \mathbf{P}^{1/2} \mathbf{x}^*$ and so using the fact that $\|\mathbf{x} - \hat{\mathbf{x}}^*\|_{\hat{\mathbf{M}}_\lambda} \leq \epsilon \|\hat{\mathbf{x}}^*\|_{\hat{\mathbf{M}}_\lambda}$ can write:

$$\begin{aligned} \|\mathbf{x} - \mathbf{P}^{1/2} \mathbf{x}^*\|_{\hat{\mathbf{M}}_\lambda} &\leq \epsilon \|\mathbf{P}^{1/2} \mathbf{x}^*\|_{\hat{\mathbf{M}}_\lambda} \\ \|\mathbf{P}^{-1/2} \mathbf{x} - \mathbf{x}^*\|_{\mathbf{M}_\lambda} &\leq \epsilon \|\mathbf{x}^*\|_{\mathbf{M}_\lambda}. \end{aligned}$$

□

By Fact 47, we can find a near optimal solution to our preconditioned system and by multiplying by $\mathbf{P}^{-1/2}$ obtain a near optimal solution to the original ridge regression problem. We now show that our preconditioned problem can be solved quickly. We first give a variance bound as we did in Lemma 42 in the non-preconditioned case.

Lemma 48 (Preconditioned Variance Bound). *Let $\mathbf{B}^T = [\mathbf{P}^{-1/2} \mathbf{A}^T, \sqrt{\lambda} \mathbf{P}^{-1/2}]$ so that $\mathbf{B}^T \mathbf{B} = \mathbf{P}^{-1/2}(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I}) \mathbf{P}^{-1/2} = \hat{\mathbf{M}}_\lambda$. For $i \in 1, \dots, (n+d)$, let*

$$\hat{\psi}_i(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T (\mathbf{b}_i \mathbf{b}_i^T) \mathbf{x} - \frac{1}{n+d} \mathbf{b}_i^T \mathbf{P}^{-1/2} \mathbf{x}$$

so that $\sum_{i=1}^{n+d} \hat{\psi}_i(\mathbf{x}) = \hat{f}(\mathbf{x})$. Set $p_i = \frac{\|\mathbf{b}_i\|_2^2}{\|\mathbf{B}\|_F^2}$ then letting $\bar{S} = \text{tr}(\mathbf{P}^{-1/2}(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I}) \mathbf{P}^{-1/2})$:

$$\sum_{i=1}^n \frac{1}{p_i} \left\| \nabla \hat{\psi}_i(\mathbf{x}) - \nabla \hat{\psi}_i(\mathbf{x}^*) \right\|_2^2 \leq 2\bar{S} [f(\mathbf{x}) - f(\mathbf{x}^*)].$$

Proof. Write $\mathbf{x} - \mathbf{x}^* = \mathbf{y}$ for simplicity of notation. Then we have:

$$\begin{aligned} \sum_{i=1}^{n+d} \frac{1}{p_i} \left\| \nabla \hat{\psi}_i(\mathbf{x}) - \nabla \hat{\psi}_i(\mathbf{x}^*) \right\|_2^2 &= \|\mathbf{B}\|_F^2 \sum_{i=1}^{n+d} \frac{\|\mathbf{b}_i \mathbf{b}_i^T \mathbf{y}\|_2^2}{\|\mathbf{b}_i\|_2^2} = \|\mathbf{B}\|_F^2 \sum_{i=1}^{n+d} \frac{\mathbf{y}^T \mathbf{b}_i (\mathbf{b}_i^T \mathbf{b}_i) \mathbf{b}_i^T \mathbf{y}}{\|\mathbf{b}_i\|_2^2} \\ &= \|\mathbf{B}\|_F^2 \cdot \mathbf{y}^T \mathbf{B}^T \mathbf{B} \mathbf{y} \\ &= 2 \|\mathbf{B}\|_F^2 \cdot [\hat{f}(\mathbf{x}) - \hat{f}(\mathbf{x}^*)] \end{aligned}$$

where the last step uses Fact 39. Finally we write:

$$\bar{S} = \|\mathbf{B}\|_F^2 = \text{tr}(\mathbf{B}^T \mathbf{B}) = \text{tr}(\mathbf{P}^{-1/2}(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I}) \mathbf{P}^{-1/2}).$$

□

We can now combine this variance bound with Lemmas 41 and 45 to prove our preconditioned SVRG runtime.

Proof of Theorem 43. Using Lemma 48 we can instantiate Lemma 41 with $\bar{S} = \|\mathbf{B}\|_F^2 = \text{tr}(\mathbf{P}^{-1/2}(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I}) \mathbf{P}^{-1/2})$. Letting $\mu = \lambda_d(\mathbf{P}^{-1/2}(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I}) \mathbf{P}^{-1/2})$ be a lower bound on the strong convexity of $\hat{f}$, by Lemma 45 we have:

$$\begin{aligned} m &= O\left(\frac{\bar{S}}{\mu}\right) = O\left(\frac{\text{tr}(\mathbf{P}^{-1/2}(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I}) \mathbf{P}^{-1/2})}{\lambda_d(\mathbf{P}^{-1/2}(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I}) \mathbf{P}^{-1/2})}\right) \\ &= O\left(\frac{k\sigma_k^2(\mathbf{A}) + \sum_{i=k+1}^d \sigma_i^2(\mathbf{A})}{\lambda} + d\right). \end{aligned}$$

If we set the step size $\eta = O\left(\frac{1}{\bar{S}}\right) = O\left(\frac{1}{\|\mathbf{B}\|_F^2}\right)$ which we will compute explicitly, then after an m step SVRG epoch running on the functions $\hat{\psi}_i(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T (\mathbf{b}_i \mathbf{b}_i^T) \mathbf{x} - \frac{1}{n+d} \mathbf{b}_i^T \mathbf{P}^{-1/2} \mathbf{x}$ each selected with probability $p_i = \frac{\|\mathbf{b}_i\|_2^2}{\|\mathbf{B}\|_F^2}$, we make constant progress in expectation on $\hat{f}(\mathbf{x})$. After $\log(1/\epsilon)$ iterations we have $\mathbf{x}$ with $\mathbb{E} \|\mathbf{x} - \hat{\mathbf{x}}^*\|_{\hat{\mathbf{M}}_\lambda} \leq \epsilon \|\hat{\mathbf{x}}^*\|_{\hat{\mathbf{M}}_\lambda}$ and hence, $\mathbb{E} \|\mathbf{P}^{-1/2} \mathbf{x} - \mathbf{x}^*\|_{\mathbf{M}_\lambda} \leq \epsilon \|\mathbf{x}^*\|_{\mathbf{M}_\lambda}$ by Fact 47. It remains to bound the cost of each SVRG epoch applied to $\hat{f}(\mathbf{x})$.

For dense inputs, the argument is relatively simple. We can compute the approximate top singular vector space $\mathbf{Z} \in \mathbb{R}^{d \times k}$ in $O(nd^{\omega(1,1,\log_d k)-1} \log d)$ time by Lemma 44. Using the factored structure of $\mathbf{P}^{-1/2}$, we can explicitly form $\mathbf{A} \mathbf{P}^{-1/2}$ and $\mathbf{P}^{-1/2}$ also in $(O(nd^{\omega(1,1,\log_d k)-1}))$ time. This allows us to compute a full gradient in $O(nd)$ time (which we do once per epoch) and perform each stochastic gradient step in $O(d)$ time (which we do m times per epoch). This gives final runtime

$$O((nd + md) \log(1/\epsilon)) = O(\log(1/\epsilon)(nd + d^2 \lceil \bar{\kappa} \rceil))$$

where $\bar{\kappa} \stackrel{\text{def}}{=} \frac{k\sigma_k^2(\mathbf{A}) + \sum_{i=k+1}^d \sigma_i^2(\mathbf{A})}{d\lambda}$.

For sparse $\mathbf{A}$ we have to be much more careful to fully exploit sparsity when preconditioning. Computing $\mathbf{Z}$ takes time $O(\text{nnz}(\mathbf{A})k \log d + dk^{\omega-1})$ by Lemma 44. We will not explicitly form $\mathbf{P}^{-1/2}$ but will show how to apply it when needed. Recall that $\mathbf{P}^{-1/2} = \mathbf{Z} \tilde{\Sigma}^{-1/2} \mathbf{Z}^T + \delta(\mathbf{I} - \mathbf{Z} \mathbf{Z}^T)$ where we denote $\delta \stackrel{\text{def}}{=} 1/(\tilde{\sigma}_i^2 + \lambda)$. First, in order to determine our step size η we must compute $\|\mathbf{B}\|_F^2$.

$$\begin{aligned} \|\mathbf{B}\|_F^2 &= \left\| \mathbf{A} \mathbf{P}^{-1/2} \right\|_F^2 + \lambda \left\| \mathbf{P}^{-1/2} \right\|_F^2 \\ &= \left\| \mathbf{A} \mathbf{Z} \tilde{\Sigma}^{-1/2} \mathbf{Z}^T \right\|_F^2 + \delta \left\| \mathbf{A}(\mathbf{I} - \mathbf{Z} \mathbf{Z}^T) \right\|_F^2 + \lambda \left\| \mathbf{Z} \tilde{\Sigma}^{-1/2} \mathbf{Z}^T \right\|_F^2 + \lambda \delta \left\| \mathbf{I} - \mathbf{Z} \mathbf{Z}^T \right\|_F^2. \end{aligned}$$

Applying Pythagorean theorem and submultiplicativity we have:

$$\|\mathbf{B}\|_F^2 = \|\mathbf{AZ}\tilde{\Sigma}^{-1/2}\|_F^2 + \delta \|\mathbf{A}\|_F^2 - \delta \|\mathbf{AZ}\|_F^2 + \lambda \|\mathbf{Z}\tilde{\Sigma}^{-1/2}\|_F^2 + \lambda \delta (d - k)$$

which can be computed in $O(\text{nnz}(\mathbf{A})k)$ time with just a single multiplication of $\mathbf{A}$ by $\mathbf{Z}$.

We can similarly compute our sampling probabilities $p_i = \frac{\|\mathbf{b}_i\|_2^2}{\mathbf{B}_F^2}$ quickly by noting that for $i = 1, \dots, n$, $\|\mathbf{b}_i\|_2^2 = \|\mathbf{a}_i^T \mathbf{Z} \tilde{\Sigma}^{-1/2}\|_2^2 + \delta \|\mathbf{a}_i\|_2^2 - \delta \|\mathbf{a}_i^T \mathbf{Z}\|_2^2$ and for $i = n + 1, \dots, n + d$, $\|\mathbf{b}_i\|_2^2 = \lambda \left(\|\mathbf{z}_i^T \tilde{\Sigma}^{-1/2}\|_2^2 + \delta (1 - \|\mathbf{z}_i\|_2^2) \right)$. This again requires just $O(\text{nnz}(\mathbf{A})k)$ time.

In each SVRG epoch we must compute one full gradient of the form:

$$\nabla \hat{f}(\mathbf{x}_0) = \mathbf{B}^T \mathbf{B} \mathbf{x}_0 - \mathbf{P}^{-1/2} \mathbf{b} = \mathbf{P}^{-1/2} \mathbf{A}^T \mathbf{A} \mathbf{P}^{-1/2} \mathbf{x}_0 + \lambda \mathbf{P}^{-1} \mathbf{x}_0 - \mathbf{P}^{-1/2} \mathbf{b}$$

which takes $O(\text{nnz}(\mathbf{A}) + dk)$ time as $\mathbf{P}^{-1/2}$ can be applied to a vector in $O(dk)$ time. Naively, we then must make m stochastic gradient steps, each requiring $O(dk)$ time. However we can do better. Instead of storing our iterate $\mathbf{x}_k$ explicitly, we will store it as the sum:

$$\mathbf{x}_k = \left(\mathbf{x}_k^{(0)} - \mathbf{Z} \mathbf{x}_k^{(1)} \right) + \mathbf{Z} \mathbf{x}_k^{(2)} + x_k^{(3)} \cdot \nabla \hat{f}(\mathbf{x}_0).$$

Note that $\mathbf{x}_k^{(0)} \in \mathbb{R}^d$, $\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)} \in \mathbb{R}^k$ and $x_k^{(3)}$ is a scalar. At the beginning of each epoch we will set $\mathbf{x}_0^{(0)} = \mathbf{x}_0$, $\mathbf{x}_0^{(1)} = \mathbf{Z}^T \mathbf{x}_0$, $\mathbf{x}_0^{(2)} = \mathbf{Z}^T \mathbf{x}_0$ and $x_0^{(3)} = 0$. We can compute $\mathbf{Z}^T \mathbf{x}_0$ in $O(dk)$ time. We will maintain the invariant that $\left(\mathbf{x}_k^{(0)} - \mathbf{Z} \mathbf{x}_k^{(1)} \right)$ is perpendicular to the span of $\mathbf{Z}$ while $\mathbf{x}_k^{(2)}$ will give the component of $\mathbf{x}_k$ within this span.

For ease of notation let $\mathbf{a}_{n+i}$ denote $\sqrt{\lambda} \mathbf{e}_i$ where $\mathbf{e}_i \in \mathbb{R}^d$ the i^{th} standard basis vector. In this way we have $\mathbf{b}_i = \mathbf{P}^{-1/2} \mathbf{a}_i$ for all $i \in 1, \dots, n + d$. Each stochastic gradient step is of the form:

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{x}_k - \frac{\eta}{p_i} \left(\nabla \hat{\psi}_i(\mathbf{x}_k) - \nabla \hat{\psi}_i(\mathbf{x}_0) \right) + \eta \nabla \hat{f}(\mathbf{x}_0) \\ &= \mathbf{x}_k + \frac{\eta}{p_i} \mathbf{P}^{-1/2} (\mathbf{a}_i \mathbf{a}_i^T) \mathbf{P}^{-1/2} (\mathbf{x}_k - \mathbf{x}_0) + \eta \nabla \hat{f}(\mathbf{x}_0) \\ &= \mathbf{x}_k + \frac{\eta}{p_i} (\mathbf{P}^{-1/2} \mathbf{a}_i) \left((\mathbf{a}_i^T \mathbf{P}^{-1/2}) \mathbf{x}_k - \mathbf{a}_i^T (\mathbf{P}^{-1/2} \mathbf{x}_0) \right) + \eta \nabla \hat{f}(\mathbf{x}_0). \end{aligned}$$

We can precompute $\mathbf{P}^{-1/2} \mathbf{x}_0$ in $O(dk)$ time per epoch and then can compute the dot product $\mathbf{a}_i^T (\mathbf{P}^{-1/2} \mathbf{x}_0)$ in $\text{nnz}(\mathbf{a}_i) = O(d_s(\mathbf{A}))$ time. Using the fact that $\left(\mathbf{x}_k^{(0)} - \mathbf{Z} \mathbf{x}_k^{(1)} \right)$ is always perpendicular to the span of $\mathbf{Z}$, we can write the dot product $(\mathbf{a}_i^T \mathbf{P}^{-1/2}) \mathbf{x}_k$ as:

$$\begin{aligned} \mathbf{a}_i^T \left(\mathbf{Z} \tilde{\Sigma}^{-1/2} \mathbf{Z}^T + \delta (\mathbf{I} - \mathbf{Z} \mathbf{Z}^T) \right) \left(\left(\mathbf{x}_k^{(0)} - \mathbf{Z} \mathbf{x}_k^{(1)} \right) + \mathbf{Z} \mathbf{x}_k^{(2)} + x_k^{(3)} \nabla \hat{f}(\mathbf{x}_0) \right) \\ = \delta \mathbf{a}_i^T \left(\mathbf{x}_k^{(0)} - \mathbf{Z} \mathbf{x}_k^{(1)} \right) + \mathbf{a}_i^T \mathbf{Z} \tilde{\Sigma}^{-1/2} \mathbf{x}_k^{(2)} + x_k^{(3)} \mathbf{a}_i^T \mathbf{Z} \tilde{\Sigma}^{-1/2} \mathbf{Z}^T \nabla \hat{f}(\mathbf{x}_0) + \delta x_k^{(3)} \mathbf{a}_i^T (\mathbf{I} - \mathbf{Z} \mathbf{Z}^T) \nabla \hat{f}(\mathbf{x}_0). \end{aligned}$$

We can precompute $\mathbf{AZ}$ in $O(\text{nnz}(\mathbf{A})k)$ time. We can also precompute $\mathbf{Z} \tilde{\Sigma}^{-1/2} \mathbf{Z}^T \nabla \hat{f}(\mathbf{x}_0)$ and $(\mathbf{I} - \mathbf{Z} \mathbf{Z}^T) \nabla \hat{f}(\mathbf{x}_0)$ in $O(dk)$ time. With these values in hand, computing the above dot product takes just $O(\text{nnz}(\mathbf{a}_i) + k) = O(d_s(\mathbf{A}) + k)$ time.

Now, we can write $\mathbf{P}^{-1/2} \mathbf{a}_i = \mathbf{Z} (\tilde{\Sigma} \mathbf{Z}^T \mathbf{a}_i) - \delta (\mathbf{I} - \mathbf{Z} \mathbf{Z}^T) \mathbf{a}_i$. Since we have precomputed $\mathbf{AZ}$, $\tilde{\Sigma} \mathbf{Z}^T \mathbf{a}_i$ can be computed and added to $\mathbf{x}_k^{(2)}$ with the appropriate weight in $O(k)$ time. We can then

compute for the appropriate weight w :

$$\begin{aligned} \left(\mathbf{x}_{k+1}^{(0)} - \mathbf{Z} \mathbf{x}_{k+1}^{(1)} \right) &= \left(\mathbf{x}_k^{(0)} - \mathbf{Z} \mathbf{x}_k^{(1)} \right) + w \cdot (\mathbf{I} - \mathbf{Z} \mathbf{Z}^T) \mathbf{a}_i \\ &= \left(\mathbf{x}_k^{(0)} + w \mathbf{a}_i \right) - \mathbf{Z} \left(\mathbf{x}_k^{(1)} + w \cdot \mathbf{Z}^T \mathbf{a}_i \right) \end{aligned}$$

which takes time $O(d_s(\mathbf{A}) + k)$. Finally, we set $x_{k+1}^{(3)} = x_k^{(3)} + \eta$. Overall, our runtime per epoch is $O(\text{nnz}(\mathbf{A}) + dk + m(d_s(\mathbf{A}) + k))$ and so our total runtime is:

$$O(\text{nnz}(\mathbf{A}) \log(1/\epsilon) + \log(1/\epsilon) [d \cdot d_s(\mathbf{A}) + dk] \bar{\kappa})$$

where $\bar{\kappa} \stackrel{\text{def}}{=} \frac{k\sigma_k^2(\mathbf{A}) + \sum_{i=k+1}^d \sigma_i^2(\mathbf{A})}{d\lambda}$. □

A.3 Accelerated and Preconditioned SVRG

We now combine the deflation-based preconditioning described above with accelerated gradient methods to give our strongest runtime bound for ridge regression using via stochastic solvers.

Theorem 49 (Accelerated Preconditioned SVRG Runtime). *For any $\mathbf{A} \in \mathbb{R}^{n \times d}$, $\mathbf{b} \in \mathbb{R}^d$, and $\lambda > 0$, let $\mathbf{M}_\lambda \stackrel{\text{def}}{=} \mathbf{A}^T \mathbf{A} + \lambda \mathbf{I}$ and $\bar{\kappa} \stackrel{\text{def}}{=} \frac{k\sigma_k^2(\mathbf{A}) + \sum_{i=k+1}^d \sigma_i^2(\mathbf{A})}{d\lambda}$ where $k \in [0, \dots, d]$ is an input parameter. There is an algorithm that uses $O(\text{nnz}(\mathbf{A})k \log d + dk\omega^{-1})$ precomputation time for sparse matrices or $O(nd^{\omega(1,1,\log_d k)-1} \log d)$ for dense matrices after which, given any input $\mathbf{y} \in \mathbb{R}^d$, letting $\mathbf{x}^* \stackrel{\text{def}}{=} \mathbf{M}_\lambda^{-1} \mathbf{y}$ the algorithm returns $\mathbf{x}$ with $\mathbb{E} \|\mathbf{x} - \mathbf{x}^*\|_{\mathbf{M}_\lambda} \leq \epsilon \|\mathbf{x}^*\|_{\mathbf{M}_\lambda}$ in*

$$O\left(\text{nnz}(\mathbf{A}) \log(1/\epsilon) + \log(1/\epsilon) \log(d\bar{\kappa}) \cdot \sqrt{\text{nnz}(\mathbf{A})[d \cdot d_s(\mathbf{A}) + dk] \bar{\kappa}}\right)$$

time for sparse $\mathbf{A}$ or

$$O\left(\log(1/\epsilon)(nd + n^{1/2} d^{3/2} \log(\bar{\kappa}) \sqrt{\bar{\kappa}})\right)$$

time for dense $\mathbf{A}$.

The above runtime will follow from applying a blackbox technique for accelerating the runtime of convex optimization methods. While there are a number of improvements over this method [AZ17], we at most lose logarithmic factors and gain significant simplicity in our proofs.

Lemma 50 (Theorem 1.1 of [FGKS15]). *Let $f(\mathbf{x})$ be μ -strongly convex and let $\mathbf{x}^* \stackrel{\text{def}}{=} \arg \min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$. For any $\gamma > 0$ and any $\mathbf{x}_0 \in \mathbb{R}^d$, let $f_{\gamma, \mathbf{x}_0}(\mathbf{x}) \stackrel{\text{def}}{=} f(\mathbf{x}) + \frac{\gamma}{2} \|\mathbf{x} - \mathbf{x}_0\|_2^2$. Let $\mathbf{x}_{\gamma, \mathbf{x}_0}^* \stackrel{\text{def}}{=} \arg \min_{\mathbf{x} \in \mathbb{R}^d} f_{\gamma, \mathbf{x}_0}(\mathbf{x})$. Suppose that, for all $\mathbf{x}_0 \in \mathbb{R}^d$, $c > 0$, $\gamma > 2\mu$, we can compute $\mathbf{x}_c$ such that*

$$\mathbb{E} [f_{\gamma, \mathbf{x}_0}(\mathbf{x}_c) - f_{\gamma, \mathbf{x}_0}(\mathbf{x}_{\gamma, \mathbf{x}_0}^*)] \leq \frac{1}{c} [f_{\gamma, \mathbf{x}_0} - f_{\gamma, \mathbf{x}_0}(\mathbf{x}_{\gamma, \mathbf{x}_0}^*)]$$

in time $\mathcal{T}_c$. Then we can compute x_1 such that $\mathbb{E} [f(\mathbf{x}_1) - f(\mathbf{x}^*)] \leq \frac{1}{c} [f(\mathbf{x}_0) - f(\mathbf{x}^*)]$ in time

$$O\left(\mathcal{T}_{4\left(\frac{2\gamma+\mu}{\mu}\right)^{3/2} \sqrt{\gamma/\mu} \log c}\right).$$

Proof of Theorem 49. We focus again on optimizing the preconditioned function $\hat{f}$ in Definition 46 as by Fact 47 a near optimal minimizer for this function yields a near optimal solution for our original ridge regression problem. We split $\hat{f}_{\gamma, \mathbf{x}_0} = \frac{1}{2} \mathbf{x}^T \hat{\mathbf{M}}_\lambda \mathbf{x} - \mathbf{b}^T \mathbf{P}^{-1/2} \mathbf{x} + \frac{\gamma}{2} \|\mathbf{x} - \mathbf{x}_0\|_2^2$ as $\hat{f}_{\gamma, \mathbf{x}_0} = \sum_{i=1}^{n+d} \hat{\psi}_{\gamma, \mathbf{x}_0}^i(\mathbf{x})$ where

$$\hat{\psi}_{\gamma, \mathbf{x}_0}^i(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T (\mathbf{b}_i \mathbf{b}_i^T) \mathbf{x} - \frac{1}{n+d} \mathbf{b}^T \mathbf{P}^{-1/2} \mathbf{x} + \frac{\gamma \|\mathbf{b}_i\|_2^2}{2 \|\mathbf{B}_F^2\|} \|\mathbf{x} - \mathbf{x}_0\|_2^2.$$

$\mathbf{B}$ is as defined in Lemma 48 with $\mathbf{B}^T = [\mathbf{P}^{-1/2} \mathbf{A}^T, \sqrt{\lambda} \mathbf{P}^{-1/2}]$.

We have $\nabla \hat{\psi}_{\gamma, \mathbf{x}_0}^i(\mathbf{x}) = (\mathbf{b}_i \mathbf{b}_i^T) \mathbf{x} - \frac{1}{n+d} \mathbf{b} \mathbf{P}^{-1/2} + \frac{\gamma \|\mathbf{b}_i\|_2^2}{\|\mathbf{B}\|_F^2} (\mathbf{x} - \mathbf{x}_0)$. Letting $\mathbf{y} = \mathbf{x} - \mathbf{x}_{\gamma, \mathbf{x}_0}^*$, we can follow a similar calculation to Lemma 42 to show:

$$\begin{aligned} & \sum_{i=1}^n \frac{1}{p_i} \left\| \nabla \hat{\psi}_{\gamma, \mathbf{x}_0}^i(\mathbf{x}) - \nabla \hat{\psi}_{\gamma, \mathbf{x}_0}^i(\mathbf{x}_{\gamma, \mathbf{x}_0}^*) \right\|_2^2 \\ &= \|\mathbf{B}\|_F^2 \sum_{i=1}^{n+d} \frac{\left\| (\mathbf{b}_i \mathbf{b}_i^T + \frac{\gamma \|\mathbf{b}_i\|_2^2}{\|\mathbf{B}\|_F^2} \mathbf{I}) \mathbf{y} \right\|_2^2}{\|\mathbf{b}_i\|_2^2} \\ &\leq 2(\|\mathbf{B}\|_F^2 + 2\gamma) [f(\mathbf{x}) - f(\mathbf{x}_{\gamma, \mathbf{x}_0}^*)]. \end{aligned}$$

Let $\bar{S} = \|\mathbf{B}\|_F^2$ and $\mu = \lambda_d(\mathbf{P}^{-1/2}(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I}) \mathbf{P}^{-1/2})$ be an lower bound on the strong convexity of $\hat{f}$. Denote $\frac{\gamma}{\mu} = r$. The strong convexity of $\hat{f}_{\gamma, \mathbf{x}_0}$ is lower bounded by $\mu_\gamma = \lambda_d(\mathbf{P}^{-1/2}(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I}) \mathbf{P}^{-1/2}) + \gamma = \Theta(r \cdot \lambda_d(\mathbf{P}^{-1/2}(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I}) \mathbf{P}^{-1/2}))$ which gives:

$$\frac{\bar{S}}{\mu_\gamma} = \frac{\text{tr}(\mathbf{P}^{-1/2}(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I}) \mathbf{P}^{-1/2}) + 2\gamma}{\lambda_d(\mathbf{P}^{-1/2}(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I}) \mathbf{P}^{-1/2}) + \gamma} = O\left(\frac{k\sigma_k^2(\mathbf{A}) + \sum_{i=k+1}^d \sigma_i^2(\mathbf{A})}{r\lambda} + \frac{d}{r}\right)$$

So, by Theorem 43, for dense inputs, ignoring the $O(nd^{\omega(1,1,\log_d k)-1} \log d)$ precomputation cost to compute $\mathbf{Z}$, $\mathbf{P}^{-1/2}$ and $\mathbf{A} \mathbf{P}^{-1/2}$, which we only pay once, letting $\hat{\kappa} \stackrel{\text{def}}{=} \frac{k\sigma_k^2(\mathbf{A}) + \sum_{i=k+1}^d \sigma_i^2(\mathbf{A})}{d\lambda}$, we have $\mathcal{T}_c = O\left(\log c \cdot \left(nd + \frac{d^2 \hat{\kappa}}{r}\right)\right)$. If $nd \geq d^2 \hat{\kappa}$ then we already have runtime $O(nd \log(1/\epsilon))$ by Theorem 43. Otherwise, setting $\frac{\gamma}{\mu} \stackrel{\text{def}}{=} r = \frac{d\hat{\kappa}}{n}$ we can solve $\hat{f}$ up to ϵ accuracy in time:

$$O\left(\mathcal{T}_{4\left(\frac{2\gamma+\mu}{\mu}\right)^{3/2}} \sqrt{\gamma/\mu} \log(1/\epsilon)\right) = O\left(\log(1/\epsilon) n^{1/2} d^{3/2} \log(\bar{\kappa}) \sqrt{\bar{\kappa}}\right).$$

For sparse inputs, we again ignore the one time precomputation cost of $O(\text{nnz}(\mathbf{A})k \log d + dk^{\omega-1})$ to compute $\mathbf{Z}$ and $\mathbf{A} \mathbf{Z}$. We have:

$$\mathcal{T}_c = O\left(\log c \cdot \left(\text{nnz}(\mathbf{A}) + [d \cdot d_s(\mathbf{A}) + dk] \frac{\bar{\kappa}}{r}\right)\right).$$

If $\text{nnz}(\mathbf{A}) \geq [d \cdot d_s(\mathbf{A}) + dk] \bar{\kappa}$ then we already have runtime $O(\text{nnz}(\mathbf{A}) \log(1/\epsilon))$ by Theorem 43. Otherwise, $\text{nnz}(\mathbf{A}) \leq [d \cdot d_s(\mathbf{A}) + dk] \bar{\kappa}$. So setting $r = \frac{\bar{\kappa}[d \cdot d_s(\mathbf{A}) + dk]}{\sqrt{\text{nnz}(\mathbf{A})}}$ we have:

$$O\left(\mathcal{T}_{4\left(\frac{2\gamma+\mu}{\mu}\right)^{3/2}} \sqrt{\gamma/\mu} \log(1/\epsilon)\right) = O\left(\log(1/\epsilon) \log(d\bar{\kappa}) \cdot \sqrt{\text{nnz}(\mathbf{A})[d \cdot d_s(\mathbf{A}) + dk] \bar{\kappa}}\right).$$

□

A.4 Preconditioned Iterative Methods

Finally, we describe how to combine deflation-based preconditioning with standard iterative methods, which can give runtime advantages over Theorem 49 in some parameter regimes.

Theorem 51 (Preconditioned Iterative Method). *For any $\mathbf{A} \in \mathbb{R}^{n \times d}$, $\mathbf{b} \in \mathbb{R}^d$, and $\lambda > 0$, let $\mathbf{M}_\lambda \stackrel{\text{def}}{=} \mathbf{A}^T \mathbf{A} + \lambda \mathbf{I}$ and $\hat{\kappa} = \frac{\sigma_{k+1}^2(\mathbf{A})}{\lambda}$ where $k \in [0, \dots, d]$ is an input parameter. There is an algorithm that uses $O(\text{nnz}(\mathbf{A})k \log d + dk^{\omega-1})$ precomputation time after which, given any input $\mathbf{y} \in \mathbb{R}^d$, letting $\mathbf{x}^* \stackrel{\text{def}}{=} \mathbf{M}_\lambda^{-1} \mathbf{y}$ the algorithm returns $\mathbf{x}$ such that with high probability $\|\mathbf{x} - \mathbf{x}^*\|_{\mathbf{M}_\lambda} \leq \epsilon \|\mathbf{x}^*\|_{\mathbf{M}_\lambda}$ in time*

$$O\left(\log(1/\epsilon)(\text{nnz}(\mathbf{A}) + dk)\sqrt{\hat{\kappa}}\right).$$

Proof. If we form the preconditioner $\mathbf{P}^{-1/2}$ as in Lemma 45, by Lemmas 2 and 4 of [GOSS16] we have the preconditioned condition number bound:

$$\hat{\kappa} \stackrel{\text{def}}{=} \frac{\lambda_1(\mathbf{P}^{-1/2} \mathbf{M}_\lambda \mathbf{P}^{-1/2})}{\lambda_d(\mathbf{P}^{-1/2} \mathbf{M}_\lambda \mathbf{P}^{-1/2})} = O\left(\frac{\sigma_{k+1}^2(\mathbf{A})}{\lambda}\right)$$

Note that the bound in [GOSS16] is actually in terms of $\sigma_k^2(\mathbf{A})$, however we write $\sigma_{k+1}^2(\mathbf{A})$ so that our theorem holds in the case that $k = 0$. This can be achieved with no effect on the asymptotic runtime by setting k to $k + 1$.

We can now apply any accelerated linear system solver, such as Chebyshev iteration, Conjugate Gradient, or Accelerated Gradient Descent [Saa03, Nes13] to obtain $\mathbf{x}$ with $\|\mathbf{x} - \mathbf{x}^*\|_{\mathbf{M}_\lambda} \leq \epsilon \|\mathbf{x}^*\|_{\mathbf{M}_\lambda}$ in $O\left(\log(1/\epsilon)\sqrt{\hat{\kappa}} \cdot \text{matvec}(\mathbf{P}^{-1/2}(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})\mathbf{P}^{-1/2})\right)$ where $\text{matvec}(\mathbf{P}^{-1/2}(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})\mathbf{P}^{-1/2})$ is the time required to multiply a single vector by this matrix. $\text{matvec}(\mathbf{P}^{-1/2}(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})\mathbf{P}^{-1/2}) = O(\text{nnz}(\mathbf{A}) + dk)$ since $\mathbf{P}^{-1/2} = \mathbf{Z}\tilde{\Sigma}^{-1/2}\mathbf{Z}^T + \delta(\mathbf{I} - \mathbf{Z}\mathbf{Z}^T)$ can be applied in $O(dk)$ time. This gives the result combined with the precomputation time for $\mathbf{P}^{-1/2}$ from Lemma 44. $\square$

B Additional Proofs: Lower Bounds

Lemma 52 (Determinant Hardness). *Given algorithm $\mathcal{A}$ which given $\mathbf{B} \in \mathbb{R}^{n \times n}$ returns $X \in (1 \pm \epsilon) \det(\mathbf{B})$ in $O(n^\gamma \epsilon^{-c})$ time, we can detect if an n -node graph contains a triangle in $O(n^{\gamma+12c})$ time.*

Proof. Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be the adjacency matrix of an n -node graph G . Let $\lambda_1, \dots, \lambda_n$ denote its eigenvalues. Let $\mathbf{B} = \mathbf{I} + \delta \mathbf{A}$ for some δ which we will set later. We can write:

$$\det(\mathbf{B}) = \prod_{i=1}^n \lambda_i(\mathbf{B}) = \prod_{i=1}^n (1 + \delta \lambda_i) = \sum_{k=0}^n \left(\delta^k \cdot \sum_{i_1 < i_2 < \dots < i_k} \lambda_{i_1} \lambda_{i_2} \dots \lambda_{i_k} \right). \quad (31)$$

The $k = 0$ term in (31) is 1, and the next two are easy to compute. $\delta \sum_{i=1}^n \lambda_i = \delta \text{tr}(\mathbf{A}) = 0$, and $\delta^2 \sum_{i < j} \lambda_i \lambda_j = \frac{\delta^2}{2} \left(\sum_{i,j} \lambda_i \lambda_j - \sum_i \lambda_i^2 \right) = \frac{\delta^2}{2} \sum_i \lambda_i \text{tr}(\mathbf{A}) - \frac{\delta^2}{2} \|\mathbf{A}\|_F^2 = -\delta^2 \|\mathbf{A}\|_F^2 / 2$. For $k = 3$ we

have:

$$\begin{aligned}
\delta^3 \sum_{i < j < k} \lambda_i \lambda_j \lambda_k &= \frac{\delta^3}{3} \left(\sum_{i < j} \lambda_i \lambda_j \operatorname{tr}(\mathbf{A}) - \sum_{i \neq j} \lambda_i^2 \lambda_j \right) \\
&= 0 - \frac{\delta^3}{3} \|\mathbf{A}\|_F^2 \cdot \operatorname{tr}(\mathbf{A}) + \frac{\delta^3}{3} \operatorname{tr}(\mathbf{A}^3) \\
&= \frac{\delta^3}{3} \operatorname{tr}(\mathbf{A}^3).
\end{aligned}$$

We will bound the $k > 3$ terms by: $\left| \delta^k \cdot \sum_{i_1 < i_2 < \dots < i_k} \lambda_{i_1} \lambda_{i_2} \dots \lambda_{i_k} \right| \leq \binom{n}{k} \delta^k \lambda_1^k \leq (n\delta\lambda_1)^k \leq (n^2\delta)^k$ since $\lambda_1 \leq n$. However, in order to obtain a tighter result, we will use stronger bounds for $k = 4, 5$. These bounds are very tedious but straightforward. Specifically:

$$\begin{aligned}
\left| \delta^4 \sum_{i < j < k < l} \lambda_i \lambda_j \lambda_k \lambda_l \right| &= \frac{\delta^4}{4} \left| \operatorname{tr}(\mathbf{A}) \sum_{i < j < k} \lambda_i \lambda_j \lambda_k - \frac{1}{2} \sum_{i \neq j \neq k} \lambda_i^2 \lambda_j \lambda_k \right| \\
&= \frac{\delta^4}{8} \left| \operatorname{tr}(\mathbf{A}) \sum_{i \neq j} \lambda_i^2 \lambda_j - \sum_{i \neq j} \lambda_i^2 \lambda_j^2 - \sum_{i \neq j} \lambda_i^3 \lambda_j \right| \\
&= \frac{\delta^4}{8} \left| \sum_{i \neq j} \lambda_i^2 \lambda_j^2 + \operatorname{tr}(\mathbf{A}) \sum_{i \neq j} \lambda_i^3 - \sum_i \lambda_i^4 \right| \\
&= \frac{\delta^4}{8} \left| \|\mathbf{A}\|_F^2 - 2 \operatorname{tr}(\mathbf{A}^4) \right| \leq \frac{\delta^4 n^4}{4}.
\end{aligned}$$

And similarly:

$$\begin{aligned}
\left| \delta^5 \sum_{i < j < k < l < m} \lambda_i \lambda_j \lambda_k \lambda_l \lambda_m \right| &= \frac{\delta^5}{30} \left| \sum_{i \neq j \neq k \neq l} \lambda_i^2 \lambda_j \lambda_k \lambda_l \right| \\
&= \frac{\delta^5}{30} \left| 2 \sum_{i \neq j \neq k} \lambda_i^2 \lambda_j^2 \lambda_k + \sum_{i \neq j \neq k} \lambda_i^3 \lambda_j \lambda_k \right| \\
&= \frac{\delta^5}{30} \left| 5 \sum_{i \neq j} \lambda_i^2 \lambda_j^3 + \sum_{i \neq j} \lambda_i^4 \lambda_j \right| \\
&= \frac{\delta^5}{30} \left| 5 \left(\sum \lambda_i^2 \right) \left(\sum \lambda_i^3 \right) - 6 \sum \lambda_i^5 \right| \\
&\leq \frac{\delta^5 n^2}{6} \operatorname{tr}(\mathbf{A}^3) + \frac{\delta^5}{5} \lambda_1 \sum \lambda_i^4 \\
&\leq \frac{\delta^5 n^2}{6} \operatorname{tr}(\mathbf{A}^3) + \frac{\delta^5 n^5}{5}.
\end{aligned}$$

Finally, if we set $\delta = \frac{1}{10n^4}$ then we have:

$$\begin{aligned} \left| \sum_{k=4}^n \left(\delta^k \cdot \sum_{i_1 < i_2 < \dots < i_k} \lambda_{i_1} \lambda_{i_2} \dots \lambda_{i_k} \right) \right| &\leq \frac{\delta^4 n^4}{4} + \frac{\delta^5 n^5}{5} + \frac{\delta^5 n^2}{6} \text{tr}(\mathbf{A}^3) + \sum_{k=6}^{\infty} (n^2 \delta)^k \\ &\leq \delta^3 \left(\frac{1}{40} + \frac{1}{500} + \frac{1}{600} \text{tr}(\mathbf{A}^3) + \left(\frac{1}{10^3} + \frac{1}{10^5} + \dots \right) \right) \\ &\leq \frac{\delta^3}{30} + \frac{\delta^3}{600} \text{tr}(\mathbf{A}^3). \end{aligned}$$

We then write:

$$\det(\mathbf{B}) = 1 - \frac{\delta^2 \|\mathbf{A}\|_F^2}{2} + \frac{\delta^3 \text{tr}(\mathbf{A}^3)}{3} \pm \frac{\delta^3}{30} \pm \frac{\delta^3}{600} \text{tr}(\mathbf{A}^3).$$

Since $1 \leq \delta^3 \cdot 10^3 n^{12}$ and $\frac{\delta^2 \|\mathbf{A}\|_F^2}{2} \leq \delta^3 \cdot 5n^6$ if we compute $X \in (1 \pm c_1/n^{12}) \det(\mathbf{B})$ for sufficiently small constant c_1 and subtract off $\left(1 - \frac{\delta^2 \|\mathbf{A}\|_F^2}{2}\right)$, we will be able to determine if $\text{tr}(\mathbf{A}^3) > 0$ and hence detect if G has a triangle. So any algorithm approximating $\det(\mathbf{B})$ to $(1 \pm \epsilon)$ error in $O(n^\gamma \epsilon^{-c})$ time yields a triangle detection algorithm running in $O(n^{\gamma+12c})$ time. $\square$

C Krylov and Power Methods preserve Invariant Norms

In this section, we show that when the Krylov method [MM15] or the power method is used to deflate the top singular vectors of the matrix, any unitarily invariant norm of the tail (remaining part of the spectrum) is preserved. Let $\mathbf{P} = \mathbf{Z}\mathbf{Z}^\top$ be the projector obtained for the top k singular vectors of $\mathbf{A}$ using the Krylov method [MM15] or the power method. Then, for any invariant norm of the tail to be preserved, we just need to show the following:

$$\sigma_i((\mathbf{I} - \mathbf{P})\mathbf{A}) \leq (1 + \epsilon)\sigma_i(\mathbf{A} - \mathbf{A}_k) + \frac{\epsilon}{n}\sigma_1(\mathbf{A}). \quad (32)$$

Since $\mathbf{P}$ is a $(1 + \epsilon)$ approximation obtained in terms of the spectral norm, we have by the spectral low rank approximation guarantee of [MM15] that $\sigma_1((\mathbf{I} - \mathbf{P})\mathbf{A}) \leq (1 + \epsilon)\sigma_1(\mathbf{A} - \mathbf{A}_k)$, where $\mathbf{A}_k$ is the best rank k approximation of $\mathbf{A}$. Consider first any i for which $\sigma_i(\mathbf{A} - \mathbf{A}_k) \geq (1 - \epsilon)\sigma_1(\mathbf{A} - \mathbf{A}_k) = (1 - \epsilon)\sigma_{k+1}(\mathbf{A})$. Then we have,

$$\sigma_i((\mathbf{I} - \mathbf{P})\mathbf{A}) \leq \sigma_1((\mathbf{I} - \mathbf{P})\mathbf{A}) \leq (1 + \epsilon)\sigma_{k+1}(\mathbf{A}) \leq \frac{(1 + \epsilon)}{(1 - \epsilon)}\sigma_i(\mathbf{A} - \mathbf{A}_k) \leq (1 + 3\epsilon)\sigma_i(\mathbf{A} - \mathbf{A}_k).$$

For any i with $\sigma_i(\mathbf{A} - \mathbf{A}_k) < (1 - \epsilon)\sigma_{k+1}(\mathbf{A})$, i.e., with $\sigma_{i+k}(\mathbf{A}) \leq (1 - \epsilon)\sigma_{k+1}(\mathbf{A})$ there is a large ($> \epsilon$) relative gap between this singular value and σ_k . We have by the min-max characterization of singular values:

$$\sigma_i((\mathbf{I} - \mathbf{P})\mathbf{A}) = \min_{\mathbf{Y} | \text{rank}(\mathbf{Y}) = n - i + 1} \left(\max_{\mathbf{y} \in \text{span}(\mathbf{Y}) \| \mathbf{y} \|_2 = 1} \| \mathbf{y}^T (\mathbf{I} - \mathbf{P})\mathbf{A} \|_2 \right)$$

If we just set $\mathbf{Y} = \mathbf{Z}$, then we have:

$$\sigma_{n-k+1}((\mathbf{I} - \mathbf{P})\mathbf{A}) \leq \max_{\mathbf{y} \in \text{span}(\mathbf{Z}) \| \mathbf{y} \|_2 = 1} \| \mathbf{y}^T (\mathbf{I} - \mathbf{P})\mathbf{A} \|_2 = 0.$$

It follows that the bottom k singular values are all 0, and so equal to the bottom k singular values of $\sigma_i(\mathbf{A} - \mathbf{A}_k)$.

Now, for $i < n - k + 1$, let $\mathbf{U}_U$ denote the top $i + k - 1$ singular vectors of $\mathbf{A}$ and $\mathbf{U}_L$ denote the bottom $n - (i + k - 1)$ singular vectors. We set $\mathbf{Y} = [\mathbf{Z}, \mathbf{U}_L]$. First, we note that $\mathbf{Y}$ has $k + n - (i + k - 1) = n - i + 1$ columns and also $\text{rank}(\mathbf{Y}) = n - i + 1$. This is because, if we consider $\mathbf{Y}^T \mathbf{Y}$, the top left $k \times k$ blocks is $\mathbf{Z}^T \mathbf{Z} = \mathbf{I}$ and the bottom right $n - i - k + 1$ block is $\mathbf{U}_L^T \mathbf{U}_L = \mathbf{I}$. The off-diagonal entries in the top right and bottom left blocks are all bounded by $1/\text{poly}(n)$; the proof of this latter statement is given in the latter part of this section, where we bound $\|\mathbf{Z}^T \mathbf{U}_L\|_2$. By the Gershgorin circle theorem [GVL12], all eigenvalues of $\mathbf{Y}^T \mathbf{Y}$ are in the range $1 \pm 1/\text{poly}(n)$ and so the matrix is full rank. So, we have:

$$\sigma_i((\mathbf{I} - \mathbf{P})\mathbf{A}) \leq \max_{\mathbf{y} \in \text{span}(\mathbf{Y}) \|\mathbf{y}\|_2=1} \|\mathbf{y}^T (\mathbf{I} - \mathbf{P})\mathbf{A}\|_2. \quad (33)$$

Next, we can write $\mathbf{y} = [\mathbf{Z}, \mathbf{U}_L]\mathbf{w}$ for some $\mathbf{w}$. By our argument above, every singular value of $[\mathbf{Z}, \mathbf{U}_L]$ lies in $1 \pm 1/\text{poly}(n)$. Then, we have $\|\mathbf{w}\|_2 \leq 1 + 1/\text{poly}(n)$. Splitting $\mathbf{w} = \mathbf{w}_1 + \mathbf{w}_2$, where $\mathbf{w}_1$ contains the first k coordinates of the vector and $\mathbf{w}_2$ contains the rest, we have:

$$\begin{aligned} (\mathbf{I} - \mathbf{P})\mathbf{y} &= (\mathbf{I} - \mathbf{P})\mathbf{Z}\mathbf{w}_1 + (\mathbf{I} - \mathbf{P})\mathbf{U}_L\mathbf{w}_2 \\ &= 0 + (\mathbf{I} - \mathbf{P})\mathbf{U}_L\mathbf{w}_2 \\ &= \mathbf{U}_L\mathbf{w}_2 - \mathbf{Z}\mathbf{Z}^T\mathbf{U}_L\mathbf{w}_2. \end{aligned}$$

Then, we have

$$\|\mathbf{Z}\mathbf{Z}^T\mathbf{U}_L\mathbf{w}_2\|_2 \leq \|\mathbf{Z}^T\mathbf{U}_L\|_2 \cdot \|\mathbf{w}_2\|_2 \leq 1/\text{poly}(n),$$

where the inequality comes from bounding $\|\mathbf{Z}^T\mathbf{U}_L\|_2$ using the fact that all its entries are at most $1/\text{poly}(n)$ (see the latter part of this section), and bounding $\|\mathbf{w}_2\|_2 \leq \|\mathbf{w}\|_2 \leq 1 + 1/\text{poly}(n)$. Thus, finally we obtain:

$$\begin{aligned} \|\mathbf{y}^T (\mathbf{I} - \mathbf{P})\mathbf{A}\|_2 &\leq \|\mathbf{w}_2^T \mathbf{U}_L^T \mathbf{A}\|_2 + \|\mathbf{w}_2^T \mathbf{U}_L^T \mathbf{Z}\mathbf{Z}^T \mathbf{A}\|_2 \\ &\leq (1 + 1/\text{poly}(n)) \cdot \|\mathbf{U}_L^T \mathbf{A}\|_2 + 1/\text{poly}(n) \cdot \|\mathbf{A}\|_2 \\ &\leq (1 + 1/\text{poly}(n))\sigma_i(\mathbf{A} - \mathbf{A}_k) + \frac{1}{\text{poly}(n)}\sigma_1(\mathbf{A}). \end{aligned}$$

Plugging into (33) gives the proof of (32). Note that we can assume $\frac{1}{\text{poly}(n)} \leq \frac{\epsilon}{n}$ since if $\epsilon = o(1/\text{poly}(n))$, we can just compute the SVD.

Bound for $\|\mathbf{Z}^T\mathbf{U}_L\|_2$: Let $\mathbf{Z}$ be an orthonormal basis for $\mathbf{A}^q\mathbf{\Omega}$ from the power method (or $p_q(\mathbf{A})\mathbf{\Omega}$ from block Krylov method). Then, for any singular vector $\mathbf{u}_i$ with corresponding singular value $\sigma_i(\mathbf{A})$ with $\sigma_i(\mathbf{A}) \leq (1 - \epsilon)\sigma_k(\mathbf{A})$, we have $\|\mathbf{u}_i^T \mathbf{Z}\|_2 \leq 1/\text{poly}(n)$.

To see this, we first write $\mathbf{A}^q\mathbf{\Omega} = \mathbf{Z}\mathbf{T}\mathbf{W}^T$ in its SVD form. Then, we have $\|\mathbf{u}_i^T \mathbf{A}^q\mathbf{\Omega}\|_2 = \|\mathbf{u}_i^T \mathbf{Z}\mathbf{T}\|_2 \geq \|\mathbf{u}_i^T \mathbf{Z}\|_2 \tau_{k,k}$, where $\tau_{k,k}$ is the k th diagonal entry of $\mathbf{T}$. On the other hand, we also have $\|\mathbf{u}_i^T \mathbf{A}^q\mathbf{\Omega}\|_2 = \sigma_i(\mathbf{A})^q \|\mathbf{v}_i^T \mathbf{\Omega}\|_2$, where $\mathbf{v}_i^T$ is the right singular vector corresponding to $\mathbf{u}_i$ of $\mathbf{A}$. Since $\mathbf{\Omega}$ is i.i.d. Gaussian, we have $\|\mathbf{v}_i^T \mathbf{\Omega}\|_2^2 = O(k \log n)$ with probability $1 - 1/n^2$ for a fixed $\mathbf{v}_i$ (see proof of Lemma 2.12 in [Woo14]). So, we can union bound over all n right singular vectors and the relation holds for all right singular vectors $\mathbf{v} \in \mathbf{V}$. We condition on this event.

This event implies that $\|\mathbf{u}_i^T \mathbf{A}^q\mathbf{\Omega}\|_2 = O(\sigma_i(\mathbf{A})^q \sqrt{k \log n})$. It follows that, if we show $\tau_{k,k} \geq \sigma_i(\mathbf{A})^q \text{poly}(n)$, then $\|\mathbf{u}_i^T \mathbf{Z}\|_2 \leq 1/\text{poly}(n)$.

Next, to show that $\tau_{k,k} \geq \sigma_i(\mathbf{A})^q \text{poly}(n)$, it suffices to show $\tau_{k,k} \geq \sigma_k(\mathbf{A})^q \sqrt{k}$. Indeed, note that for $q = \Theta((\log n)/\epsilon)$ (for the block Krylov method, we need to use higher number of iterations than required, which is $q = \Theta((\log n)/\sqrt{\epsilon})$ [MM15]), by definition of i , we have $\sigma_k(\mathbf{A}^q)/\sigma_i(\mathbf{A}^q) \geq \text{poly}(n)$.

For this, we have $\tau_{k,k} = \sigma_k(\mathbf{U}\Sigma^q\mathbf{V}^T\mathbf{\Omega}) \geq \sigma_k(\mathbf{U}_k\mathbf{U}_k^T\mathbf{U}\Sigma^q\mathbf{V}^T\mathbf{\Omega})$ since $\sigma_i(\mathbf{P}\mathbf{A}) \leq \sigma_i(\mathbf{A})$ for any projection matrix $\mathbf{P}$ and any matrix $\mathbf{A}$. Then, we have $\sigma_k(\mathbf{U}_k\mathbf{U}_k^T\mathbf{U}\Sigma^q\mathbf{V}^T\mathbf{\Omega}) = \sigma_k(\mathbf{U}_k\Sigma_k^q\mathbf{V}_k^T\mathbf{\Omega}) = \sigma_k(\Sigma_k^q\mathbf{V}_k^T\mathbf{\Omega})$ by the definition of the SVD, and that the columns of $\mathbf{U}_k$ are orthonormal. So, we get $\tau_{k,k} \geq \sigma_k(\Sigma_k^q\mathbf{H})$, where $\mathbf{H} = \mathbf{V}_k^T\mathbf{\Omega}$ is a $k \times k$ matrix with i.i.d. Gaussian distribution due to the rotational invariance of the Gaussian distribution. Since the k -th singular value is the smallest singular value of $\mathbf{H}$ and Σ_k , we then have $\sigma_k(\Sigma_k^q\mathbf{H}) \geq \sigma_k^q(\mathbf{A})\sigma_k(\mathbf{H}) \geq \sigma_k^q\sqrt{k} \cdot C$ with probability at least 9/10, for an arbitrary constant $C > 0$, using standard properties of minimum singular values of squared Gaussian matrices (see e.g., Fact 6 in section 4.3 of [Woo14]). Thus, we have $\tau_{k,k} \geq C\sqrt{k}\sigma_k^q(\mathbf{A})$, which completes the proof.

D Additional Proofs

Claim 10. *Let $f : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ be a δ_f -multiplicatively smooth function on the range $[a, b]$. For any $x, y \in [a, b]$ and $c \in (0, \frac{1}{3\delta_f})$*

$$y \in [(1-c)x, (1+c)x] \Rightarrow f(y) \in [(1-3\delta_fc)f(x), (1+3\delta_fc)f(x)].$$

Proof. Let R denote the range $[\min(x, y), \max(x, y)]$. For $y \in [(1-c)x, (1+c)x]$ we have:

$$|f(x) - f(y)| \leq |x - y| \cdot \sup_{z \in R} |f'(z)| \leq cx \cdot \frac{\delta_f \sup_{z \in R} f(z)}{\min(x, y)} \leq \frac{c\delta_f \sup_{z \in R} f(z)}{1-c}. \quad (34)$$

Similarly, letting $Z = \sup_{z \in R} f(z)$ we have:

$$Z - f(x) \leq |x - y| \frac{\delta_f Z}{\min(x, y)} \leq \frac{c\delta_f Z}{1-c} \leq \frac{c}{\delta_f} 1 - 1/3 - 1/3$$

and hence $Z \leq \frac{1-c}{1-(\delta_f+1)c} f(x)$. So plugging into (34) we have:

$$|f(x) - f(y)| \leq \frac{c\delta_f}{1-(\delta_f+1)c} \cdot f(x) \leq \frac{c\delta_f}{1-1/3-1/3} f(x) = 3\delta_f f(x)$$

which gives the result. □