FIXED-RANK APPROXIMATION OF A POSITIVE-SEMIDEFINITE MATRIX FROM STREAMING DATA

JOEL A. TROPP, ALP YURTSEVER, MADELEINE UDELL, AND VOLKAN CEVHER

Technical Report No. 2017-03
May 2017

APPLIED & COMPUTATIONAL MATHEMATICS
CALIFORNIA INSTITUTE OF TECHNOLOGY
mail code 9-94 · pasadena, ca 91125

Fixed-Rank Approximation of a Positive-Semidefinite Matrix from Streaming Data

Including supplementary appendix

Joel A. Tropp	**Alp Yurtsever**	**Madeleine Udell**	**Volkan Cevher**
Caltech	EPFL	Cornell	EPFL
jtropp@caltech.edu	alp.yurtsever@epfl.ch	mru8@cornell.edu	volkan.cevher@epfl.ch

Abstract

Several important applications, such as streaming PCA and semidefinite programming, involve a large-scale positive-semidefinite (psd) matrix that is presented as a sequence of linear updates. Because of storage limitations, it may only be possible to retain a sketch of the psd matrix. This paper develops a new algorithm for fixed-rank psd approximation from a sketch. The approach combines the Nyström approximation with a novel mechanism for rank truncation. Theoretical analysis establishes that the proposed method can achieve any prescribed relative error in the Schatten 1-norm and that it exploits the spectral decay of the input matrix. Computer experiments show that the proposed method dominates alternative techniques for fixed-rank psd matrix approximation across a wide range of examples.

1 Motivation

In recent years, researchers have studied many applications where a large positive-semidefinite (psd) matrix is presented as a series of linear updates. A recurring theme is that we only have space to store a small summary of the psd matrix, and we must use this information to construct an accurate psd approximation with specified rank. Here are two important cases where this problem arises.

Streaming Covariance Estimation. Suppose that we receive a stream $\boldsymbol{h}_1, \boldsymbol{h}_2, \boldsymbol{h}_3, \cdots \in \mathbb{R}^n$ of high-dimensional vectors. The psd sample covariance matrix of these vectors has the linear dynamics

$$\boldsymbol{A}^{(0)} \leftarrow \mathbf{0} \quad \text{and} \quad \boldsymbol{A}^{(i)} \leftarrow (1 - i^{-1})\boldsymbol{A}^{(i-1)} + i^{-1}\boldsymbol{h}_i\boldsymbol{h}_i^*.$$

When the dimension n and the number of vectors are both large, it is not possible to store the vectors or the sample covariance matrix. Instead, we wish to maintain a small summary that allows us to compute the rank-r psd approximation of the sample covariance matrix $\boldsymbol{A}^{(i)}$ at a specified instant i. This problem and its variants are often called *streaming PCA* [4, 13, 15, 16, 25, 32].

Convex Low-Rank Matrix Optimization with Optimal Storage. A primary application of semidefinite programming (SDP) is to search for a rank-r psd matrix that satisfies additional constraints. Because of storage costs, SDPs are difficult to solve when the matrix variable is large. Recently, Yurtsever et al. [42] exhibited the first provable algorithm, called SketchyCGM, that produces a rank-r approximate solution to an SDP *using optimal storage*.

Implicitly, SketchyCGM forms a sequence of approximate psd solutions to the SDP via the iteration

$$\boldsymbol{A}^{(0)} \leftarrow \mathbf{0} \quad \text{and} \quad \boldsymbol{A}^{(i)} \leftarrow (1 - \eta_i)\boldsymbol{A}^{(i-1)} + \eta_i\boldsymbol{h}_i\boldsymbol{h}_i^*.$$

The step size $\eta_i = 2/(i+2)$, and the vectors $\boldsymbol{h}_i$ do not depend on the matrices $\boldsymbol{A}^{(i)}$. In fact, SketchyCGM only maintains a small summary of the evolving solution $\boldsymbol{A}^{(i)}$. When the iteration terminates, SketchyCGM computes a rank-r psd approximation of the final iterate using the method described by Tropp et al. [36, Alg. 9].

Date: 18 May 2017

1.1 Notation and Background

The scalar field $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$. Define $\alpha(\mathbb{R}) = 1$ and $\alpha(\mathbb{C}) = 0$. The asterisk * is the (conjugate) transpose, and the dagger † denotes the Moore–Penrose pseudoinverse. The notation $\boldsymbol{A}^{1/2}$ refers to the unique psd square root of a psd matrix $\boldsymbol{A}$. For $p \in [1, \infty]$, the Schatten p-norm $\|\cdot\|_p$ returns the ℓ_p norm of the singular values of a matrix. As usual, σ_r refers to the rth largest singular value.

For a nonnegative integer r, the phrase "rank-r" and its variants mean "rank at most r." For a matrix $\boldsymbol{M}$, the symbol $[\![\boldsymbol{M}]\!]_r$ denotes a (simultaneous) best rank-r approximation of the matrix $\boldsymbol{M}$ with respect to any Schatten p-norm. We can take $[\![\boldsymbol{M}]\!]_r$ to be any r-truncated singular value decomposition (SVD) of $\boldsymbol{M}$ [24, Sec. 6]. Every best rank-r approximation of a psd matrix is psd.

2 Sketching and Fixed-Rank PSD Approximation

We begin with a streaming data model for a psd matrix that evolves via a sequence of general linear updates, and it describes a randomized linear sketch for tracking the psd matrix. To compute a fixed-rank psd approximation, we develop an algorithm based on the Nyström method [38], a technique from the literature on kernel methods. In contrast to previous approaches, **our algorithm uses a distinct mechanism to truncate the rank of the approximation.**

The Streaming Model. Fix a rank parameter r in the range $1 \leq r \leq n$. Initially, the psd matrix $\boldsymbol{A} \in \mathbb{F}^{n \times n}$ equals a known psd matrix $\boldsymbol{A}_{\text{init}} \in \mathbb{F}^{n \times n}$. Then $\boldsymbol{A}$ evolves via a series of linear updates:

$$\boldsymbol{A} \leftarrow \theta_1 \boldsymbol{A} + \theta_2 \boldsymbol{H} \quad \text{where} \quad \theta_i \in \mathbb{R}, \quad \boldsymbol{H} \in \mathbb{F}^{n \times n} \text{ is (conjugate) symmetric.} \tag{2.1}$$

In many applications, the innovation $\boldsymbol{H}$ is low-rank and/or sparse. We assume that the evolving matrix $\boldsymbol{A}$ always remains psd. At one given instant, we must produce an accurate rank-r approximation of the psd matrix $\boldsymbol{A}$ induced by the stream of linear updates.

The Sketch. Fix a sketch size parameter k in the range $r \leq k \leq n$. Independent from $\boldsymbol{A}$, we draw and fix a random test matrix

$$\boldsymbol{\Omega} \in \mathbb{F}^{n \times k}. \tag{2.2}$$

See Sec. 3 for a discussion of possible distributions. The sketch of the matrix $\boldsymbol{A}$ takes the form

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{\Omega} \in \mathbb{F}^{n \times k}. \tag{2.3}$$

The sketch (2.3) supports updates of the form (2.1):

$$\boldsymbol{Y} \leftarrow \theta_1 \boldsymbol{Y} + \theta_2 \boldsymbol{H}\boldsymbol{\Omega}. \tag{2.4}$$

To find a good rank-r approximation, we must set the sketch size k larger than r. But storage costs and computation also increase with k. One of our main contributions is to clarify the role of k.

Under the model (2.1), it is more or less necessary to use a randomized linear sketch to track $\boldsymbol{A}$ [28]. For psd matrices, sketches of the form (2.2)–(2.3) appear explicitly in Gittens's work [17, 18, 20]. Tropp et al. [36] relies on a more complicated sketch developed in [8, 40].

The Nyström Approximation. The Nyström method is a general technique for low-rank psd matrix approximation. Various instantiations appear in the papers [6, 12, 14, 17, 18, 20, 23, 27, 34, 38].

Here is the application to the present situation. Given the test matrix $\boldsymbol{\Omega}$ and the sketch $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{\Omega}$, the Nyström method constructs a rank-k psd approximation of the psd matrix $\boldsymbol{A}$ via the formula

$$\hat{\boldsymbol{A}}^{\text{nys}} = \boldsymbol{Y}(\boldsymbol{\Omega}^*\boldsymbol{Y})^\dagger \boldsymbol{Y}^*. \tag{2.5}$$

In most work on the Nyström method, the test matrix $\boldsymbol{\Omega}$ depends adaptively on $\boldsymbol{A}$, so these approaches are not valid in the streaming setting. Gittens's framework [17, 18, 20] covers the streaming case.

Fixed-Rank Nyström Approximation: Prior Art. To construct a Nyström approximation with exact rank r from a sketch of size k, the standard approach is to truncate the center matrix to rank r:

$$\hat{\boldsymbol{A}}_r^{\text{nysfix}} = \boldsymbol{Y}([\![\boldsymbol{\Omega}^*\boldsymbol{Y}]\!]_r)^\dagger \boldsymbol{Y}^*. \tag{2.6}$$

The truncated Nyström approximation (2.6) appears in the many papers, including [6, 12, 19, 34]. We have found (Sec. 5) that the truncation method (2.6) performs poorly in the present setting. This observation motivated us to search for more effective techniques.

Fixed-Rank Nyström Approximation: Proposal. The purpose of this paper is to develop, analyze, and evaluate a new approach for fixed-rank approximation of a psd matrix under the streaming model. We propose a more intuitive rank-r approximation:

$$\hat{\boldsymbol{A}}_r = [\![\hat{\boldsymbol{A}}^{\mathrm{nys}}]\!]_r. \tag{2.7}$$

That is, we report a best rank-r approximation of the full Nyström approximation (2.5).

This "matrix nearness" approach to fixed-rank approximation appears in the papers [22, 23, 36]. The combination with the Nyström method (2.5) seems totally natural. Even so, we were unable to find a reference after an exhaustive literature search and inquiries to experts on this subject.

Summary of Contributions. This paper contains a number of advances over the prior art:

1. We propose a distinct technique (2.7) for truncating the Nyström approximation to rank r. This formulation differs from earlier work on fixed-rank Nyström approximations.
2. We present a stable numerical implementation of (2.7) based on the best practices outlined in the paper [27]. This approach is essential for achieving high precision! (Sec. 3)
3. We establish informative error bounds for the method (2.7). In particular, we prove that it attains $(1 + \varepsilon)$-relative error in the Schatten 1-norm when $k = \Theta(r/\varepsilon)$. (Sec. 4)
4. We document numerical experiments on real and synthetic data to demonstrate that our method dominates existing techniques [19, 36] for fixed-rank psd approximation. (Sec. 5)

Psd matrix approximation is a ubiquitous problem, so we expect these results to have a broad impact.

Related Work. Randomized algorithms for **low-rank matrix approximation** were proposed in the late 1990s and developed into a technology in the 2000s; see [23, 30, 39] for more background. In the absence of constraints, such as streaming, we recommend the general-purpose methods from [23, 27].

Algorithms for low-rank matrix approximation in the important **streaming data** setting are discussed in [5, 8, 9, 16, 23, 36, 39, 40]. Few of these methods are designed for psd matrices.

Nyström methods for low-rank psd matrix approximation appear in [12, 14, 17, 18, 20, 23, 26, 34, 36, 38, 41]. These works mostly concern kernel matrices; they do not focus on the streaming model.

We are only aware of a few papers [17, 18, 20, 36] on algorithms for **psd matrix approximation** that operate under the **streaming model** (2.1). These papers form the comparison group.

Finally, let us mention two very recent **theoretical papers** [7, 33] that present existential results on algorithms for fixed-rank psd matrix approximation. The approach in [7] is only appropriate for sparse input matrices, while the work [33] is not valid in the streaming setting.

3 Implementation

Distributions for the Test Matrix. To ensure that the sketch is informative, we must draw the test matrix (2.2) at random from a suitable distribution. The choice of distribution determines the computational requirements for the sketch (2.3), the linear updates (2.4), and the matrix approximation (2.7). It also affects the quality of the approximation (2.7). Let us outline some of the most useful distributions. An exhaustive discussion is outside the scope of our work, but see [18, 20, 23, 29, 30, 36, 39].

Isotropic Models. Mathematically, the most natural model is to construct a test matrix $\boldsymbol{\Omega} \in \mathbb{F}^{n \times k}$ whose range is a uniformly random k-dimensional subspace in $\mathbb{F}^n$. There are two approaches:

1. **Gaussian.** Draw each entry of the matrix $\boldsymbol{\Omega} \in \mathbb{F}^{n \times k}$ independently at random from the standard normal distribution on $\mathbb{F}$.
2. **Orthonormal.** Draw a Gaussian matrix $\boldsymbol{G} \in \mathbb{F}^{n \times k}$, as above. Compute a thin orthogonal–triangular factorization $\boldsymbol{G} = \boldsymbol{\Omega}\boldsymbol{R}$ to obtain the test matrix $\boldsymbol{\Omega} \in \mathbb{F}^{n \times k}$. Discard $\boldsymbol{R}$.

Gaussian and orthonormal test matrices both require storage of kn floating-point numbers in $\mathbb{F}$ for the test matrix $\boldsymbol{\Omega}$ and another kn floating-point numbers for the sketch $\boldsymbol{Y}$. In both cases, the cost of multiplying a vector in $\mathbb{F}^n$ into $\boldsymbol{\Omega}$ is $\Theta(kn)$ floating-point operations.

For isotropic models, we can analyze the approximation (2.7) in detail. In exact arithmetic, Gaussian and isotropic test matrices yield identical Nyström approximations (Proposition A.2). In floating-point

Algorithm 1 *Sketch Initialization.* Implements (2.2)–(2.3) with a random orthonormal test matrix.

Input: Positive-semidefinite input matrix $\boldsymbol{A} \in \mathbb{F}^{n \times n}$; sketch size parameter k
Output: Constructs test matrix $\boldsymbol{\Omega} \in \mathbb{F}^{n \times k}$ and sketch $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{\Omega} \in \mathbb{F}^{n \times k}$

local: $\boldsymbol{\Omega}, \boldsymbol{Y}$ ▷ Internal variables for NYSTROMSKETCH
function NYSTROMSKETCH($\boldsymbol{A}; k$) ▷ Constructor
  if $\mathbb{F} = \mathbb{R}$ **then**
    $\boldsymbol{\Omega} \leftarrow \texttt{randn}(n, k)$
  if $\mathbb{F} = \mathbb{C}$ **then**
    $\boldsymbol{\Omega} \leftarrow \texttt{randn}(n, k) + \texttt{i} * \texttt{randn}(n, k)$
  $\boldsymbol{\Omega} \leftarrow \texttt{orth}(\boldsymbol{\Omega})$ ▷ Improve numerical stability
  $\boldsymbol{Y} \leftarrow \boldsymbol{A}\boldsymbol{\Omega}$

Algorithm 2 *Linear Update.* Implements (2.4).

Input: Scalars $\theta_1, \theta_2 \in \mathbb{R}$ and conjugate symmetric $\boldsymbol{H} \in \mathbb{F}^{n \times n}$
Output: Updates sketch to reflect linear innovation $\boldsymbol{A} \leftarrow \theta_1 \boldsymbol{A} + \theta_2 \boldsymbol{H}$

local: $\boldsymbol{\Omega}, \boldsymbol{Y}$ ▷ Internal variables for NYSTROMSKETCH
function LINEARUPDATE($\theta_1, \theta_2, \boldsymbol{H}$)
  $\boldsymbol{Y} \leftarrow \theta_1 \boldsymbol{Y} + \theta_2 \boldsymbol{H}\boldsymbol{\Omega}$

arithmetic, orthonormal matrices are more stable for large k, but we can generate Gaussian matrices with less arithmetic and communication. References for isotropic test matrices include [22, 23, 31].

Subsampled Scrambled Fourier Transform (SSFT). One shortcoming of the isotropic models is the cost of storing the test matrix and the cost of multiplying a vector into the test matrix. We can often reduce these costs using an SSFT test matrix. An SSFT takes the form

$$\boldsymbol{\Omega} = \boldsymbol{\Pi}_1 \boldsymbol{F} \boldsymbol{\Pi}_2 \boldsymbol{F} \boldsymbol{R} \in \mathbb{F}^{n \times k}. \tag{3.1}$$

The $\boldsymbol{\Pi}_i \in \mathbb{F}^{n \times n}$ are independent, signed permutation matrices,[1] chosen uniformly at random. The matrix $\boldsymbol{F} \in \mathbb{F}^{n \times n}$ is a discrete Fourier transform ($\mathbb{F} = \mathbb{C}$) or a discrete cosine transform ($\mathbb{F} = \mathbb{R}$). The matrix $\boldsymbol{R} \in \mathbb{F}^{n \times k}$ is a restriction to k coordinates, chosen uniformly at random.

An SSFT $\boldsymbol{\Omega}$ requires only $\Theta(n)$ storage, but the sketch $\boldsymbol{Y}$ still requires storage of kn numbers. We can multiply a vector in $\mathbb{F}^n$ into $\boldsymbol{\Omega}$ using $\Theta(n \log n)$ arithmetic operations via an FFT or FCT algorithm. Thus, for most choices of sketch size k, the SSFT improves over the isotropic models.

In practice, the SSFT yields matrix approximations whose quality is identical to those we obtain with an isotropic test matrix (Sec. 5). Although the analysis for SSFTs is less complete, the empirical evidence confirms that the theory for isotropic models also offers excellent guidance for SSFTs. References for SSFTs and related test matrices include [1, 3, 10, 23, 29, 35, 40].

Numerically Stable Implementation. It requires care to compute the fixed-rank approximation (2.7). App. B shows that a poor implementation may produce an approximation with 100% error!

Let us outline a numerically stable and very accurate implementation of (2.7), based on an idea from [27, 37]. Fix a small parameter $\nu > 0$. Instead of approximating the psd matrix $\boldsymbol{A}$ directly, we approximate the shifted matrix $\boldsymbol{A}_\nu = \boldsymbol{A} + \nu\mathbf{I}$ and then remove the shift. Here are the steps:

1. Construct the shifted sketch $\boldsymbol{Y}_\nu = \boldsymbol{Y} + \nu\boldsymbol{\Omega}$.
2. Form the matrix $\boldsymbol{B} = \boldsymbol{\Omega}^* \boldsymbol{Y}_\nu$.
3. Compute a Cholesky decomposition $\boldsymbol{B} = \boldsymbol{C}\boldsymbol{C}^*$.
4. Compute $\boldsymbol{E} = \boldsymbol{Y}_\nu \boldsymbol{C}^{-1}$ by back-substitution.
5. Compute the (thin) singular value decomposition $\boldsymbol{E} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$.
6. Form $\hat{\boldsymbol{A}}_r = \boldsymbol{U}[\![\boldsymbol{\Sigma}^2 - \nu\mathbf{I}]\!]_r \boldsymbol{U}^*$.

The pseudocode addresses some additional implementation details. Related, but distinct, methods were proposed by Williams & Seeger [38] and analyzed in Gittens's thesis [18].

[1] A signed permutation has exactly one nonzero entry in each row and column; the nonzero has modulus one.

Algorithm 3 *Fixed-Rank PSD Approximation.* Implements (2.7).

Input: Matrix $\boldsymbol{A}$ in sketch must be psd; rank parameter $1 \leq r \leq k$
Output: Returns factors $\boldsymbol{U} \in \mathbb{F}^{n \times r}$ with orthonormal columns and nonnegative, diagonal $\boldsymbol{\Lambda} \in \mathbb{F}^{r \times r}$ that form a rank-r psd approximation $\hat{\boldsymbol{A}}_r = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^*$ of the sketched matrix $\boldsymbol{A}$

local: $\boldsymbol{\Omega}, \boldsymbol{Y}$ ▷ Internal variables for NYSTROMSKETCH
function FIXEDRANKPSDAPPROX(r)
  $\nu \leftarrow \mu \, \texttt{norm}(\boldsymbol{Y})$ ▷ $\mu = 2.2 \cdot 10^{-16}$ in double precision
  $\boldsymbol{Y} \leftarrow \boldsymbol{Y} + \nu\boldsymbol{\Omega}$ ▷ Sketch of shifted matrix $\boldsymbol{A} + \nu\mathbf{I}$
  $\boldsymbol{B} \leftarrow \boldsymbol{\Omega}^*\boldsymbol{Y}$
  $\boldsymbol{C} \leftarrow \texttt{chol}((\boldsymbol{B} + \boldsymbol{B}^*)/2)$ ▷ Force symmetry
  $(\boldsymbol{U}, \boldsymbol{\Sigma}, \sim) \leftarrow \texttt{svd}(\boldsymbol{Y}/\boldsymbol{C}, \texttt{'econ'})$ ▷ Solve least squares problem; form thin SVD
  $\boldsymbol{U} \leftarrow \boldsymbol{U}(\texttt{:, 1:r})$ and $\boldsymbol{\Sigma} \leftarrow \boldsymbol{\Sigma}(\texttt{1:r, 1:r})$ ▷ Truncate to rank r
  $\boldsymbol{\Lambda} \leftarrow \max\{0, \boldsymbol{\Sigma}^2 - \nu\mathbf{I}\}$ ▷ Square to get eigenvalues; remove shift
  return $(\boldsymbol{U}, \boldsymbol{\Lambda})$

Pseudocode. We present detailed pseudocode for the sketch (2.2)–(2.4) and the implementation of the fixed-rank psd approximation (2.7) described above. For simplicity, we only elaborate the case of a random orthonormal test matrix; we have also developed an SSFT implementation for empirical testing. The pseudocode uses both mathematical notation and MATLAB 2017A functions.

Algorithms and Computational Costs. Algorithm 1 constructs a random orthonormal test matrix, and computes the sketch (2.3) of an input matrix. The test matrix and sketch require the storage of $2kn$ floating-point numbers. Owing to the orthogonalization step, the construction of the test matrix requires $\Theta(k^2 n)$ floating-point operations. For a general input matrix, the sketch requires $\Theta(kn^2)$ floating-point operations; this cost can be removed by initializing the input matrix to zero.

Algorithm 2 implements the linear update (2.4) to the sketch. Nominally, the computation requires $\Theta(kn^2)$ arithmetic operations, but this cost can be reduced when $\boldsymbol{H}$ has structure (e.g., low rank). Using the SSFT test matrix (3.1) also reduces this cost.

Algorithm 3 computes the rank-r psd approximation (2.7). This method requires additional storage of $\Theta(kn)$. The arithmetic cost is $\Theta(k^2 n)$ operations, which is dominated by the SVD of the matrix $\boldsymbol{E}$.

4 Theoretical Results

Relative Error Bound. Our first result is an accurate bound for the expected Schatten 1-norm error in the fixed-rank psd approximation (2.7).

Theorem 4.1 (Fixed-Rank Nyström: Relative Error)**.** *Assume $1 \leq r < k \leq n$. Let $\boldsymbol{A} \in \mathbb{F}^{n \times n}$ be a psd matrix. Draw a test matrix $\boldsymbol{\Omega} \in \mathbb{F}^{n \times k}$ from the Gaussian or orthonormal distribution, and form the sketch $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{\Omega}$. Then the approximation $\hat{\boldsymbol{A}}_r$ given by* (2.5) *and* (2.7) *satisfies*

$$\mathbb{E}\,\|\boldsymbol{A} - \hat{\boldsymbol{A}}_r\|_1 \leq \left(1 + \frac{r}{k - r - \alpha}\right) \cdot \|\boldsymbol{A} - [\![\boldsymbol{A}]\!]_r\|_1; \tag{4.1}$$

$$\mathbb{E}\,\|\boldsymbol{A} - \hat{\boldsymbol{A}}_r\|_\infty \leq \|\boldsymbol{A} - [\![\boldsymbol{A}]\!]_r\|_\infty + \frac{r}{k - r - \alpha} \cdot \|\boldsymbol{A} - [\![\boldsymbol{A}]\!]_r\|_1. \tag{4.2}$$

The quantity $\alpha(\mathbb{R}) = 1$ and $\alpha(\mathbb{C}) = 0$. Similar results hold with high probability.

The proof of Theorem 4.1 appears in App. A.

In contrast to previous analyses of Nyström methods, Theorem 4.1 yields explicit, sharp constants. As a consequence, the formulae (4.1)–(4.2) offer an *a priori* mechanism for selecting the sketch size k to achieve a desired error bound. In particular, for each $\varepsilon > 0$,

$$k = (1 + \varepsilon^{-1})r + \alpha \quad \text{implies} \quad \mathbb{E}\,\|\boldsymbol{A} - \hat{\boldsymbol{A}}_r\|_1 \leq (1 + \varepsilon) \cdot \|\boldsymbol{A} - [\![\boldsymbol{A}]\!]_r\|_1.$$

Thus, we can attain an arbitrarily small relative error in the Schatten 1-norm. In the streaming setting, the scaling $k = \Theta(r/\varepsilon)$ is optimal for this result [15, Thm. 4.2]. Furthermore, it is impossible [39, Sec. 6.2] to obtain "pure" relative error bounds in the Schatten ∞-norm unless $k = \Omega(n)$.

The Role of Spectral Decay. To circumvent these limitations, it is necessary to develop a different kind of error bound. Our second result shows that the fixed-rank psd approximation (2.7) automatically exploits decay in the spectrum of the input matrix.

Theorem 4.2 (Fixed-Rank Nyström: Spectral Decay)**.** *Instate the notation and assumptions of Theorem 4.1. Then*

$$\mathbb{E}\,\|\boldsymbol{A} - \hat{\boldsymbol{A}}_r\|_1 \le \|\boldsymbol{A} - [\![\boldsymbol{A}]\!]_r\|_1 + 2 \min_{\varrho < k-\alpha} \left[\left(1 + \frac{\varrho}{k - \varrho - \alpha}\right) \cdot \|\boldsymbol{A} - [\![\boldsymbol{A}]\!]_\varrho\|_1 \right]; \tag{4.3}$$

$$\mathbb{E}\,\|\boldsymbol{A} - \hat{\boldsymbol{A}}_r\|_\infty \le \|\boldsymbol{A} - [\![\boldsymbol{A}]\!]_r\|_\infty + 2 \min_{\varrho < k-\alpha} \left[\left(1 + \frac{\varrho}{k - \varrho - \alpha}\right) \cdot \|\boldsymbol{A} - [\![\boldsymbol{A}]\!]_\varrho\|_1 \right]. \tag{4.4}$$

The index ϱ ranges over the natural numbers.

The proof of Theorem 4.2 appears in App. A.

Here is one way to understand this result. As the index ϱ increases, the quantity $\varrho/(k-\varrho-\alpha)$ increases while the rank-ϱ approximation error decreases. Theorem 4.2 states that the approximation (2.7) automatically achieves the best tradeoff between these two terms. When the spectrum of $\boldsymbol{A}$ decays, the rank-ϱ approximation error may be far smaller than the rank-r approximation error. In this case, Theorem 4.2 is tighter than Theorem 4.1, although the prediction is more qualitative.

Additional Results. The proofs can be extended to obtain high-probability bounds, as well as results for other Schatten norms or for other test matrices (App. A).

5 Numerical Performance

Experimental Setup. In many streaming applications, such as [42], it is essential that the sketch uses as little memory as possible and that the psd approximation achieves the best possible error. For the methods we consider, the arithmetic costs of linear updates and psd approximation are roughly comparable. Therefore, we only assess storage and accuracy.

For the numerical experiments, the field $\mathbb{F} = \mathbb{C}$ except when noted explicitly. Choose a psd input matrix $\boldsymbol{A} \in \mathbb{F}^{n\times n}$ and a target rank r. Then fix a sketch size parameter k with $r \le k \le n$. For each trial, draw the test matrix $\boldsymbol{\Omega}$ from the orthonormal or the SSFT distribution, and form the sketch $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{\Omega}$ of the input matrix. Using Algorithm 3, compute the rank-r psd approximation $\hat{\boldsymbol{A}}_r$ defined in (2.7). We evaluate the performance using the relative error metric:

$$\text{Schatten } p\text{-norm relative error} \quad = \quad \frac{\|\boldsymbol{A} - \hat{\boldsymbol{A}}_r\|_p}{\|\boldsymbol{A} - [\![\boldsymbol{A}]\!]_r\|_p} - 1. \tag{5.1}$$

We perform 20 independent trials and report the average error.

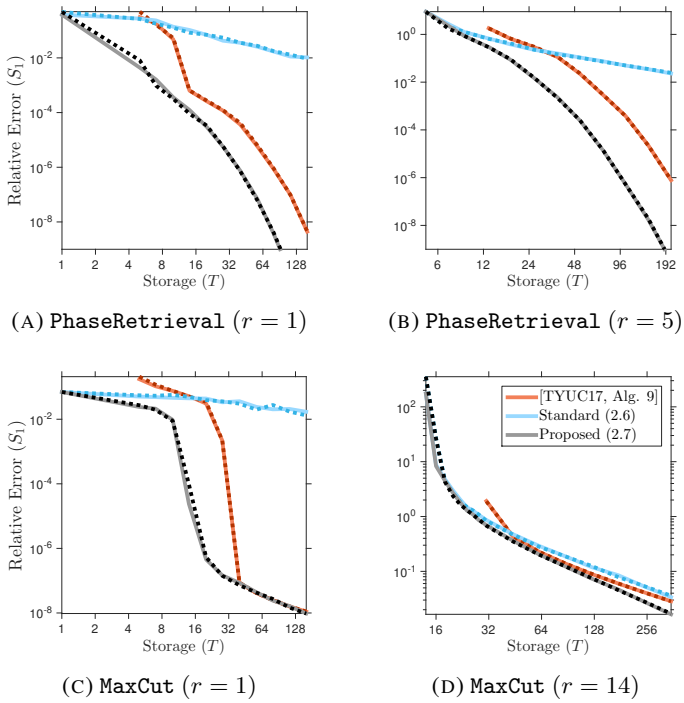
We compare our method (2.7) with the standard truncated Nyström approximation (2.6); the best reference for this type of approach is [19, Sec. 2.2]. The approximation (2.6) is constructed from the same sketch as (2.7), so the experimental procedure is identical.

We also consider the sketching method and psd approximation algorithm [36, Alg. 9] based on earlier work from [8, 23, 40]. We implemented this sketch with orthonormal matrices and also with SSFT matrices. The sketch has two different parameters (k, ℓ), so we select the parameters that result in the minimum relative error. Otherwise, the experimental procedure is the same.

We apply the methods to representative input matrices; see Figure B.1 for the spectra.

Synthetic Examples. The synthetic examples are **diagonal** with dimension $n = 10^3$; results for larger and non-diagonal matrices are similar. These matrices are parameterized by an effective rank parameter R, which takes values in $\{5, 10, 20\}$. We compute approximations with rank $r = 10$.

1. **Low-Rank + PSD Noise.** These matrices take the form
$$\boldsymbol{A} = \operatorname{diag}(\underbrace{1, \dots, 1}_{R}, 0, \dots, 0) + \xi n^{-1} \boldsymbol{W} \in \mathbb{F}^{n\times n}.$$
The matrix $\boldsymbol{W} \in \mathbb{F}^{n\times n}$ has the $\textsc{Wishart}(n, n; \mathbb{F})$ distribution; that is, $\boldsymbol{W} = \boldsymbol{G}\boldsymbol{G}^*$ where $\boldsymbol{G} \in \mathbb{F}^{n\times n}$ is standard normal. The parameter ξ controls the signal-to-noise ratio. We consider three examples: `LowRankLowNoise` ($\xi = 10^{-4}$), `LowRankMedNoise` ($\xi = 10^{-2}$), `LowRankHiNoise` ($\xi = 10^{-1}$).



(A) `PhaseRetrieval` ($r = 1$) (B) `PhaseRetrieval` ($r = 5$)

(C) `MaxCut` ($r = 1$) (D) `MaxCut` ($r = 14$)

FIGURE 5.1: **Application Examples, Approximation Rank r, Schatten 1-Norm Error.** The data series show the performance of three algorithms for rank-r psd approximation. **Solid lines** are generated from the Gaussian sketch; **dashed lines** are from the SSFT sketch. Each panel displays the Schatten 1-norm relative error (5.1) as a function of storage cost T. See Sec. 5 for details.

2. **Polynomial Decay.** These matrices take the form
$$\boldsymbol{A} = \operatorname{diag}(\underbrace{1, \dots, 1}_{R}, 2^{-p}, 3^{-p}, \dots, (n-R+1)^{-p}) \in \mathbb{F}^{n \times n}.$$
The parameter $p > 0$ controls the rate of polynomial decay. We consider three examples: `PolyDecaySlow` ($p = 0.5$), `PolyDecayMed` ($p = 1$), `PolyDecayFast` ($p = 2$).
3. **Exponential Decay.** These matrices take the form
$$\boldsymbol{A} = \operatorname{diag}(\underbrace{1, \dots, 1}_{R}, 10^{-q}, 10^{-2q}, \dots, 10^{-(n-R)q}) \in \mathbb{F}^{n \times n}.$$
The parameter $q > 0$ controls the rate of exponential decay. We consider three examples: `ExpDecaySlow` ($q = 0.1$), `ExpDecayMed` ($q = 0.25$), `ExpDecayFast` ($q = 1$).

Application Examples. We also consider **non-diagonal** matrices inspired by the SDP algorithm [42].

1. `MaxCut`: This is a **real-valued** psd matrix with dimension $n = 2\,000$, and its effective rank $R = 14$. We form approximations with rank $r \in \{1, 14\}$. The matrix is an approximate solution to the MAXCUT SDP [21] for the sparse graph `G40` [11].
2. `PhaseRetrieval`: This is a psd matrix with dimension $n = 25\,921$. It has exact rank 250, but its effective rank $R = 5$. We form approximations with rank $r \in \{1, 5\}$. The matrix is an approximate solution to a phase retrieval SDP; it was provided by the authors of [42].

Experimental Results. Figures 5.1–5.2 display the performance of the three fixed-rank psd approximation methods for a subcollection of the input matrices. The vertical axis is the Schatten 1-norm relative error (5.1). The variable T on the horizontal axis is proportional to the storage required for the sketch only. For the Nyström-based approximations (2.6)–(2.7), we have the correspondence $T = k$. For the approximation [36, Alg. 9], we set $T = k + \ell$.

The experiments demonstrate that the proposed method (2.7) has a significant benefit over the alternatives for input matrices that admit a good low-rank approximation. It equals or improves on the

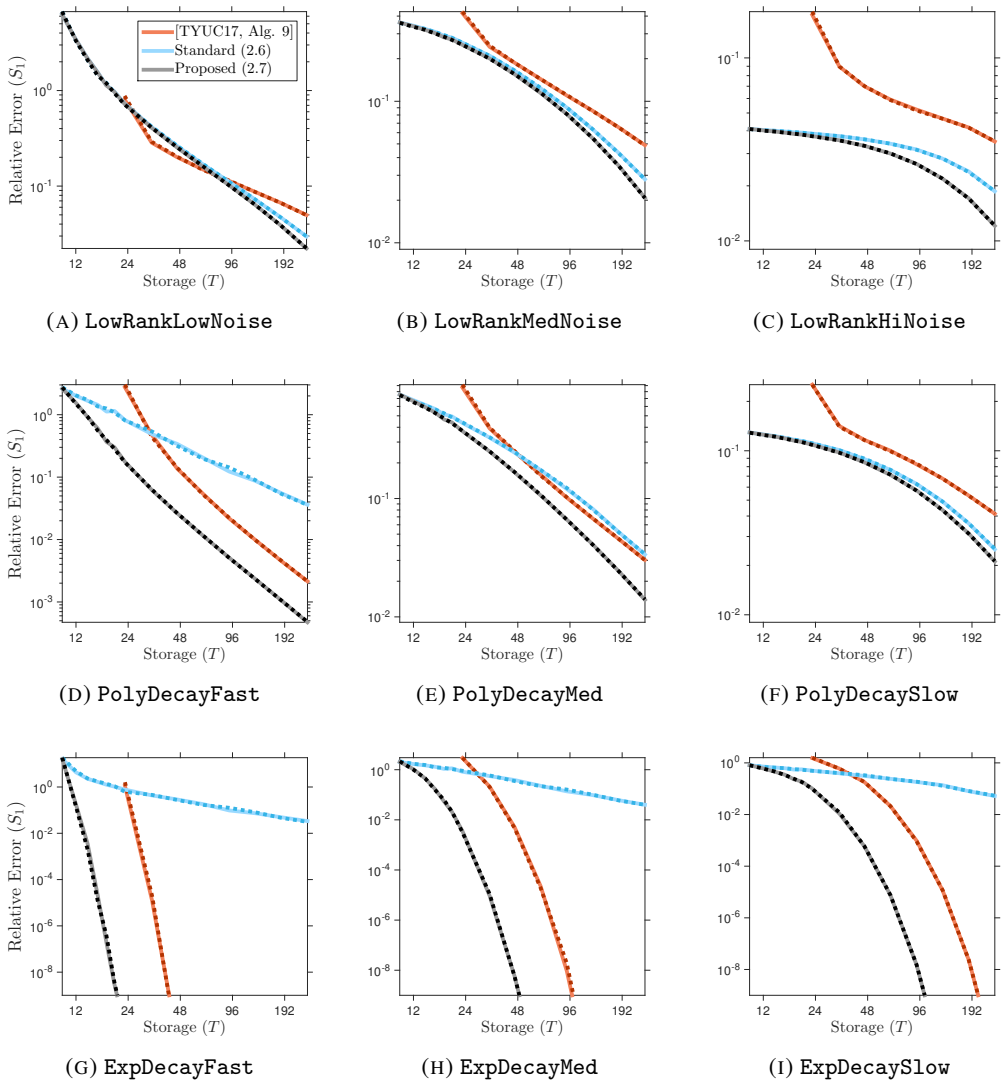


FIGURE 5.2: **Synthetic Examples with Effective Rank $R = 10$, Approximation Rank $r = 10$, Schatten 1-Norm Error.** The data series show the performance of three algorithms for rank-r psd approximation with $r = 10$. **Solid lines** are generated from the Gaussian sketch; **dashed lines** are from the SSFT sketch. Each panel displays the Schatten 1-norm relative error (5.1) as a function of storage cost T.

competitors for almost all other examples and storage budgets. App. B contains additional numerical results; these experiments only reinforce the message of Figures 5.1–5.2.

Conclusions. This paper makes the case for using the proposed fixed-rank psd approximation (2.7) in lieu of the alternatives (2.6) or [36, Alg. 9]. Theorem 4.1 shows that the proposed fixed-rank psd approximation (2.7) can attain any prescribed relative error, and Theorem 4.2 shows that it can exploit spectral decay. Furthermore, our numerical work demonstrates that the proposed approximation improves (almost) uniformly over the competitors for a range of examples. These results are timely because of the recent arrival of compelling applications, such as [42], for sketching psd matrices.

Acknowledgments. The authors wish to thank Mark Tygert and Alex Gittens for helpful feedback on preliminary versions of this work. JAT gratefully acknowledges partial support from ONR Award N00014-17-1-2146 and the Gordon & Betty Moore Foundation. VC and AY were supported in part by the European Commission under Grant ERC Future Proof, SNF 200021-146750, and SNF CRSII2-147633. MU was supported in part by DARPA Award FA8750-17-2-0101.

A Details of the Theoretical Analysis

This appendix contains a new theoretical analysis of the simple Nyström approximation (2.5) and the proposed fixed-rank Nyström approximation (2.7).

A.1 Best Approximation in Schatten Norms

Let us introduce compact notation for the optimal rank-r approximation error in the Schatten p-norm:

$$\sigma_{r+1}^{(p)}(\boldsymbol{M}) = \|\boldsymbol{M} - [\![\boldsymbol{M}]\!]_r\|_p = \Big[\sum\nolimits_{i>r} \sigma_i(\boldsymbol{M})^p\Big]^{1/p}. \tag{A.1}$$

Ordinary singular values correspond to the case $p = \infty$.

A.2 Analysis of the Nyström Approximation

The first result gives a very accurate error bound for the basic Nyström approximation $\hat{\boldsymbol{A}}^{\mathrm{nys}}$ with respect to the Schatten 1-norm. This estimate is the key ingredient in the proof of Theorem 4.2.

Theorem A.1 (Error in Nyström Approximation)**.** Assume $1 \le r \le k \le n$. Let $\boldsymbol{A} \in \mathbb{F}^{n\times n}$ be a psd matrix. Draw the test matrix $\boldsymbol{\Omega} \in \mathbb{F}^{n\times k}$ from the Gaussian or orthonormal distribution, and form the sketch $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{\Omega}$. Then the rank-$k$ Nyström approximation $\hat{\boldsymbol{A}}^{\mathrm{nys}}$ determined by (2.5) satisfies the error bound

$$\mathbb{E}\,\|\boldsymbol{A} - \hat{\boldsymbol{A}}^{\mathrm{nys}}\|_1 \le \min_{\varrho<k-\alpha} \left[\left(1 + \frac{\varrho}{k-\varrho-\alpha}\right)\sigma_{\varrho+1}^{(1)}(\boldsymbol{A})\right]. \tag{A.2}$$

The index ϱ ranges over natural numbers. The quantity $\alpha(\mathbb{R}) = 1$ and $\alpha(\mathbb{C}) = 0$. The optimal rank-ϱ Schatten 1-norm approximation error is defined in (A.1).

To the best of our knowledge, Theorem A.1 is new. The proof appears below in App. A.3.

Let us situate Theorem A.1 with respect to the results in Gittens's work [18, 20]. Gittens develops error bounds for the Nyström approximation (2.5) that hold with high probability, rather than in expectation. He measures errors in the Schatten p-norm for $p = 1, 2, \infty$. He also obtains results for several types of test matrices, including isotropic models and a relative of the SSFT. In contrast to Theorem A.1, Gittens's bounds are more complicated, and the constants are much larger.

A.3 Proof of Theorem A.1

We begin with the proof of Theorem A.1. Gittens [17, 18, 20] uses a related argument to obtain bounds on the *probability* that the Nyström approximation achieves a given error.

The first step is to write the Nyström approximation in terms of an orthogonal projector. This expression allows us to exploit the analysis from [23, 36].

Proposition A.2 (Representation of Nyström Approximation)**.** *Let $\boldsymbol{P}$ be the orthogonal projector onto* $\mathrm{range}(\boldsymbol{A}^{1/2}\boldsymbol{\Omega})$*:*

$$\boldsymbol{P} = (\boldsymbol{A}^{1/2}\boldsymbol{\Omega})(\boldsymbol{\Omega}^*\boldsymbol{A}\boldsymbol{\Omega})^\dagger(\boldsymbol{A}^{1/2}\boldsymbol{\Omega})^*. \tag{A.3}$$

Then the Nyström approximation (2.5) *can be expressed as*

$$\hat{\boldsymbol{A}}^{\mathrm{nys}} = \boldsymbol{A}^{1/2}\boldsymbol{P}\boldsymbol{A}^{1/2} \tag{A.4}$$

In particular, the Nyström approximation only depends on $\boldsymbol{\Omega}$ through $\mathrm{range}(\boldsymbol{\Omega})$.

We believe that Proposition A.2 first appeared explicitly in the work of Gittens [17].

Proof. This argument follows from a direct calculation:

$$\begin{aligned}
\hat{\boldsymbol{A}}^{\mathrm{nys}} &= \boldsymbol{A}\boldsymbol{\Omega}(\boldsymbol{\Omega}^*\boldsymbol{A}\boldsymbol{\Omega})^\dagger\boldsymbol{\Omega}^*\boldsymbol{A} \\
&= \boldsymbol{A}^{1/2}(\boldsymbol{A}^{1/2}\boldsymbol{\Omega})\big[(\boldsymbol{A}^{1/2}\boldsymbol{\Omega})^*(\boldsymbol{A}^{1/2}\boldsymbol{\Omega})\big]^\dagger(\boldsymbol{A}^{1/2}\boldsymbol{\Omega})^*\boldsymbol{A}^{1/2} \\
&= \boldsymbol{A}^{1/2}\boldsymbol{P}\boldsymbol{A}^{1/2}.
\end{aligned}$$

To reach the last line, we identified the orthogonal projector (A.3). □

With Proposition A.2 at hand, the proof of Theorem A.1 is straightforward.

We may assume that $\boldsymbol{\Omega}$ is a Gaussian matrix because the reconstruction $\hat{\boldsymbol{A}}$ only depends on $\mathrm{range}(\boldsymbol{\Omega})$. The range of a random orthonormal matrix has the same distribution as a Gaussian matrix up to a set of measure zero.

Let $\boldsymbol{P}$ be the orthogonal projector (A.3). In view of the formula (A.4) for $\hat{\boldsymbol{A}}^{\mathrm{nys}}$, we have

$$\boldsymbol{A} - \hat{\boldsymbol{A}}^{\mathrm{nys}} = \boldsymbol{A}^{1/2}(\mathbf{I} - \boldsymbol{P})\boldsymbol{A}^{1/2}. \tag{A.5}$$

We can now express the Schatten 1-norm of the error in terms of the Schatten 2-norm:

$$\|\boldsymbol{A} - \hat{\boldsymbol{A}}^{\mathrm{nys}}\|_1 = \|\boldsymbol{A}^{1/2}(\mathbf{I} - \boldsymbol{P})(\mathbf{I} - \boldsymbol{P})\boldsymbol{A}^{1/2}\|_1 = \|(\mathbf{I} - \boldsymbol{P})\boldsymbol{A}^{1/2}\|_2^2.$$

The first identity follows from (A.5) and the fact that the orthogonal projector $\mathbf{I} - \boldsymbol{P}$ is idempotent.

Fix a natural number $\varrho < k - \alpha$. We can use established results from the literature to control the expectation of the error. In particular, we invoke a slight generalization [36, Fact. 8.3] of a result [23, Thm. 10.5] of Halko et al. We arrive at the bound

$$\begin{aligned}\mathbb{E}\,\|(\mathbf{I} - \boldsymbol{P})\boldsymbol{A}^{1/2}\|_2^2 &\le \left(1 + \frac{\varrho}{k - \varrho - \alpha}\right) \sum_{i > \varrho} \sigma_i(\boldsymbol{A}^{1/2})^2 \\ &= \left(1 + \frac{\varrho}{k - \varrho - \alpha}\right) \sum_{i > \varrho} \sigma_i(\boldsymbol{A}) = \left(1 + \frac{\varrho}{k - \varrho - \alpha}\right) \sigma_{\varrho+1}^{(1)}(\boldsymbol{A}).\end{aligned}$$

Combine the last two displays and minimize over eligible ϱ to complete the argument.

Remark A.3 (Spectral-Norm Error)**.** When $\mathbb{F} = \mathbb{R}$, we can also obtain a spectral-norm error bound by combining this argument with another result [23, Thm. 10.6] of Halko et al.:

$$\mathbb{E}\,\sqrt{\|\boldsymbol{A} - \hat{\boldsymbol{A}}^{\mathrm{nys}}\|} \le \min_{\varrho < k-1} \left[\left(1 + \sqrt{\frac{\varrho}{k - \varrho - 1}}\right) \sqrt{\sigma_{\varrho+1}(\boldsymbol{A})} + \frac{\mathrm{e}\sqrt{k}}{k - \varrho} \sqrt{\sigma_{\varrho+1}^{(1)}(\boldsymbol{A})}\right].$$

It takes a surprising amount of additional work to obtain an accurate bound for the first moment of the error (instead of the 1/2 moment). We have chosen not to include this argument.

Remark A.4 (High-Probability Bounds)**.** As noted by Gittens [18, 20], we can obtain high-probability error bounds in the real setting by combining the approach here with results [23, Thms. 10.7–10.8] from Halko et al. We omit the details.

Remark A.5 (Other Test Matrices)**.** As noted by Gittens [18, 20], we can obtain results for other types of test matrices by replacing parts of the analysis that depend on Gaussian matrices. These changes result in bounds that are quantitatively and qualitatively worse. The numerical evidence suggests that many types of test matrices have the same empirical performance, so we omit this development.

A.4 Theorem 4.2: Schatten 1-Norm Bound

Let us continue with the proof of the Schatten 1-norm bound (4.3) from Theorem 4.2. We require a basic result on rank-r approximation adapted from [36, Prop. 7.1].

Proposition A.6 (Fixed-Rank Projection)**.** *Let $\boldsymbol{A} \in \mathbb{F}^{n \times n}$ and $\hat{\boldsymbol{A}} \in \mathbb{F}^{n \times n}$ be arbitrary matrices. For each natural number r and number $p \in [1, \infty]$,*

$$\|\boldsymbol{A} - [\![\hat{\boldsymbol{A}}]\!]_r\|_p \le \sigma_{r+1}^{(p)}(\boldsymbol{A}) + 2\|\boldsymbol{A} - \hat{\boldsymbol{A}}\|_p.$$

Proof. The argument follows from a short calculation based on the triangle inequality:

$$\begin{aligned}\|\boldsymbol{A} - [\![\hat{\boldsymbol{A}}]\!]_r\|_p &\le \|\boldsymbol{A} - \hat{\boldsymbol{A}}\|_p + \|\hat{\boldsymbol{A}} - [\![\hat{\boldsymbol{A}}]\!]_r\|_p \\ &\le \|\boldsymbol{A} - \hat{\boldsymbol{A}}\|_p + \|\hat{\boldsymbol{A}} - [\![\boldsymbol{A}]\!]_r\|_p \\ &\le 2\|\boldsymbol{A} - \hat{\boldsymbol{A}}\|_p + \|\boldsymbol{A} - [\![\boldsymbol{A}]\!]_r\|_p.\end{aligned}$$

In the second line, we have used the fact that $[\![\hat{\boldsymbol{A}}]\!]_r$ is a *best* rank-r approximation of $\hat{\boldsymbol{A}}$. To complete the argument, we identify the last term (A.1) as the best rank-r approximation error in the Schatten p-norm. □

The bound (4.3) from Theorem 4.2 is now an immediate consequence of Theorem A.1 and Proposition A.6:

$$\begin{aligned}\mathbb{E}\,\|\boldsymbol{A} - \hat{\boldsymbol{A}}_r\|_1 &\le \sigma_{r+1}^{(1)}(\boldsymbol{A}) + 2\,\mathbb{E}\,\|\boldsymbol{A} - \hat{\boldsymbol{A}}^{\mathrm{nys}}\|_1 \\ &\le \sigma_{r+1}^{(1)}(\boldsymbol{A}) + 2 \min_{\varrho < k-\alpha} \left(1 + \frac{\varrho}{k - \varrho - \alpha}\right) \sigma_{\varrho+1}^{(1)}(\boldsymbol{A}).\end{aligned}$$

We have used the definition (2.7) of our fixed-rank approximation: $\hat{\boldsymbol{A}}_r = [\![\hat{\boldsymbol{A}}^{\mathrm{nys}}]\!]_r$.

Remark A.7 (Extensions)**.** Given a bound on the error in the Nyström approximation (2.5) in the Schatten p-norm for any test matrix, this approach automatically yields an estimate for the associated fixed-rank psd approximation (2.7).

A.5 Theorem 4.1: Schatten 1-Norm Bound

Next, we turn to the proof of the Schatten 1-norm bound (4.1) from Theorem 4.1. This argument is based on the same approach as Theorem A.1, but we require several additional ingredients from [18, 22, 23, 36].

As before, we may assume that $\boldsymbol{\Omega}$ is Gaussian. With probability one, the nonzero eigenvalues of $\hat{\boldsymbol{A}}^{\mathrm{nys}}$ are all distinct, so the best rank-r approximation $\hat{\boldsymbol{A}}_r$ of $\hat{\boldsymbol{A}}^{\mathrm{nys}}$ is determined uniquely.

Let $\boldsymbol{P}$ be the orthogonal projector (A.3). According to (A.4), the Nyström approximation takes the form

$$\hat{\boldsymbol{A}}^{\mathrm{nys}} = \boldsymbol{A}^{1/2}\boldsymbol{P}\boldsymbol{A}^{1/2} = (\boldsymbol{A}^{1/2}\boldsymbol{P})(\boldsymbol{P}\boldsymbol{A}^{1/2}).$$

Let $\boldsymbol{Q}$ denote the orthogonal projector onto the range of $[\![\boldsymbol{P}\boldsymbol{A}^{1/2}]\!]_r$. Using the (truncated) SVD of the matrix $\boldsymbol{P}\boldsymbol{A}^{1/2}$, we can verify that the best rank-r approximation $\hat{\boldsymbol{A}}_r$ of $\hat{\boldsymbol{A}}^{\mathrm{nys}}$ satisfies

$$\hat{\boldsymbol{A}}_r = [\![\boldsymbol{A}^{1/2}\boldsymbol{P}]\!]_r [\![\boldsymbol{P}\boldsymbol{A}^{1/2}]\!]_r = \boldsymbol{A}^{1/2}\boldsymbol{P}\boldsymbol{Q}\boldsymbol{P}\boldsymbol{A}^{1/2}$$

As in the proof of Theorem A.1, the Schatten 1-norm of the error satisfies

$$\|\boldsymbol{A} - \hat{\boldsymbol{A}}_r\|_1 = \|\boldsymbol{A} - \boldsymbol{A}^{1/2}\boldsymbol{P}\boldsymbol{Q}\boldsymbol{P}\boldsymbol{A}^{1/2}\|_1 = \|(\mathbf{I} - \boldsymbol{Q}\boldsymbol{P})\boldsymbol{A}^{1/2}\|_2^2.$$

Since $\mathrm{range}(\boldsymbol{Q}) \subset \mathrm{range}(\boldsymbol{P})$, we can rewrite this expression as

$$\|(\mathbf{I} - \boldsymbol{Q}\boldsymbol{P})\boldsymbol{A}^{1/2}\|_2^2 = \|(\mathbf{I} - \boldsymbol{P}\boldsymbol{Q}\boldsymbol{P})\boldsymbol{A}^{1/2}\|_2^2 = \|\boldsymbol{A}^{1/2} - \boldsymbol{P}[\![\boldsymbol{P}\boldsymbol{A}^{1/2}]\!]_r\|_2^2.$$

The last identity holds because $\boldsymbol{Q}\boldsymbol{P}\boldsymbol{A}^{1/2} = [\![\boldsymbol{P}\boldsymbol{A}^{1/2}]\!]_r$. A direct application of Gu's result [22, Thm. 3.5] yields

$$\|\boldsymbol{A}^{1/2} - \boldsymbol{P}[\![\boldsymbol{P}\boldsymbol{A}^{1/2}]\!]_r\|_2^2 \le \|(\mathbf{I} - \boldsymbol{P})[\![\boldsymbol{A}^{1/2}]\!]_r\|_2^2 + \sum_{i>r} \sigma_i(\boldsymbol{A}^{1/2})^2.$$

A direct application of the result [36, Prop. 9.2] shows that

$$\mathbb{E}\,\|(\mathbf{I} - \boldsymbol{P})[\![\boldsymbol{A}^{1/2}]\!]_r\|_2^2 = \frac{r}{k - r - \alpha} \sum_{i>r} \sigma_i(\boldsymbol{A}^{1/2})^2.$$

As before, we note that

$$\sum_{i>r} \sigma_i(\boldsymbol{A}^{1/2})^2 = \sigma_{r+1}^{(1)}(\boldsymbol{A}).$$

Taking an expectation and sequencing these displays, we arrive at

$$\mathbb{E}\,\|\boldsymbol{A} - \hat{\boldsymbol{A}}_r\|_1 \le \left(1 + \frac{r}{k - r - \alpha}\right) \sigma_{r+1}^{(1)}(\boldsymbol{A}).$$

This is the stated result (4.1).

A.6 Theorems 4.1 and 4.2: Schatten ∞-Norm Bounds

Last, we develop the bounds (4.2) and (4.4) on the Schatten ∞-norm of the fixed-rank psd approximation (2.7) using a formal argument. We require the following result.

Proposition A.8 (Reversed Eckart–Young)**.** *Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{F}^{n\times n}$ be matrices, and assume that $\operatorname{rank}(\boldsymbol{B}) \leq r$. Then*

$$\|\boldsymbol{A} - \boldsymbol{B}\|_\infty \leq \sigma_{r+1}(\boldsymbol{A}) + \left[\|\boldsymbol{A} - \boldsymbol{B}\|_1 - \sigma_{r+1}^{(1)}(\boldsymbol{A})\right].$$

The proof of Proposition A.8 follows from a minor change to [22, Thm. 3.4].

Proof. As a consequence of Weyl's inequalities [2, Thm. III.2.1], we have the bound

$$\sigma_{i+r}(\boldsymbol{A}) \leq \sigma_i(\boldsymbol{A} - \boldsymbol{B}) + \sigma_{r+1}(\boldsymbol{B}) = \sigma_i(\boldsymbol{A} - \boldsymbol{B}). \tag{A.6}$$

The last identity holds because $\operatorname{rank}(\boldsymbol{B}) \leq r$. It follows that

$$\begin{aligned}
\|\boldsymbol{A} - \boldsymbol{B}\|_1 &= \sum_{i\geq 1} \sigma_i(\boldsymbol{A} - \boldsymbol{B}) \\
&= \sigma_1(\boldsymbol{A} - \boldsymbol{B}) + \sum_{i\geq 2} \sigma_i(\boldsymbol{A} - \boldsymbol{B}) \\
&\geq \|\boldsymbol{A} - \boldsymbol{B}\|_\infty + \sum_{i\geq 2} \sigma_{r+i}(\boldsymbol{A}) \\
&= \|\boldsymbol{A} - \boldsymbol{B}\|_\infty - \sigma_{r+1}(\boldsymbol{A}) + \sigma_{r+1}^{(1)}(\boldsymbol{A}).
\end{aligned}$$

The first expression is the representation of the Schatten 1-norm in terms of singular values. The inequality is (A.6). Finally, we identify the best Schatten 1-norm error from (A.1). □

To obtain the Schatten ∞-norm bound (4.2), we combine Proposition A.8 with the Schatten 1-norm bound (4.1):

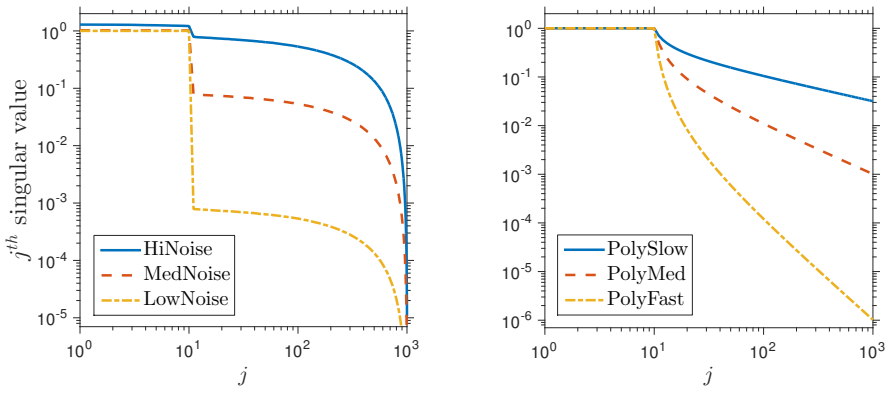
$$\begin{aligned}
\mathbb{E}\,\|\boldsymbol{A} - \hat{\boldsymbol{A}}_r\|_\infty &\leq \sigma_{r+1}(\boldsymbol{A}) + \left[\mathbb{E}\,\|\boldsymbol{A} - \hat{\boldsymbol{A}}_r\|_1 - \sigma_{r+1}^{(1)}(\boldsymbol{A})\right] \\
&\leq \sigma_{r+1}(\boldsymbol{A}) + \frac{r}{k - r - \alpha} \cdot \sigma_{r+1}^{(1)}(\boldsymbol{A}).
\end{aligned}$$

Similarly, to obtain the Schatten ∞-norm bound (4.4), we combine Proposition A.8 with the Schatten 1-norm bound (4.3).

B Supplemental Numerics

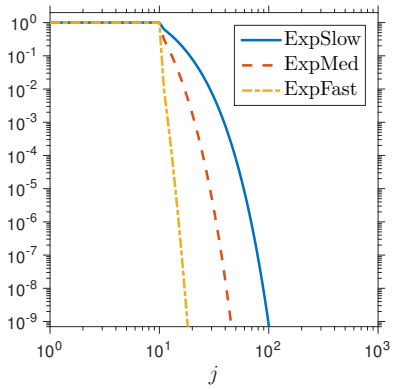
This appendix documents additional numerical work. These experiments provide a more complete picture of the performance of the psd approximation methods.

- Figure B.1 contains a plot of the singular-value spectrum of each input matrix described in Sec. 5.
- Figures B.2–B.10 document the results of numerical experiments for the remaining parameter regimes outlined in Sec. 5. In particular, we consider all Schatten p-norm relative error measures for $p \in \{1, 2, \infty\}$ and all effective rank parameters $R \in \{5, 10, 20\}$ for the synthetic data. We omit the case $p = \infty, R = 20$ because the plots are uninformative.
- Figure B.11 gives evidence about the numerical challenges involved in implementing Nyström approximations, such as (2.7). Our implementation in Algorithm 3 is based on the Nyström approximation routine `eigenn` released by Tygert [37] to accompany the paper [27]. We compare with another implementation strategy described in the text of the same paper [27, Eqn. (13)]. It is surprising to discover very different levels of precision in two implementations designed by professional numerical analysts.

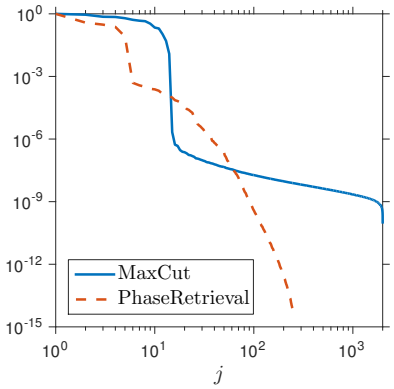


(A) Low-Rank + PSD Noise

(B) Polynomial Decay

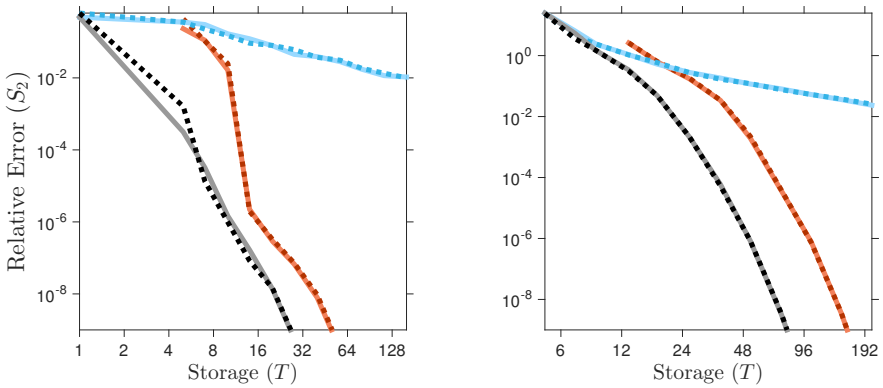


(C) Exponential Decay



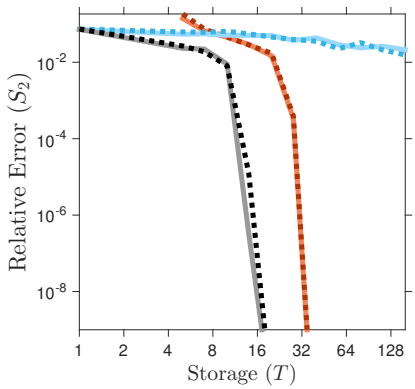
(D) `MaxCut` and `PhaseRetrieval`

FIGURE B.1: **Singular Values of Input Matrices.** These plots display the singular value spectra of the input matrices that appear in the experiments. See Sec. 5 for descriptions of the matrices.

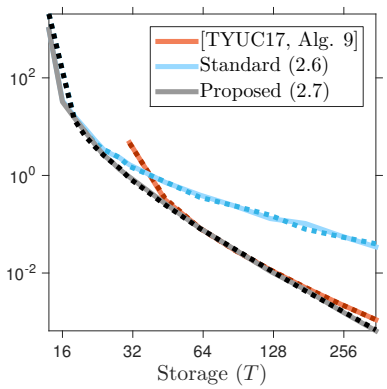


(A) `PhaseRetrieval` ($r = 1$)

(B) `PhaseRetrieval` ($r = 5$)

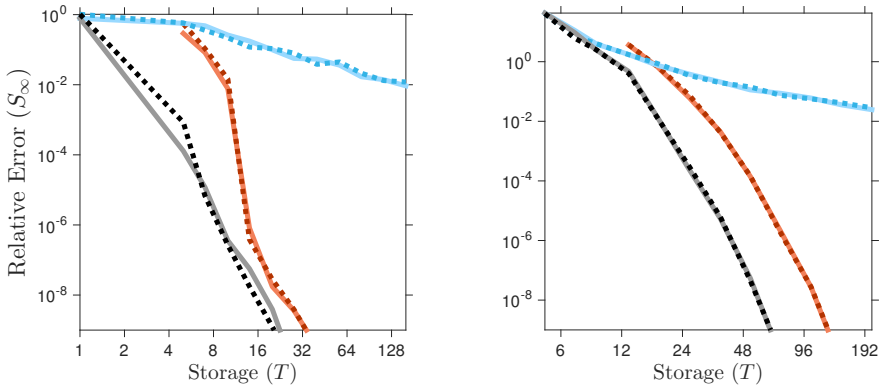


(C) `MaxCut` ($r = 1$)



(D) `MaxCut` ($r = 14$)

FIGURE B.2: **Application Examples, Approximation Rank r, Schatten 2-Norm Error.** The data series are generated by three algorithms for rank-r psd approximation. **Solid lines** are generated from the Gaussian sketch; **dashed lines** are from the SSFT sketch. Each panel displays the Schatten 2-norm relative error (5.1) as a function of storage cost T. See Sec. 5 for details.



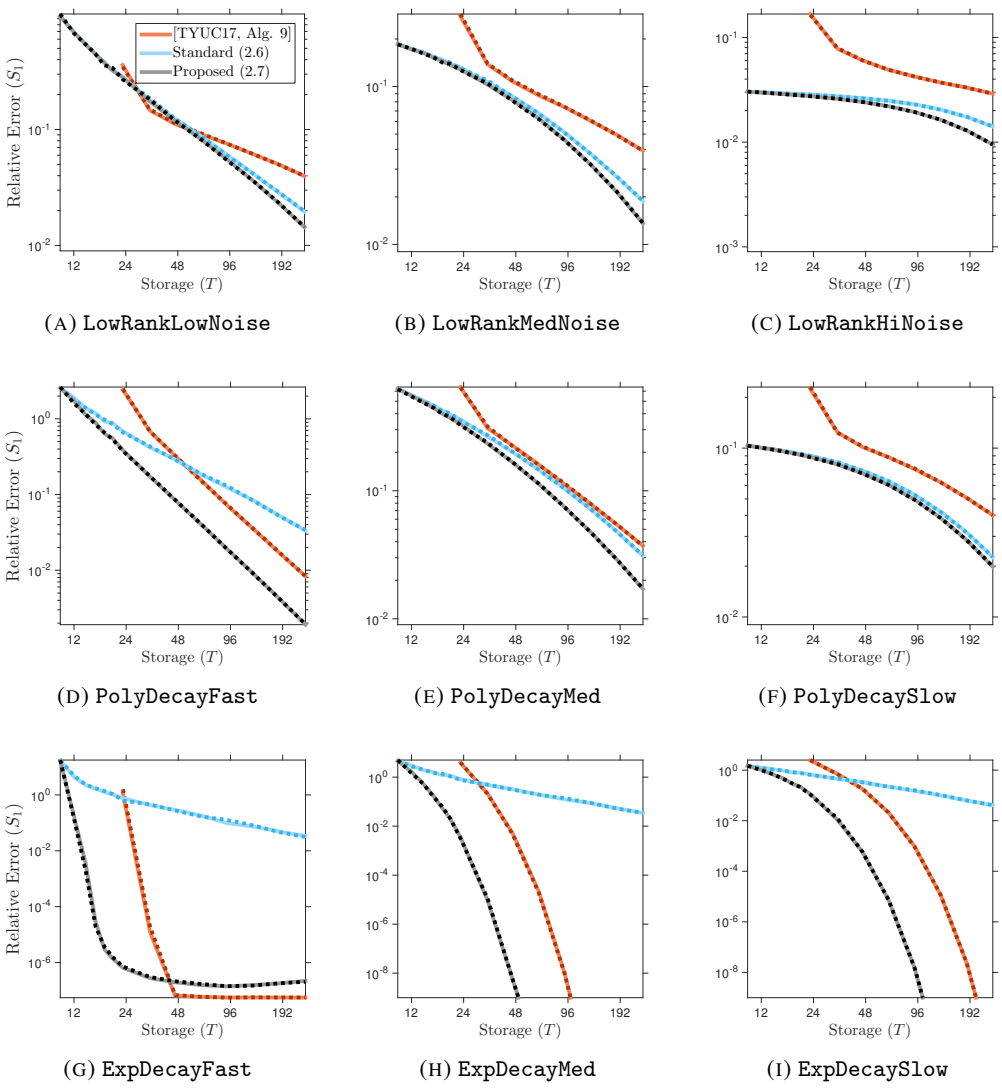
(A) `PhaseRetrieval` ($r = 1$)

(B) `PhaseRetrieval` ($r = 5$)

(C) `MaxCut` ($r = 1$)

(D) `MaxCut` ($r = 14$)

FIGURE B.3: **Application Examples, Approximation Rank r, Schatten ∞-Norm Error.** The data series are generated by three algorithms for rank-r psd approximation. **Solid lines** are generated from the Gaussian sketch; **dashed lines** are from the SSFT sketch. Each panel displays the Schatten ∞-norm relative error (5.1) as a function of storage cost T. See Sec. 5 for details.

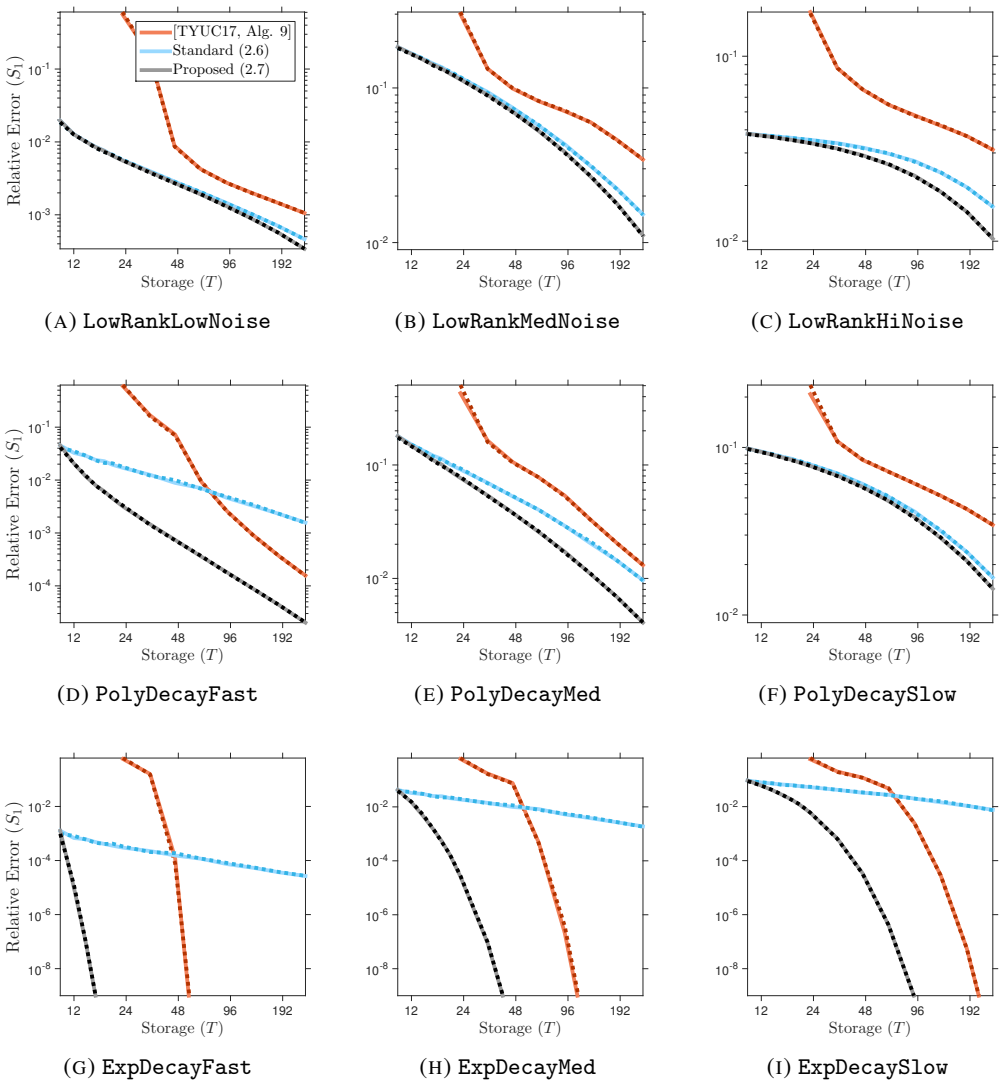


(A) `LowRankLowNoise` (B) `LowRankMedNoise` (C) `LowRankHiNoise`

(D) `PolyDecayFast` (E) `PolyDecayMed` (F) `PolyDecaySlow`

(G) `ExpDecayFast` (H) `ExpDecayMed` (I) `ExpDecaySlow`

FIGURE B.4: **Synthetic Examples with Effective Rank $R = 5$, Approximation Rank $r = 10$, Schatten 1-Norm Error.** The series are generated by three algorithms for rank-r psd approximation with $r = 10$. **Solid lines** are generated from the Gaussian sketch; **dashed lines** are from the SSFT sketch. Each panel displays the Schatten 1-norm relative error (5.1) as a function of storage cost T. See Sec. 5 for details.

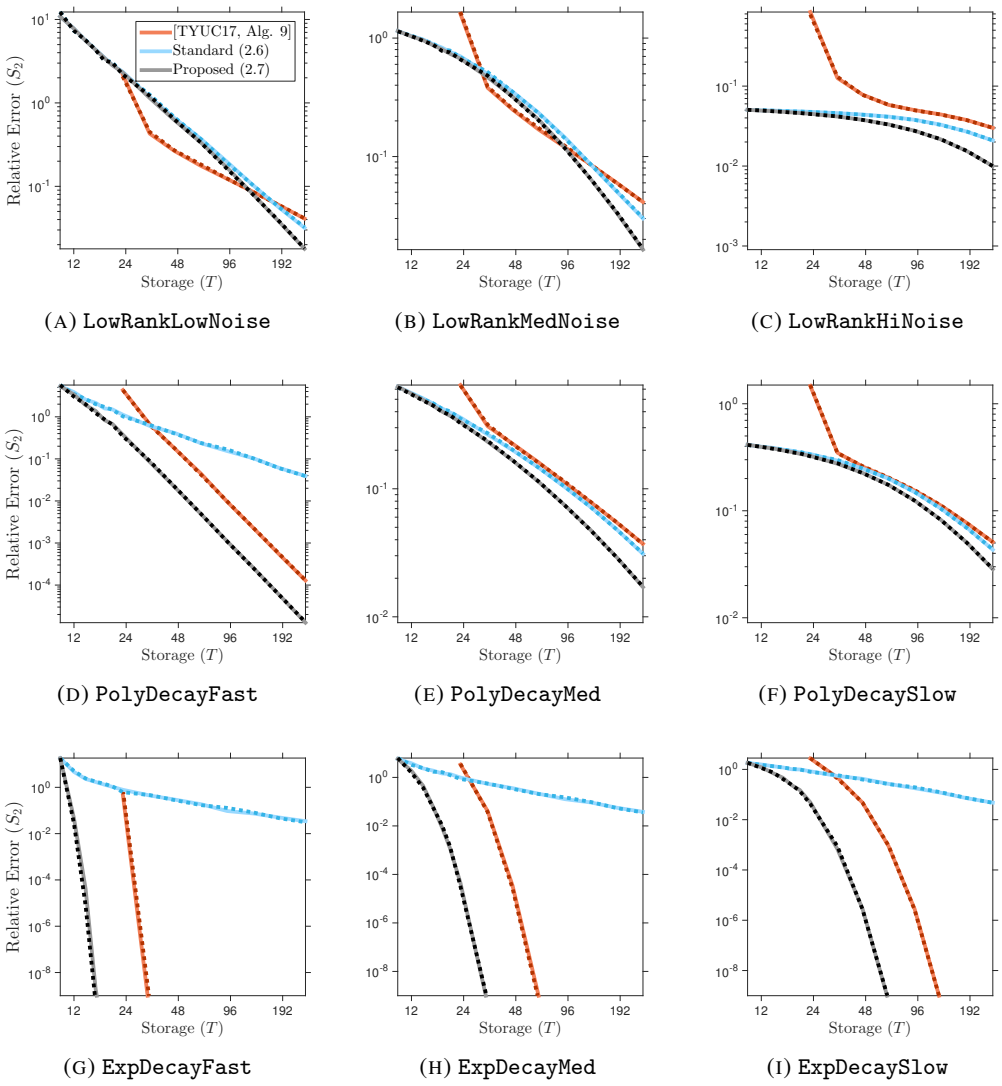


(A) `LowRankLowNoise` (B) `LowRankMedNoise` (C) `LowRankHiNoise`

(D) `PolyDecayFast` (E) `PolyDecayMed` (F) `PolyDecaySlow`

(G) `ExpDecayFast` (H) `ExpDecayMed` (I) `ExpDecaySlow`

FIGURE B.5: **Synthetic Examples with Effective Rank $R = 20$, Approximation Rank $r = 10$, Schatten 1-Norm Error.** The series are generated by three algorithms for rank-r psd approximation with $r = 10$. **Solid lines** are generated from the Gaussian sketch; **dashed lines** are from the SSFT sketch. Each panel displays the Schatten 1-norm relative error (5.1) as a function of storage cost T. See Sec. 5 for details.

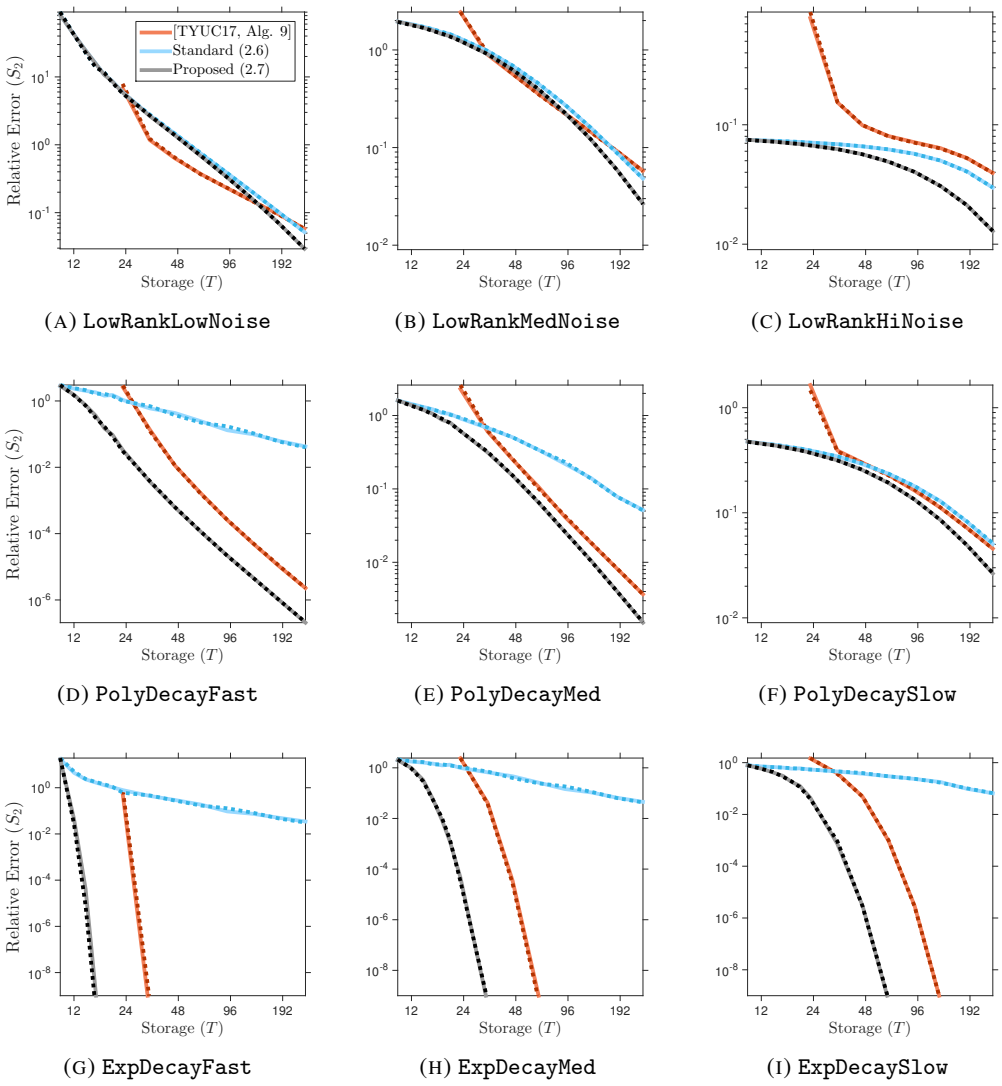


(A) `LowRankLowNoise` (B) `LowRankMedNoise` (C) `LowRankHiNoise`

(D) `PolyDecayFast` (E) `PolyDecayMed` (F) `PolyDecaySlow`

(G) `ExpDecayFast` (H) `ExpDecayMed` (I) `ExpDecaySlow`

FIGURE B.6: **Synthetic Examples with Effective Rank $R = 5$, Approximation Rank $r = 10$, Schatten 2-Norm Error.** The series are generated by three algorithms for rank-r psd approximation with $r = 10$. **Solid lines** are generated from the Gaussian sketch; **dashed lines** are from the SSFT sketch. Each panel displays the Schatten 2-norm relative error (5.1) as a function of storage cost T. See Sec. 5 for details.

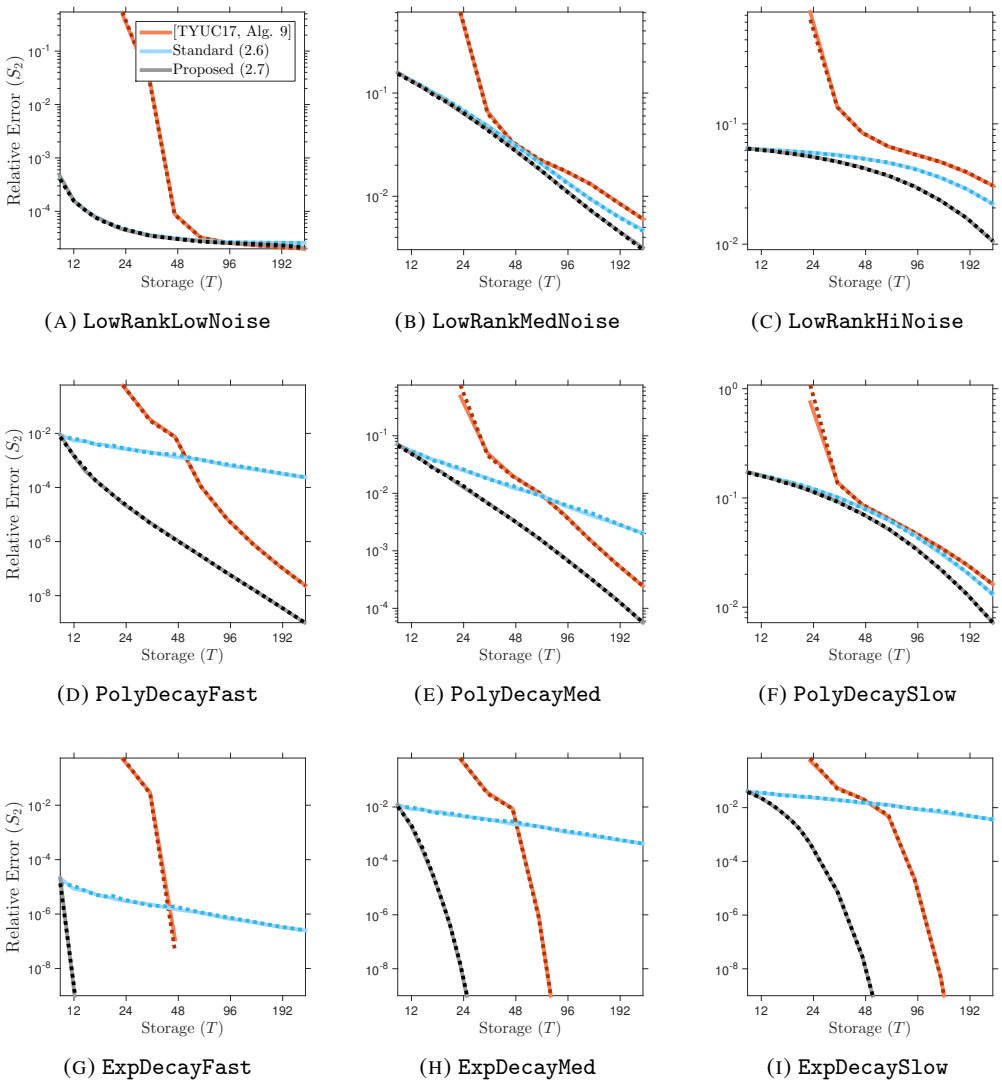


(A) `LowRankLowNoise` (B) `LowRankMedNoise` (C) `LowRankHiNoise`

(D) `PolyDecayFast` (E) `PolyDecayMed` (F) `PolyDecaySlow`

(G) `ExpDecayFast` (H) `ExpDecayMed` (I) `ExpDecaySlow`

FIGURE B.7: **Synthetic Examples with Effective Rank $R = 10$, Approximation Rank $r = 10$, Schatten 2-Norm Error.** The series are generated by three algorithms for rank-r psd approximation with $r = 10$. **Solid lines** are generated from the Gaussian sketch; **dashed lines** are from the SSFT sketch. Each panel displays the Schatten 2-norm relative error (5.1) as a function of storage cost T. See Sec. 5 for details.

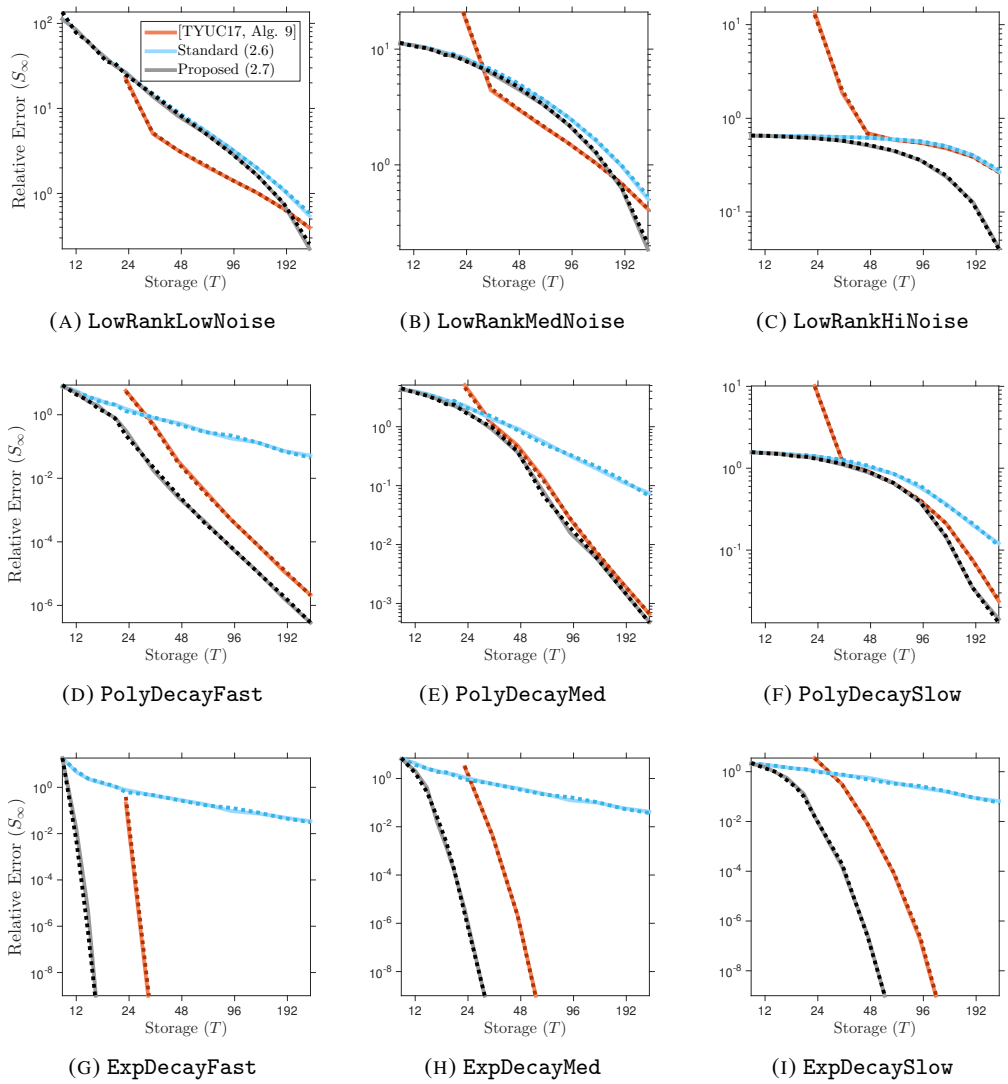


(A) `LowRankLowNoise` (B) `LowRankMedNoise` (C) `LowRankHiNoise`

(D) `PolyDecayFast` (E) `PolyDecayMed` (F) `PolyDecaySlow`

(G) `ExpDecayFast` (H) `ExpDecayMed` (I) `ExpDecaySlow`

FIGURE B.8: **Synthetic Examples with Effective Rank $R = 20$, Approximation Rank $r = 10$, Schatten 2-Norm Error.** The series are generated by three algorithms for rank-r psd approximation with $r = 10$. **Solid lines** are generated from the Gaussian sketch; **dashed lines** are from the SSFT sketch. Each panel displays the Schatten 2-norm relative error (5.1) as a function of storage cost T. See Sec. 5 for details.

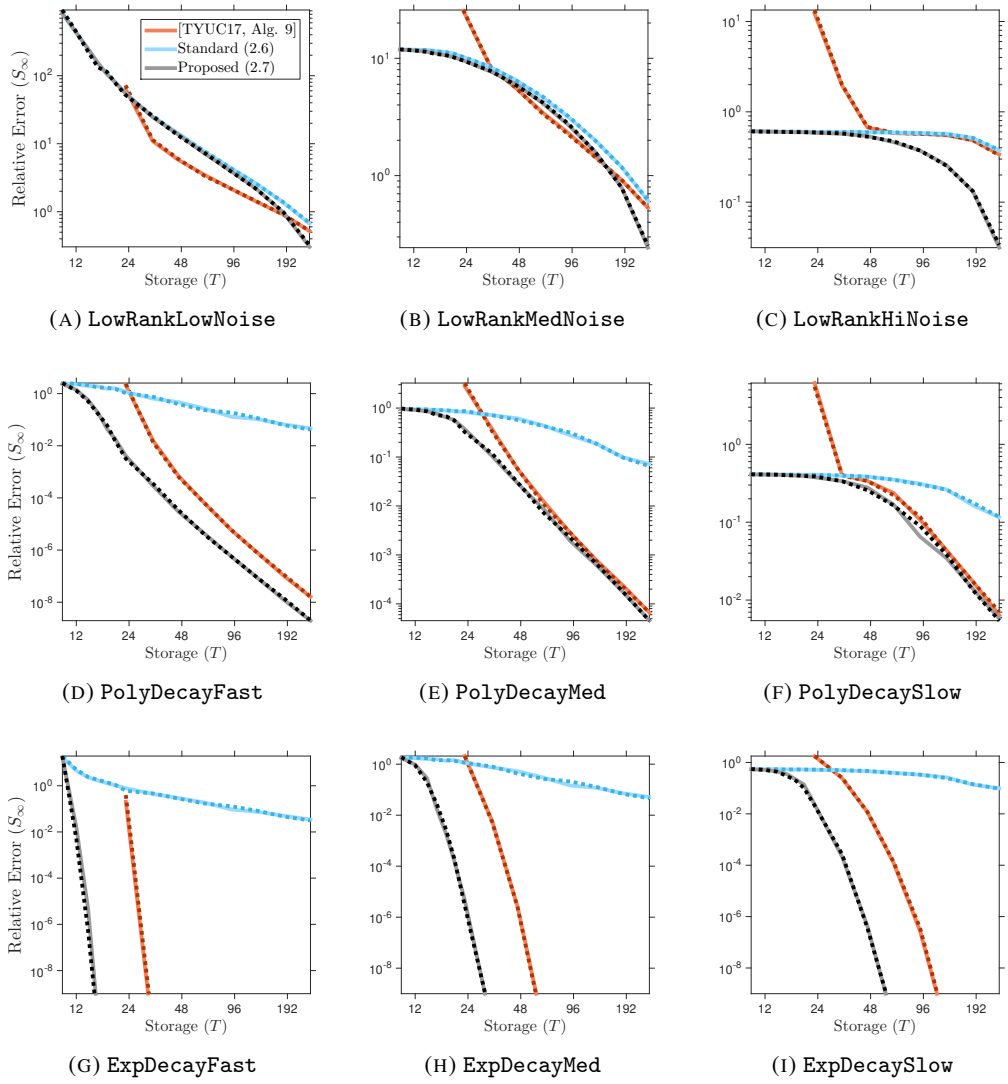


(A) `LowRankLowNoise` (B) `LowRankMedNoise` (C) `LowRankHiNoise`

(D) `PolyDecayFast` (E) `PolyDecayMed` (F) `PolyDecaySlow`

(G) `ExpDecayFast` (H) `ExpDecayMed` (I) `ExpDecaySlow`

FIGURE B.9: **Synthetic Examples with Effective Rank $R = 5$, Approximation Rank $r = 10$, Schatten ∞-Norm Error.** The series are generated by three algorithms for rank-r psd approximation with $r = 10$. **Solid lines** are generated from the Gaussian sketch; **dashed lines** are from the SSFT sketch. Each panel displays the Schatten ∞-norm relative error (5.1) as a function of storage cost T. See Sec. 5 for details.

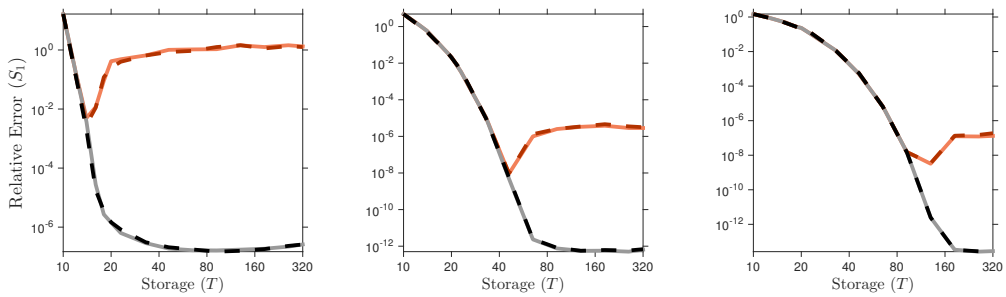


(A) `LowRankLowNoise` (B) `LowRankMedNoise` (C) `LowRankHiNoise`

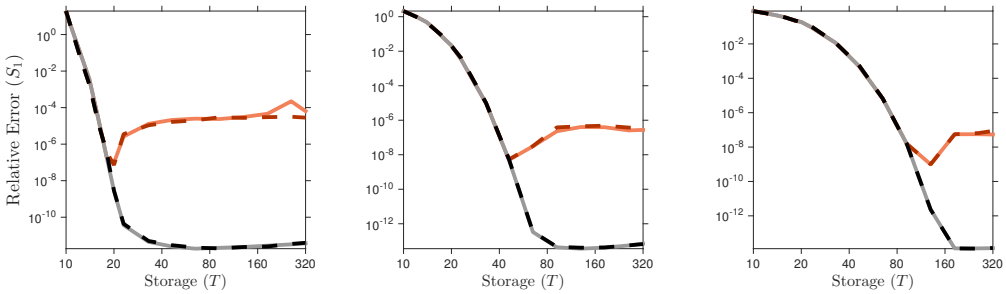
(D) `PolyDecayFast` (E) `PolyDecayMed` (F) `PolyDecaySlow`

(G) `ExpDecayFast` (H) `ExpDecayMed` (I) `ExpDecaySlow`

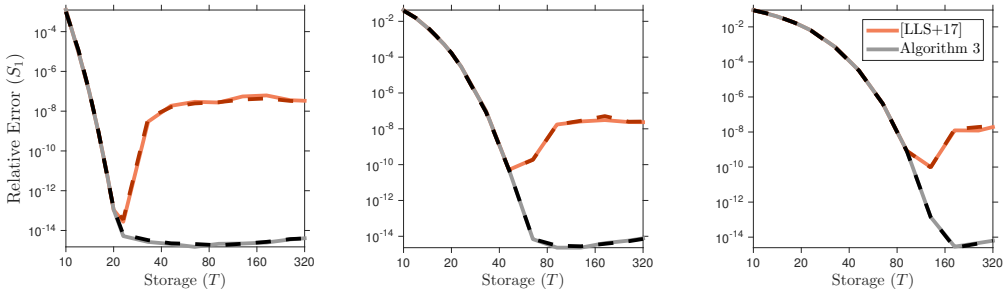
FIGURE B.10: **Synthetic Examples with Effective Rank $R = 10$, Approximation Rank $r = 10$, Schatten ∞-Norm Error.** The series are generated by three algorithms for rank-r psd approximation with $r = 10$. **Solid lines** are generated from the Gaussian sketch; **dashed lines** are from the SSFT sketch. Each panel displays the Schatten ∞-norm relative error (5.1) as a function of storage cost T. See Sec. 5 for details.



(A) `ExpDecayFast`, $R = 5$ (B) `ExpDecayMed`, $R = 5$ (C) `ExpDecaySlow`, $R = 5$



(D) `ExpDecayFast`, $R = 10$ (E) `ExpDecayMed`, $R = 10$ (F) `ExpDecaySlow`, $R = 10$



(G) `ExpDecayFast`, $R = 20$ (H) `ExpDecayMed`, $R = 20$ (I) `ExpDecaySlow`, $R = 20$

FIGURE B.11: **Bad Numerics, Approximation Rank $r = 10$, Schatten 1-Norm Error.** The series are generated by two implementations of the fixed-rank psd approximation (2.7). We compare Algorithm 3 with another approach [LLS+17] proposed in [27, Eqn. (13)]. **Solid lines** are generated from the Gaussian sketch; **dashed lines** are from the SSFT sketch. Each panel displays the Schatten 1-norm relative error (5.1) as a function of storage cost T. See App. B for details.

References

[1] N. Ailon and B. Chazelle. The fast Johnson-Lindenstrauss transform and approximate nearest neighbors. *SIAM J. Comput.*, 39(1):302–322, 2009.

[2] R. Bhatia. *Matrix analysis*. Springer-Verlag, New York, 1997.

[3] C. Boutsidis and A. Gittens. Improved matrix algorithms via the subsampled randomized Hadamard transform. *SIAM J. Matrix Anal. Appl.*, 34(3):1301–1340, 2013.

[4] C. Boutsidis, D. Garber, Z. Karnin, and E. Liberty. Online principal components analysis. In *Proc. 26th Ann. ACM-SIAM Symp. Discrete Algorithms (SODA)*, pages 887–901, 2015.

[5] C. Boutsidis, D. Woodruff, and P. Zhong. Optimal principal component analysis in distributed and streaming models. In *Proc. 48th ACM Symp. Theory of Computing (STOC)*, 2016.

[6] J. Chiu and L. Demanet. Sublinear randomized algorithms for skeleton decompositions. *SIAM J. Matrix Anal. Appl.*, 34(3):1361–1383, 2013.

[7] K. Clarkson and D. Woodruff. Low-rank PSD approximation in input-sparsity time. In *Proc. 28th Ann. ACM-SIAM Symp. Discrete Algorithms (SODA)*, pages 2061–2072, Jan. 2017.

[8] K. L. Clarkson and D. P. Woodruff. Numerical linear algebra in the streaming model. In *Proc. 41st ACM Symp. Theory of Computing (STOC)*, 2009.

[9] M. B. Cohen, S. Elder, C. Musco, C. Musco, and M. Persu. Dimensionality reduction for k-means clustering and low rank approximation. In *Proc. 47th ACM Symp. Theory of Computing (STOC)*, pages 163–172. ACM, New York, 2015.

[10] M. B. Cohen, J. Nelson, and D. P. Woodruff. Optimal Approximate Matrix Product in Terms of Stable Rank. In *43rd Int. Coll. Automata, Languages, and Programming (ICALP)*, volume 55, pages 11:1–11:14, 2016.

[11] T. A. Davis and Hu. The University of Florida sparse matrix collection. *ACM Trans. Math. Softw.*, 3(1): 1:1–1:25, 2011.

[12] P. Drineas and M. W. Mahoney. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *J. Mach. Learn. Res.*, 6:2153–2175, 2005.

[13] D. Feldman, M. Volkov, and D. Rus. Dimensionality reduction of massive sparse datasets using coresets. In *Adv. Neural Information Processing Systems 29 (NIPS)*, 2016.

[14] C. Fowlkes, S. Belongie, F. Chung, and J. Malik. Spectral grouping using the Nyström method. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(2):214–225, Jan. 2004.

[15] M. Ghasemi, E. Liberty, J. M. Phillips, and D. P. Woodruff. Frequent directions: Simple and deterministic matrix sketching. *SIAM J. Comput.*, 45(5):1762–1792, 2016.

[16] A. C. Gilbert, J. Y. Park, and M. B. Wakin. Sketched SVD: Recovering spectral features from compressed measurements. Available at `http://arXiv.org/abs/1211.0361`, Nov. 2012.

[17] A. Gittens. The spectral norm error of the naïve Nyström extension. Available at `http:arXiv.org/abs/1110.5305`, Oct. 2011.

[18] A. Gittens. *Topics in Randomized Numerical Linear Algebra*. PhD thesis, California Institute of Technology, 2013.

[19] A. Gittens and M. W. Mahoney. Revisiting the Nyström method for improved large-scale machine learning. Available at `http://arXiv.org/abs/1303.1849`, Mar. 2013.

[20] A. Gittens and M. W. Mahoney. Revisiting the Nyström method for improved large-scale machine learning. *J. Mach. Learn. Res.*, 17:Paper No. 117, 65, 2016.

[21] M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. Assoc. Comput. Mach.*, 42(6):1115–1145, 1995.

[22] M. Gu. Subspace iteration randomization and singular value problems. *SIAM J. Sci. Comput.*, 37(3): A1139–A1173, 2015.

[23] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.*, 53(2):217–288, 2011.

[24] N. J. Higham. Matrix nearness problems and applications. In *Applications of matrix theory (Bradford, 1988)*, pages 1–27. Oxford Univ. Press, New York, 1989.

[25] P. Jain, C. Jin, S. M. Kakade, P. Netrapalli, and A. Sidford. Streaming PCA: Matching matrix Bernstein and near-optimal finite sample guarantees for Oja's algorithm. In *29th Ann. Conf. Learning Theory (COLT)*, pages 1147–1164, 2016.

[26] S. Kumar, M. Mohri, and A. Talwalkar. Sampling methods for the Nyström method. *J. Mach. Learn. Res.*, 13:981–1006, Apr. 2012.

[27] H. Li, G. C. Linderman, A. Szlam, K. P. Stanton, Y. Kluger, and M. Tygert. Algorithm 971: An implementation of a randomized algorithm for principal component analysis. *ACM Trans. Math. Softw.*, 43 (3):28:1–28:14, Jan. 2017.

[28] Y. Li, H. L. Nguyen, and D. P. Woodruff. Turnstile streaming algorithms might as well be linear sketches. In *Proc. 2014 ACM Symp. Theory of Computing (STOC)*, pages 174–183. ACM, 2014.

[29] E. Liberty. *Accelerated dense random projections*. PhD thesis, Yale Univ., New Haven, 2009.

[30] M. W. Mahoney. Randomized algorithms for matrices and data. *Found. Trends Mach. Learn.*, 3(2):123–224, 2011.

[31] P.-G. Martinsson, V. Rokhlin, and M. Tygert. A randomized algorithm for the decomposition of matrices. *Appl. Comput. Harmon. Anal.*, 30(1):47–68, 2011.

[32] I. Mitliagkas, C. Caramanis, and P. Jain. Memory limited, streaming PCA. In *Adv. Neural Information Processing Systems 26 (NIPS)*, pages 2886–2894, 2013.

[33] C. Musco and D. Woodruff. Sublinear time low-rank approximation of positive semidefinite matrices. Available at `http://arXiv.org/abs/1704.03371`, Apr. 2017.

[34] J. C. Platt. FastMap, MetricMap, and Landmark MDS are all Nyström algorithms. In *Proc. 10th Int. Workshop Artificial Intelligence and Statistics (AISTATS)*, pages 261–268, 2005.

[35] J. A. Tropp. Improved analysis of the subsampled randomized Hadamard transform. *Adv. Adapt. Data Anal.*, 3(1-2):115–126, 2011.

[36] J. A. Tropp, A. Yurtsever, M. Udell, and V. Cevher. Randomized single-view algorithms for low-rank matrix approximation. ACM Report 2017-01, Caltech, Pasadena, Jan. 2017. Available at `http://arXiv.org/abs/1609.00048`, v1.

[37] M. Tygert. Beta versions of Matlab routines for principal component analysis. Available at `http://tygert.com/software.html`, 2014.

[38] C. K. I. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In *Adv. Neural Information Processing Systems 13 (NIPS)*, 2000.

[39] D. P. Woodruff. Sketching as a tool for numerical linear algebra. *Found. Trends Theor. Comput. Sci.*, 10 (1-2):iv+157, 2014.

[40] F. Woolfe, E. Liberty, V. Rokhlin, and M. Tygert. A fast randomized algorithm for the approximation of matrices. *Appl. Comput. Harmon. Anal.*, 25(3):335–366, 2008.

[41] T. Yang, Y.-F. Li, M. Mahdavi, R. Jin, and Z.-H. Zhou. Nyström method vs random Fourier features: A theoretical and empirical comparison. In *Adv. Neural Information Processing Systems 25 (NIPS)*, pages 476–484, 2012.

[42] A. Yurtsever, M. Udell, J. A. Tropp, and V. Cevher. Sketchy decisions: Convex low-rank matrix optimization with optimal storage. In *Proc. 20th Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, Fort Lauderdale, May 2017.