Simple Heuristics Yield Provable Algorithms for Masked Low-Rank Approximation

Cameron Musco
College of Information and Computer Sciences, University of Massachusetts Amherst, MA, USA
cmusco@cs.umass.edu

Christopher Musco
Department of Computer Science and Engineering,
New York University Tandon School of Engineering, NY, USA
cmusco@nyu.edu

David P. Woodruff
Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
dwoodruf@cs.cmu.edu

Abstract

In the *masked low-rank approximation problem*, one is given data matrix $A \in \mathbb{R}^{n\times n}$ and binary mask matrix $W \in \{0,1\}^{n\times n}$. The goal is to find a rank-k matrix L for which:

$$\text{cost}(L) \stackrel{\text{def}}{=} \sum_{i=1}^{n}\sum_{j=1}^{n} W_{i,j}\cdot(A_{i,j}-L_{i,j})^2 \leq OPT + \epsilon\|A\|_F^2,$$

where $OPT = \min_{\text{rank-}k\ \hat{L}} \text{cost}(\hat{L})$ and ϵ is a given error parameter. Depending on the choice of W, the above problem captures factor analysis, low-rank plus diagonal decomposition, robust PCA, low-rank matrix completion, low-rank plus block matrix approximation, low-rank recovery from monotone missing data, and a number of other important problems. Many of these problems are NP-hard, and while algorithms with provable guarantees are known in some cases, they either 1) run in time $n^{\Omega(k^2/\epsilon)}$ or 2) make strong assumptions, for example, that A is incoherent or that the entries in W are chosen independently and uniformly at random.

In this work, we show that a common polynomial time heuristic, which simply sets A to 0 where W is 0, and then finds a standard low-rank approximation, yields bicriteria approximation guarantees for this problem. In particular, for rank $k' > k$ depending on the *public coin partition number* of W, the heuristic outputs rank-k' L with $\text{cost}(L) \leq OPT + \epsilon\|A\|_F^2$. This partition number is in turn bounded by the randomized communication complexity of W, when interpreted as a two-player communication matrix. For many important cases, including all those listed above, this yields bicriteria approximation guarantees with rank $k' = k\cdot\text{poly}(\log n/\epsilon)$.

Beyond this result, we show that different notions of communication complexity yield bicriteria algorithms for natural variants of masked low-rank approximation. For example, multi-player number-in-hand communication complexity connects to masked tensor decomposition and non-deterministic communication complexity to masked Boolean low-rank factorization.

2012 ACM Subject Classification Theory of computation → Approximation algorithms analysis; Theory of computation → Communication complexity

Keywords and phrases low-rank approximation, communication complexity, weighted low-rank approximation, bicriteria approximation algorithms

Digital Object Identifier 10.4230/LIPIcs.ITCS.2021.6

Related Version Full paper available at `https://arxiv.org/abs/1904.09841`.

Funding *David P. Woodruff*: David Woodruff would like to thank support from the Office of Naval Research (ONR) grant N00014-18-1-2562.

Acknowledgements Part of this work was done while the authors were visiting the Simons Institute for the Theory of Computing.

1 Introduction

The goal of low-rank approximation is to approximate an $n \times n$ matrix A with a rank-k matrix L. L can be written as the product $L = U \cdot V$ of a "tall-and-thin" matrix U and a "short-and-wide" matrix V with k columns and rows respectively. For $k \ll n$ this approximation can lead to computational speedups: one can store the factors U and V with less memory than storing A itself, and can compute the product $U \cdot V \cdot x$ with a vector x faster than computing $A \cdot x$. Additionally, low-rank approximation is useful for denoising and can reveal low-dimensional structure in high-dimensional data (it is e.g., the basis behind principal component analysis). It thus serves as a preprocessing step in many applications, including clustering, data mining, and recommendation systems. The optimal low-rank approximation to A with distance measured in the Frobenius, spectral, or any unitarily invariant norm can be computed in polynomial time using a singular value decomposition (SVD). There are also extremely efficient approximation algorithms for finding a near optimal L under different measures, including the Frobenius norm, spectral norm, and various entrywise norms. For a comprehensive treatment, we refer the reader to the surveys [36, 47, 70].

Despite its wide applicability, in many situations standard low-rank approximation does not suffice. For example, it is common that certain entries in A either *don't obey underlying low-rank structure* or are *missing*. For example, A may be close to low-rank but with a small number of corrupted entries, or may be the sum of a low-rank matrix plus a high-rank, but still efficiently representable, diagonal or block diagonal matrix. In both cases, one must compute a low-rank approximation of A ignoring the outlying entries. One can formalize this problem, considering a binary matrix W with $W_{i,j} = 0$ for each outlying entry (i, j) of A and $W_{i,j} = 1$ otherwise.

▶ **Problem 1** (Masked Low-Rank Approximation). *Given $A \in \mathbb{R}^{n \times n}$, binary $W \in \{0,1\}^{n \times n}$, and rank parameter k, find rank-k L minimizing:*

$$\|W \circ (A - L)\|_F^2 = \sum_{i,j \in [n]} W_{i,j} \cdot (A_{i,j} - L_{i,j})^2,$$

where for two matrices M and N of the same size, $M \circ N$ denotes the entrywise (Hadamard product): with $(M \circ N)_{i,j} = M_{i,j} \cdot N_{i,j}$ and for integer n, $[n]$ denotes $\{1, \ldots, n\}$.

As stated, Problem 1 minimizes the squared Frobenius norm of $W \circ (A - L)$. However any matrix norm can be used. In any case, is unclear how to extend standard low-rank approximation algorithms to solving Problem 1, since they optimize over the full matrix A, without the ability to take into account W encoding entries that should be ignored. We note that Problem 1 is equivalent to minimizing $\|A - (L + S)\|_F^2$ where L is rank-k and S is any matrix with support restricted to the 0 entries of W. If these zeros are on the diagonal, then S is diagonal. If they are sparse, then S is sparse, etc. This is how Problem 1 is traditionally stated in many applications.

1.1 Existing Work

A common approach to solving Problem 1 is to apply alternating minimization or the EM (Expectation-Maximization) algorithm. In fact, factor analysis, a slight variant of Problem 1 when W is 0 on its diagonal and 1 off the diagonal, was one of the original motivations of the EM algorithm [21, 56]. Much recent work studies when alternating minimization for Problem 1 converges in polynomial time under the assumptions that (1) there is a solution $L = U \cdot V \approx A$ which is *incoherent*, meaning that the squared row norms of U and column

norms of V are small and (2) the entries of W are selected at random or have pseudorandom properties [73, 40, 51]. Under similar assumptions it can be shown that Problem 1 and the related problem of *robust PCA* can be solved via convex relaxation in polynomial time [13, 71, 12]. In many cases, these algorithms perform well in practice even when the above assumptions do not hold. Additionally, they can be proven to run in polynomial time in some common settings when the entries of W are not random – e.g., when W is zero only on its diagonal or at a few arbitrary locations. That is, when we want to approximate A as a low-rank plus diagonal component, or a low-rank matrix with arbitrary sparse corruptions respectively. However, these results still require assuming the existence of $U \cdot V$ that is incoherent and further that is *exact* – with $U \cdot V$

A natural question is if for common mask patterns, one can obtain provable algorithms without incoherence or other strong assumptions. This approach was taken in [54] in the context of *weighted low-rank approximation*, where W is a nonnegative matrix and the objective is still to minimize $\|W \circ (A - L)\|_F^2$. When W is binary, this reduces to Problem 1. In [54] it was shown that if W has at most r distinct columns, then it is possible to obtain a relative error guarantee in $2^{\text{poly}(rk/\epsilon)} \cdot \text{poly}(n)$ time. More generally, if the rank of W over the reals is at most r, then $n^{\text{poly}(rk/\epsilon)}$ time is achievable. Note that such algorithms are only polynomial time if k, r, and $1/\epsilon$ are very small. In many common use cases, such as when W is all 0*s* on the diagonal and 1 off-diagonal (corresponding to low-rank plus diagonal decomposition), or when W is all 0s above the diagonal and 1s on or beneath the diagonal, r is large: in fact $\text{rank}(W) = r = n$ in these cases.

When A is low-rank with sparse corruptions, i.e., when W has at most t zero entries per row and column, the algorithms of [54] can be applied if there is an exact solution (with $A = L$ on all non-corrupted entries). [54] referred to this problem as *adversarial matrix completion* and gave an $n^{O(tk^2)}$ time algorithm. This is only polynomial time for constant values of t and k, and even for constant t and k is very large. Moreover, their method cannot be used in the approximate case since it requires creating a low-rank weight matrix W' whose support matches that of W. Since W may be far from low-rank, the non-zero entries of W and W' necessarily have very different values. This introduces significant error, unless $A = L$ exactly on the support of W.

1.2 Our Contributions

With the goal of obtaining fast masked low-rank approximation algorithms, we consider *bicriteria approximation* with additive error. That is, we allow the rank k' of the output L to be slightly larger than k, but one still compares to the best rank-k approximation. Formally, given $A \in \mathbb{R}^{n \times n}$, $W \in \{0,1\}^{n \times n}$, and an error parameter ϵ, we would like to find a rank-k' matrix L for which:

$$\|W \circ (A - L)\|_F^2 \leq OPT + \epsilon \|A\|_F^2, \tag{1}$$

where $OPT = \min_{\text{rank-}k\ \hat{L}} \|W \circ (A - \hat{L})\|_F^2$ is the optimal value of Problem 1.

Assuming a variant of the Exponential Time Hypothesis, [54] shows a lower bound of $2^{\Omega(r)}$ time for finding rank-k L achieving (1) with constant ϵ when W is rank-r. Thus the relaxation to bicriteria approximation seems necessary. In many applications it is not essential for the output rank k' to be exactly k – as long as k' is small, one still obtains significant compression. Indeed, bicriteria algorithms for low-rank matrix approximation are widely studied [22, 23, 18, 17, 63, 9]. The starting point of our work is the question:

For which mask patterns $W \in \{0,1\}^{n \times n}$ can one obtain efficient bicriteria low-rank approximation algorithms with $k' \leq k \cdot \text{poly}((\log n)/\epsilon)$ satisfying (1)?

Main Results

We show that the answer to this question is related to the *randomized communication complexity* of W.[1] If the rows and columns of $W \in \{0,1\}^{n\times n}$ are indexed by strings $x \in \{0,1\}^{\log n}$ and $y \in \{0,1\}^{\log n}$, respectively, we can think of W as a two-player communication matrix for a Boolean function f, where $f(x,y) = W_{x,y}$. Here Alice has x, Bob has y, and the two parties want to exchange messages with as few bits as possible to compute $f(x,y)$ with probability at least $1-\delta$. The number of bits required is the randomized communication complexity $R_\delta(f)$. If we further require that the protocol never errs when $f(x,y) = 1$, but for any fixed pair (x,y) with $f(x,y) = 0$, it errs with probability at most δ, then the number of bits required is the 1-sided randomized communication complexity $R_\delta^{1-sided}(f)$. We show:

▶ **Theorem 1.** *Letting f be the function computed by $W \in \{0,1\}^{n\times n}$ and $\neg f$ be its negation, there is a bicriteria low-rank approximation L with rank $k' = k \cdot 2^{R_\epsilon^{1-sided}(\neg f)}$ achieving:*

$$\|W \circ (A - L)\|_F^2 \leq OPT + 2\epsilon\|A \circ W\|_F^2,$$

where $OPT = \min_{\text{rank-}k\ \hat{L}} \|W \circ (A - \hat{L})\|_F^2$. L is computable in $O(\text{nnz}(A)) + n \cdot \text{poly}(k'/\epsilon)$ time.

As we will see, for many common W, $R_\epsilon^{1-sided}(\neg f)$ is very small – with $2^{R_\epsilon^{1-sided}(\neg f)}$ at most $\text{poly}(\log n/\epsilon)$. Note that our additive error is in terms of $\|A \circ W\|_F^2$ which is only smaller than $\|A\|_F^2$, and may be much smaller, if e.g., the zeros in W correspond to corruptions in A. We also show a bound in terms of the communication complexity with 2-sided error.

▶ **Theorem 2.** *Letting f be the function computed by $W \in \{0,1\}^{n\times n}$, there is a bicriteria low-rank approximation L with rank $k' = k \cdot 2^{R_\epsilon(f)}$ achieving:*

$$\|W \circ (A - L)\|_F^2 \leq OPT + 2\epsilon\|A \circ W\|_F^2 + \epsilon\|L_{opt} \circ (1 - W)\|_F^2,$$

where $OPT = \min_{\text{rank-}k\ \hat{L}} \|W \circ (A - \hat{L})\|_F^2$ and L_{opt} is any rank-k matrix achieving OPT. L is computable in $O(\text{nnz}(A)) + n \cdot \text{poly}(k'/\epsilon)$ time.

Further, the algorithm achieving Theorems 1 and 2 is extremely simple: just zero out the entries in A corresponding to entries in W that are 0 (i.e., compute $A \circ W$), and then output a standard rank-k' approximation of the resulting matrix. This is already a widely-used heuristic for solving Problem 1 [3, 74], and we obtain the first provable guarantees. An optimal low-rank approximation of $A \circ W$ can be computed in polynomial time via an SVD. An approximation achieving relative error $(1+\epsilon)$ can be computed with high probability in $O(\text{nnz}(A)) + n \cdot \text{poly}(k/\epsilon)$ time, giving the runtime bounds of Theorems 1 and 2 [19].

1.2.1 Applications

Theorems 1 and 2 provide the first bicriteria approximation algorithms for Problem 1 with small k' for a number of important special cases of the mask matrix W:

1. If W has at most t zero entries in each row, this is Low-Rank Plus Sparse (LRPS) matrix approximation, which captures the challenge of finding a low-rank approximation when a few entries are not known, or do not obey underlying low-rank structure. It has been studied in the context of adversarial matrix completion [58], robust matrix decomposition [31, 12], optics, system identification [7], and more [15].

[1] Our bounds actually hold for the public coin partition number of W, which is upper bounded by the randomized communication complexity [34]. See Section 1.2.4 for a more detailed discussion.

2. If W is zero exactly on the diagonal entries, this is Low-Rank Plus Diagonal (LRPD) matrix approximation. This problem arises since in practice, many matrices that are not close to low-rank are close to diagonal, or contain a mixture of diagonal and low-rank components [15]. This observation has been used e.g., to construct compact representations of kernel matrices [61, 69], weight matrices in neural networks [48, 74], and covariance matrices [66, 65]. LRPD approximation also arises in applications related to source separation [44] and variational inference [49] and is closely related to factor analysis [64, 57], which adds the additional constraints that L and $A - L$ are PSD.
3. If W is the negation of a block-diagonal matrix with blocks of varying sizes, meaning that W is 0 on entries in the blocks and 1 on entries outside of the blocks, this is Low-Rank Plus Block-Diagonal (LRPBD) matrix approximation. This is a natural generalization of the LRPD problem and has been studied in the context of anomaly detection in networks [4], foreground detection [29], and robust principal component analysis [41]. We also consider the natural generalization of LRPS approximation discussed above, which we call the Low-Rank Plus Block-Sparse (LRPBS) matrix approximation problem.
4. If each row of W has a prefix of ones, followed by a suffix of zeros, this is the Monotone Missing Data Pattern (MMDP) problem. This is a common missing data pattern, arising in the event that when a variable is missing from a sample, all subsequent variables are also missing. Methods for handling this pattern are, e.g., included in the SAS/STAT package for statistical analyses [1]. We refer the reader to [68] for more examples of common missing data patterns, such as "connected" and "file matching" patterns.
5. If W is the negation of a banded matrix where $W_{i,j} = 0$ iff $|i - j| < p$ for some distance p, this is Low-Rank Plus Banded (LRPBand) matrix approximation. Variants of this problem arise in scientific computing and machine learning, in particular in the approximation of kernel matrices via fast multipole methods [55, 28, 72]. These methods approximate a kernel matrix using a low-rank "far-field" component, and a "near-field" component, which explicitly represents the kernel function between close points. If points are in one dimension and sorted, this corresponds to approximating A with a low-rank plus banded matrix. Many methods compute the low-rank component analytically (using polynomial approximations of the kernel function). A natural alternative is to seek an optimal decomposition via Problem 1. Many applications involve higher dimensional data. E.g., in the two-dimensional case, each $i \in [n]$ can be mapped to $(i_1, i_2) \in [\sqrt{n}] \times [\sqrt{n}]$ where i_1, i_2 correspond to the first and second halves of $i's$ binary expansion. $W_{i,j} = 0$ iff $|i_1 - j_1| + |i_2 - j_2| < p$. We give similar bounds for this multidimensional variant.

We summarize our results for the above weight patterns in Table 1. We detail the specific functions f used in these applications in Sections 2 and 3, but note that (1), (2), and (3) use variants of Equality, which has $O(\log(1/\epsilon))$ randomized 1-sided error communication complexity, (4) and (5) use a variant of the Greater-Than problem with $O(\log\log n + \log(1/\epsilon))$ randomized 2-sided error communication complexity for $\log n$ bit inputs.

1.2.2 Relation to Matrix Completion

Masked low-rank approximation is closely related to the well-studied matrix completion problem [13, 32, 37], however the goal is different. In masked low-rank approximation, we want to approximate A as accurately as possible on the *non-masked entries* (i.e., where $W_{ij} = 1$). In matrix completion, the support of W represents entries in A that are observed and the goal is to approximate A *on the missing entries* (i.e., where $W_{ij} = 0$). The most common approach to solving this problem is in fact to find a low-rank approximation fitting the non-missing entries (i.e., to solve Problem 1), however the two problems are not equivalent.

Table 1 Summary of applications of Theorems 1 and 2.

Mask Pattern	k'	Communication Problem	Ref.
LRPD/LRPBD	$O(k/\epsilon)$	Equality	Cor. 15 & 16
LRPBand	$k \cdot \text{poly}\left(\frac{\log n}{\epsilon}\right)$	Variant of Greater-Than	Cors. 17
LRPS/LRPBS (w/ sparsity t)	$O(kt/\epsilon)$	Variant of equality	Full paper
MMDP	$k \cdot \text{poly}\left(\frac{\log n}{\epsilon}\right)$	Greater-Than	Cor. 18
Subsampled Toeplitz	$O(\min(pk, k/\epsilon))$	Equality mod p	Full paper

For example, it is not clear that a bicriteria solution to Problem 1, as given by Theorems 1 and 2, will give anything interesting for the matrix completion problem. In fact, our proof technique implies that it likely will not.

We further note that in matrix completion, the mask W is typically assumed to be random and the goal is to recover the missing entries of A when W has as few sampled ones as possible. We do not expect that a random matrix will have low-communication complexity, unless it has further structure (e.g., few zeros or ones per row).

1.2.3 Other Communication Models

Theorems 1 and 2 connect communication complexity to the analysis of a simple heuristic for masked low-rank approximation. A natural question is:

Can other notions of communication complexity, such as multi-party communication complexity, non-deterministic communication complexity, and communication complexity of non-Boolean functions yield algorithms for masked low-rank approximation?

We answer this question affirmatively. We first look at multi-party communication complexity, which we show corresponds to masked tensor low-rank approximation. Here we focus on order-3 tensors, though our results are proven for arbitrary order-t tensors. A tensor is just an array $A \in \mathbb{R}^{n\times n\times n}$. In masked low-rank tensor approximation we are given such an A and a mask tensor $W \in \{0,1\}^{n\times n\times n}$ and the goal is to find rank-k tensor L minimizing $\|W \circ (A - L)\|_F^2$. This problem has been widely studied in the context of low-rank tensor completion [25, 43, 50] and robust tensor PCA [42, 45], which corresponds to the setting where W's zeros represent sparse corruptions of an otherwise low-rank tensor. Applications include color image and video reconstruction along with low-rank plus diagonal tensor approximation [8], where W is zero on its diagonal and one everywhere else. In the full paper we show:

▶ **Theorem 3** (Multiparty Communication Complexity → Tensor Low-Rank Approx)**.** *Let f be the function computed by $W \in \{0,1\}^{n\times n\times n}$, $\neg f$ be its negation, and $R_\epsilon^{3,1-sided}(\neg f)$ be the randomized 3-party communication complexity of $\neg f$ in the number-in-hand blackboard model with 1-sided error. A bicriteria low-rank approximation L with rank $k' = O\left((k/\epsilon)^2 \cdot 4^{R_\epsilon^{3,1-sided}(\neg f)}\right)$ achieving:*

$$\|W \circ (A - L)\|_F^2 \leq OPT + 2\epsilon\|A \circ W\|_F^2,$$

where $OPT = \inf_{\text{rank-}k\ \hat{L}} \|W \circ (A - \hat{L})\|_F^2$, can be computed in $O(\text{nnz}(A)) + n \cdot \text{poly}(k/\epsilon)$ time.

We give applications of Theorem 3 to low-rank plus diagonal tensor approximation, achieving $k' = O(k^2/\epsilon^4)$ and the low-rank plus sparse tensor approximation problem achieving $k' = O\left(\frac{k^2 \cdot t^4}{\epsilon^6}\right)$, where t is the maximum number of zeros on any face of W.

We also consider a common variant of low-rank approximation studied in data mining and information retrieval: *Boolean low-rank approximation* (binary low-rank approximation). Here one is given binary $A \in \{0,1\}^{n\times n}$ and seeks to find $U \in \{0,1\}^{n\times k}$ and $V \in \{0,1\}^{k\times n}$ minimizing $\|A - U \cdot V\|_0$ where $U \cdot V$ denotes Boolean matrix multiplication and $\|\cdot\|_0$ is the entrywise ℓ_0 norm, equal to the squared Frobenius norm in this case. While Boolean low-rank approximation is NP-hard [20, 26], there is a large body of work studying heuristic algorithms and approximation schemes, when no entries of A are masked [46, 59, 67, 10, 24]. We show in the full paper that any black-box algorithm for standard Boolean low-rank approximation yields a bicriteria algorithm for masked Boolean low-rank approximation, with rank depending on the *nondeterministic communication complexity* of the mask W.

▶ **Theorem 4** (Nondeterministic Communication Complexity → Boolean Low-Rank Approx)**.** *Let f be the function computed by W and $N(f)$ be the nondeterministic communication complexity of f. For any $k' \geq k \cdot 2^{N(f)}$, if one computes $U, V \in \{0,1\}^{n\times k'}$ satisfying $\|A \circ W - U \cdot V\|_0 \leq \min_{\hat{U},\hat{V}\in\{0,1\}^{n\times k'}} \|A \circ W - \hat{U} \cdot \hat{V}\|_0 + \Delta$ then:*

$$\|W \circ (A - U \cdot V)\|_0 \leq 2^{N(f)} \cdot OPT + \Delta,$$

where $OPT = \min_{\hat{U},\hat{V}\in\{0,1\}^{k\times n}} \|W \circ (A - U \cdot V)\|_0$ and $U \cdot V$ denotes Boolean matrix multiplication.

We can apply Theorem 4 for example, to the low-rank plus diagonal Boolean matrix approximation problem, where W is zero on its diagonal and one everywhere else. In this case we have $2^{N(f)} = \log n$ and correspondingly $k' = k \log n$.

1.2.4 Connections to Approximate Rank and Other Communication Lower Bounds

In Section 1.3 we sketch the proof of Theorem 1, which is very simple (Theorems 2, 3, and 4 are proved similarly.) The proof is based on covering W with $2^{R_\epsilon^{1-sided}(\neg f)}$ disjoint monochromatic rectangles, which match W on all but a small random subset of its 1 entries. The existence of a 1-sided error randomized communication protocol for $\neg f$ using $R_\epsilon^{1-sided}(\neg f)$ bits of communication is well known to imply the existence of such a covering with $2^{R_\epsilon^{1-sided}(\neg f)}$ rectangles. However, the optimal size of such a covering, which is known as the "public-coin partition bound" [34], may be lower than this. In fact, recent work has shown that it is provably smaller for some problems [27]. Thus, our algorithm can be stated in terms of this bound, giving improved results for these problems. However, as far as we are aware, this bound does not give any improvements for the communication problems we consider (corresponding to natural weight matrices W).

The public coin partition bound is a strengthening of the well-studied partition bound [33] for randomized communication complexity, which is itself a strengthening of the smooth rectangle bound [33]. This logarithm of the smooth rectangle bound is equivalent to the log approximate nonnegative rank of W up to constants [38]. It has been shown that the randomized communication complexity can be polynomially larger than the log partition bound [27]. Additionally, recent work refuting the log approximate rank conjecture [16] has shown that the randomized communication complexity can be exponentially larger than the log approximate nonnegative rank. Thus, improving our results to depend on these communication complexity lower bounds rather than the communication complexity itself would lead to potential improvements for some weight matrices W. However, all known separations are for W with complex structure and relatively high communication complexity, and thus not relevant to common applications. Additionally, it is unclear how to extend

our techniques to these weaker notions, or to other related notations, such as information complexity [14]. Such extensions would be interesting, e.g., connecting the difficulty of masked low-rank approximation to the approximate rank of the mask.

1.2.5 Lower Bounds

Given our results, and the above discussion, a natural question to ask is:

Is there a natural notion of the complexity of the mask W that characterizes the difficulty of the masked low-rank approximation problem?

We give some initial results, focused on how communication complexity in particular relates to the best bicriteria approximation factor for masked low-rank approximation achievable in polynomial time. We note that, since our results actually hold with rank depending on the public-coin partition bound [34], which has been separated from the randomized communication complexity, the communication complexity itself certainly does not tightly characterize the difficulty of masked low-rank approximation. However, we view our lower bounds in terms of communication complexity as a step in understanding this difficulty.

We prove two bounds based on a conjecture of the hardness of approximate 3-coloring. We show that there is a class of masks W such that any polynomial time algorithm achieving guarantee (1) and small enough ϵ requires bicriteria rank $k' = \Omega\left(\frac{D(f)}{\log D(f)}\right)$ where $D(f)$ is the deterministic communication complexity of f. Note that $D(f)$ is only greater than $R_\epsilon^{1-sided}(\neg f)$ and $R_\epsilon(f)$.

We strengthen this bound significantly for two natural variants of the masked low-rank approximation problem: when the low-rank approximation L is required to have a non-negative or binary factorization. We note that our techniques yield matching algorithmic results analogous to Theorems 1 and 2 for these variants. We show that for these variants on Problem 1, there is a class of masks W such that any polynomial time algorithm achieving guarantee (1) for small enough ϵ requires bicriteria rank which is exponential in the deterministic communication complexity, $k' = 2^{\Omega(D(\neg f))}$. This bound matches our algorithmic results for these variants. We note that in the parameter regimes considered (we just require rank $k = 3$), there exist polynomial time algorithms for the *non-masked* versions of binary and non-negative low-rank approximation. Thus, the hardness in terms of communication complexity comes from adding the mask to the low-rank cost function rather than the binary and non-negativity constraints themselves.

Our lower bounds are closely related to those of [30] on the hardness of bicriteria low-rank matrix completion. We note that for any $n \times n$ mask matrix W, we can always bound $D(f) = O(\log n)$. Thus, achieving a $2^{o(D(f))}$ bicriteria approximation factor means achieving an approximation factor sub-polynomial in n. [30] leaves open if achieving a $\sqrt{n}$ bicriteria approximation to rank-3 matrix completion is hard (Question 4.3 in [30]), and more generally asks what bicriteria approximation is achievable in polynomial time (Question 4.2 in [30]).

1.3 Our Techniques

The key ideas behind Theorems 1 and 2 are similar. We focus on Theorem 1 for exposition. We want to argue that any near optimal rank-k' approximation of $A \circ W$, gives a good bicriteria solution to the masked rank-k approximation problem. For simplicity, here we focus on showing this for the actual optimal rank-k' approximation, $L = \arg\min_{\text{rank}-k'\ \hat{L}} \|(A \circ W) - \hat{L}\|_F^2$. We show that $\|W \circ (A - L)\|_F^2 \leq OPT + O(\epsilon)\|A \circ W\|_F^2$ via a comparison method. Namely, we exhibit a rank k' matrix $\bar{L}$ that:

1. Nearly matches how well the optimum rank-k solution L_{opt} to Problem 1 approximates A on the support of W. In particular, $\|(A-\bar{L})\circ W\|_F^2 \leq \|(A-L_{opt})\circ W\|_F^2 + O(\epsilon)\|A\circ W\|_F^2$.
2. Places *no mass* outside the support of W. In particular, $\|\bar{L}\circ(1-W)\|_F^2 = 0$.

Since L minimizes the distance to $(A\circ W)$ among all rank-k' matrices, we have $\|(A\circ W) - L\|_F^2 \leq \|(A\circ W) - \bar{L}\|_F^2$. However, by (2), $\bar{L}$ *exactly matches* $A\circ W$ outside the support of W – both matrices are 0 there. Thus L must have at least as large error outside the support of W, and in turn cannot have larger error on the support of W. That is, we must have $\|(A-L)\circ W\|_F^2 \leq \|(A-\bar{L})\circ W\|_F^2$. Then by (1), L satisfies the desired bound.

1.3.1 From Communication Protocols to Low-Rank Approximations

The key question becomes how to exhibit $\bar{L}$, which we do using communication complexity. We view W as the communication matrix of some function $f : \{0,1\}^{\log n}\times\{0,1\}^{\log n}\to\{0,1\}$, with $W_{x,y} = f(x,y)$, where in f we interpret $x, y \in [n]$ as their binary representations. It is well-known that the existence of a deterministic communication protocol Π that computes f with $D(f)$ total bits of communication implies the existence of a partition of W into $2^{D(f)}$ *monochromatic combinatorial rectangles*. That is, there are $2^{D(f)}$ non-overlapping sets $R_i = S\times T$ for $S, T \in [n]$ that partition W and that satisfy $W(R_i)$ is either all 1 or all 0. We could construct $\bar{L}$ by taking the best k-rank approximation of each $A(R_i)$ where R_i is colored 1 (i.e., contains inputs with $f(x,y) = 1$). We could then sum up these approximations to produce $\bar{L}$ with rank $\leq k\cdot 2^{D(f)}$. Note that $\bar{L}$ is 0 outside the rectangles colored 1 – i.e., outside the support of W. Thus condition (2) above is satisfied. Further, $\bar{L}$ matches the optimal rank-k approximation on each R_i colored 1. So it approximates A at least as well as L_{opt} on these rectangles, and since these rectangles fully cover the support of W we have $\|(A-\bar{L})\circ W\|_F^2 \leq \|(A-L_{opt})\circ W\|_F^2$, giving the requirement of (1).

Unfortunately, essentially none of the W that are of interest in applications admit efficient deterministic communication protocols. $k' = k\cdot 2^{D(f)}$ will typically be larger than n, giving a vacuous bound. Thus we turn to randomized communication complexity with error probability ϵ, $R_\epsilon(f)$, which is much lower in these cases. A randomized protocol Π achieving this complexity corresponds to a distribution over partitions of W into $2^{R_\epsilon(f)}$ rectangles. These rectangles are not monochromatic but are close to it – letting W_Π be the communication matrix of the (random) function computed by the protocol, W_Π is partitioned into $2^{R_\epsilon(f)}$ monochromatic rectangles and further matches W on each (x,y) with probability at least $1-\epsilon$. We prove that, even with this small error, constructing $\bar{L}$ as above using the partition of W_Π instead of W itself gives a solution nearly matching L_{opt} up to small additive error. This error will involve $\|A\circ W\|_F^2$ and $\|L_{opt}\circ(1-W)\|_F^2$, depending on whether the protocol makes 1 or 2-sided error, as seen in Theorems 1 and 2.

1.3.2 Low-Rank Approximation to W Does Not Suffice

A natural view of our argument above is that the existence of an efficient randomized protocol for W implies the existence of a distribution over low-rank matrices (induced by partitions into near monochromatic rectangles) that match W on each entry with good probability. We note that this distributional view is critical – simply having a low-rank approximation to W matching all but a small fraction of entries does not suffice. The mistaken entries could in the worst case align with very heavy entries of A, which must be approximated well to solve masked low-rank approximation to small error. An approximation with small entrywise error (in the ℓ_∞ sense) would suffice. However, for important cases, e.g., when W is zero on the diagonal and one off the diagonal, such approximations provably require higher rank thank $2^{R_\epsilon(f)}$ and thus relying on them would lead to significantly weaker bounds [2].

1.3.3 Other Communication Models

In extending our results to other communication models, we first consider the connection between multiparty number-in-hand communication and tensor low-rank approximation. Protocols in this model correspond to a partition of the communication tensor $W \in \{0,1\}^{n\times n\times n}$ into $2^{R^3_\epsilon(f)}$ monochromatic (or nearly monochromatic) rectangles of the form $R_i = S \times T \times U$ for $S, T, U \subseteq [n]$, where $R^3_\epsilon(f)$ is the randomized 3-player communication complexity of W. We can again argue the existence of a rank $k' = k \cdot 2^{R^3_\epsilon(f)}$ tensor $\bar{L}$, obtained by taking a near optimal low-rank approximation to each rectangle colored 1 in W_Π, which is mostly 0 outside the support of W and at the same time competes with the best rank-k tensor approximation L_{opt} on the support of W. There are different notions of rank for tensors; here we mostly discuss canonical or CP rank. This lets us argue, as in the two player case, that the best rank-k' approximation of $A \circ W$ also competes with L_{opt}. It is not known how to find this best rank-k' approximation efficiently, however using an algorithm of [63] we can find a rank $k'' = O((k'/\epsilon)^2)$ bicriteria approximation achieving relative error $1+\epsilon$. Overall we have $k'' = O\left((k/\epsilon)^2 \cdot 2^{2R^3_\epsilon(f)}\right)$, giving Theorem 3.

We next consider the nondeterministic communication complexity. In a nondeterministic communication protocol for a function f, players can make "guesses" at any point during the protocol Π. The only requirement is that, (1) for every x, y with $f(x,y) = 1$, for some set of guesses made by the players, the protocol outputs $\Pi(x,y) = 1$ and (2) the protocol never outputs $\Pi(x,y) = 1$ for x, y with $f(x,y) = 0$. Such a protocol using $N(f)$ bits of communication corresponds to covering the communication matrix W with $2^{N(f)}$ *possibly overlapping* monochromatic rectangles. In many cases, the nondeterministic complexity is much lower than the randomized communication complexity. However, for low-rank approximation in the Frobenius norm, the overlap is a problem. We cannot construct $\bar{L}$ simply by approximating each rectangle and adding these approximations together. $\bar{L}$ will be too "heavy" where the rectangles overlap. However, for the Boolean low-rank approximation problem, the overlap is less of a problem. We simply construct $\bar{L}$ in the same way, letting it be the AND of the approximations on each rectangle. In the end, we obtain an error bound of roughly $2^{N(f)} \cdot OPT$, owing to the fact that error may still build up on the overlapping sections. Since there are $2^{N(f)}$ rectangles total, each entry is overlapped by at most $2^{N(f)}$ of them. However, since $N(f)$ can be very small, this result gives a tradeoff with Theorems 1 and 2 (which can also be extended to the Boolean case). For example, we show how to obtain error $\approx O(\log n \cdot OPT)$ for the Boolean low-rank plus diagonal approximation problem, with rank $k' = O(k \log n)$. This is smaller than the $O(k/\epsilon)$ achieved by Theorem 1 for small ϵ, which may be required to achieve good error if, e.g., $\|A\|_F^2$ is large.

1.3.4 An Alternative Approach

In the important cases when W is zero on its diagonal and one elsewhere or has a few non-zeros per row (the low-rank plus diagonal and low-rank plus sparse approximation problems, respectively) the existence of $\bar{L}$ satisfying the necessary conditions (1) and (2) above can be proven via a very different technique. The key idea is a structural result: that any low-rank matrix cannot concentrate too much weight on more than a few entries of its diagonal, or more generally, on a sparse support outside a few rows. Thus we can obtain $\bar{L}$ from L_{opt} by explicitly zero-ing out these few large entries falling outside the support of W (e.g., on its diagonal when W has zeros just on its diagonal). We detail this approach in the full paper, giving a bound matching Theorem 1 in this case. We show that the same structural result can also be used to obtain a fixed-parameter-tractable, relative error, non-bicriteria

approximation algorithm for Problem 1 in the low-rank plus diagonal case, as well as for the closely related factor analysis problem. We are unaware of any formal connection between this structural result and our communication complexity framework; however, establishing one would be very interesting.

1.4 Road Map

In Section 2 we give preliminaries, defining the communication models we use and giving communication complexity bounds for common mask matrices in these models. In Section 3 we then prove our main results, Theorems 1 and 2. We instantiate these results for the common mask matrices shown in Table 1. We defer our results connecting masked tensor approximation to multiparty communication complexity and Boolean low-rank approximation to nondeterministic communication complexity to the full paper – available at `https://arxiv.org/abs/1904.09841`. We also defer our lower bound results to the full version.

2 Preliminaries

2.1 Notation and Conventions

Throughout we use $\log z$ to denote the base-2 logarithm of z. For simplicity, so that we can associate any $W \in \mathbb{R}^{n\times n}$ with a function $f : \{0,1\}^{\log n} \times \{0,1\}^{\log n} \to \{0,1\}$ we assume that n is a power of 2 and so $\log n$ is an integer. Our results can be easily extended to general n. Given a matrix $M \in \mathbb{R}^{n\times n}$ and a combinatorial rectangle $R = S \times T$ for $S, T \subseteq [n]$, we let M_R denote the submatrix of M indexed by R. For matrix M we let $1 - M$ denote the matrix N with $N_{i,j} = 1 - M_{i,j}$. E.g., $1 - I$ is the matrix with all zeros on diagonal and all ones off diagonal.

While in the introduction we focus on low-rank approximation in the Frobenius norm, many of our results will apply to any entrywise matrix norm of the form:

▶ **Definition 5.** *An entrywise matrix norm $\|\cdot\|_\star : \mathbb{R}^{n\times n} \to \mathbb{R}$ is a function of the form:*

$$\|M\|_\star = \sum_{i=1}^{n}\sum_{j=1}^{n} g(|M_{i,j}|),$$

where $g : \mathbb{R} \to \mathbb{R}$ is some monotonically increasing nonnegative function.

$g(x) = x^2$ gives the squared Frobenius norm, $g(x) = x^p$ gives the entrywise ℓ_p norm, $g(x) = 1$ iff $x \neq 0$ gives the entrywise ℓ_0 norm, etc. See [62, 10, 17, 5] for a discussion of standard low-rank approximation algorithms for these norms. As discussed, our bicriteria results will simply require applying one of these algorithms to compute a near-optimal low-rank approximation to $A \circ W$ (i.e., A with the masked entries zeroed out).

2.2 Communication Complexity Models

We give a brief introduction to the communication models we consider, and refer the reader to the textbooks [39, 53] for more background. We mostly consider two-party communication of Boolean functions, though in the full paper discuss extensions to more than two parties.

Consider two parties, Alice and Bob, holding inputs $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ respectively. They exchange messages in order to compute a function $f : \mathcal{X} \times \mathcal{Y} \to \{0,1\}$ evaluated at (x, y). They would like to do this while minimizing the total number of bits exchanged. The communication between the parties is determined by a possibly randomized protocol, which

specifies the message of the next player to speak as a function of previous messages received by that player and that player's input. For a given protocol Π, we let $|\Pi(x,y)|$ denote the number of bits transmitted by the players on inputs x and y, and we let $|\Pi| = \max_{x,y} |\Pi(x,y)|$.

Let M be the communication matrix of f, that is, the matrix whose rows are indexed by elements of $\mathcal{X}$ and columns by elements of $\mathcal{Y}$, and for which $M_{x,y} = f(x,y)$. A well known and useful property is that Π *partitions* M into *rectangles* $R = S \times T$, where $S \subseteq \mathcal{X}$ and $T \subseteq \mathcal{Y}$, and every pair (x,y) of inputs with $(x,y) \in S \times T$ has the same output when running protocol Π. The number of rectangles in the partition is equal to $2^{|\Pi|}$. We call the unique output of Π on a rectangle $S \times T$ the *label* of the rectangle.

▶ **Definition 6** (Deterministic Communication Complexity)**.** *The deterministic communication complexity $D(f) = \min_\Pi |\Pi|$, where the minimum is taken over all protocols Π for which $\Pi(x,y) = f(x,y)$ for every pair (x,y) of inputs. Equivalently, $D(f)$ is the minimum number so that M can be partitioned via a protocol Π into $2^{D(f)}$ rectangles for which for every rectangle R and $b \in \{0,1\}$, if R is labeled b, then for all $(x,y) \in R$, $f(x,y) = b$.*

We next turn to randomized communication complexity. For the purposes of this paper, we will consider public coin randomized communication complexity, i.e., there is a shared random string r that both Alice and Bob have access to. In a randomized protocol Π, Alice and Bob see r and then run a deterministic protocol Π_r. We say a protocol Π is a (δ_1, δ_2)-error protocol if for all $x, y \in \mathcal{X} \times \mathcal{Y}$, with $f(x,y) = 1$, $\mathbb{P}_r[\Pi_r(x,y) = f(x,y)] \geq 1 - \delta_1$ and for all $x, y \in \mathcal{X} \times \mathcal{Y}$ with $f(x,y) = 0$, $\mathbb{P}_r[\Pi_r(x,y) = f(x,y)] \geq 1 - \delta_2$. We can then define a general notion of randomized communication complexity:

▶ **Definition 7** (Randomized Communication Complexity – General)**.** *The (δ_1, δ_2)-error randomized communication complexity $R_{\delta_1,\delta_2}(f) = \min_\Pi |\Pi|$, where the minimum is taken over all (δ_1, δ_2)-error protocols Π. Equivalently, $R_{\delta_1,\delta_2}(f)$ is the minimum number so that there is a distribution over protocols inducing partitions of M, each containing at most $2^{R_{\delta_1,\delta_2}(f)}$ rectangles, such that (1) for every $(x,y) \in \mathcal{X} \times \mathcal{Y}$ with $f(x,y) = 1$, with probability at least $1 - \delta_1$, (x,y) lands in a rectangle which is labeled 1 and (2) for every $(x,y) \in \mathcal{X} \times \mathcal{Y}$ with $f(x,y) = 0$, with probability at least $1 - \delta_2$, (x,y) lands in a rectangle which is labeled 0.*

Definition 7 is typically specialized to two cases: the randomized communication complexity with 2-sided error and the randomized communication complexity with 1-sided error.

▶ **Definition 8** (Randomized Communication Complexity – 2-sided)**.** *The δ-error randomized communication complexity of f is $R_\delta(f) \stackrel{\text{def}}{=} R_{\delta,\delta}(f)$.*

▶ **Definition 9** (Randomized Communication Complexity – 1-Sided)**.** *The δ-error 1-sided randomized communication complexity of f is $R_\delta^{1-sided}(f) \stackrel{\text{def}}{=} R_{0,\delta}(f)$.*

In Theorem 1 we consider the 1-sided communication complexity of $\neg f$: $R_{\delta,0}(f)$.

2.3 Specific Communication Bounds

We discuss a few problems that will be particularly useful for our applications. We only need communication upper bounds and in specific models. Note that in this section, as is standard, we state bounds for communication problems with n-bit inputs. In our applications to masked low-rank approximation, we will typically apply the bounds with input size $\log n$.

Equality

In the Equality problem, denoted EQ, there are two players Alice and Bob, holding strings $x, y \in \{0,1\}^n$, and the function $EQ(x,y) = 1$ if $x = y$, and $EQ(x,y) = 0$ otherwise.

▶ **Theorem 10** ([39], combining Corollaries 26 and 27 of [11]). *$R_\delta^{1-way}(EQ) \leq (1-\delta)\log((1-\delta)^2/\delta) + 5$, and $R_\delta^{1-way,1-sided}(EQ) \leq \log(1/\delta) + 5$.*

We also can bound the nondeterminitic communication complexity of inequality, i.e., the function $NEQ(x,y)$ with $NEQ(x,y) = 1$ iff $x \neq y$.

▶ **Theorem 11.** *$N(NEQ) \leq \lceil \log n \rceil + 2$.*

Proof. Alice simply guesses an index at which x and y differ and sends this index (using $\lceil \log n \rceil$ bits) along with the value of x at this index to Bob. Bob sends the value of y at this index and the players check if x and y differ at the index. ◀

Essentially the same protocol can be used to solve the negation of the disjointness problem, with $\neg DISJ(x,y) = 1$ only if there is some $k \in [n]$ with $x(k) = y(k) = 1$. We thus have:

▶ **Theorem 12** ([60]). *$N(\neg DISJ) \leq \lceil \log n \rceil + 2$.*

Greater-Than

In Greater-Than, denoted GT, there are two players Alice and Bob, holding integers $x, y \in \{0, 1, \ldots, n-1\}$, and the function $GT(x,y) = 1$ if $x > y$, and $GT(x,y) = 0$ otherwise.

▶ **Theorem 13** ([52]). *$R_\delta(GT) = O(\log(n/\delta))$.*

3 Bicriteria Approximation from Communication Complexity

In this section we prove our main results, Theorems 1 and 2, which connect the randomized communication complexity of the binary matrix W to the rank required to solve Problem 1 efficiently up to small additive error. We prove a general theorem connecting the rank to $R_{\delta_1,\delta_2}(f)$. Both Theorems 1 and 2 follow as corollaries if we consider the 1-sided error complexity $R_\delta^{1-sided}(\neg f) = R_{\delta,0}(f)$ and the 2-sided error complexity $R_\delta(f) \stackrel{\text{def}}{=} R_{\delta,\delta}(f)$ respectively (Definitions 8 and 9).

▶ **Theorem 14** (Randomized Communication Complexity → Bicriteria Approximation). *Consider $W \in \{0,1\}^{n\times n}$ and let f be the function computed by it. For $k' \geq k \cdot 2^{R_{\epsilon_1,\epsilon_2}(f)}$, and any entrywise norm $\|\cdot\|_\star$ (Def. 5), for any L satisfying $\|A \circ W - L\|_\star \leq \min_{rank-k'\ \hat{L}} \|A \circ W - \hat{L}\|_\star + \Delta$:*

$$\|(A-L)\circ W\|_\star \leq OPT + \epsilon_1\|A\circ W\|_\star + \epsilon_2\|L_{opt}\circ(1-W)\|_\star + \Delta,$$

where $OPT = \min_{\text{rank}-k\ \hat{L}} \|(A-\hat{L})\circ W\|_\star$ and L_{opt} is any rank-k matrix achieving OPT.

Proof. As discussed (Def. 7), $R_{\epsilon_1,\epsilon_2}(f)$ is the minimum number so that there is a distribution on protocols inducing partitions of W, each containing at most $2^{R_{\epsilon_1,\epsilon_2}(f)}$ rectangles, such that (1) for every $x, y \in \{0,1\}^{\log n}$ with $f(x,y) = 1$, (x,y) lands in a rectangle labeled 1 with probability $\geq 1-\epsilon_1$ and (2) for every $x, y \in \{0,1\}^{\log n}$ with $f(x,y) = 0$, (x,y) lands in a rectangle labeled 0 with probability $\geq 1-\epsilon_2$. In other words, letting W_Π be the (random) matrix corresponding to the function computed by the protocol: (1) $W \circ (1 - W_\Pi)$ has each entry equal to 1 with probability $\leq \epsilon_1$ and (2) $W_\Pi \circ (1-W)$ has each entry equal to 1 with probability $\leq \epsilon_2$. Thus, fixing some L_{opt}:

$$\mathop{\mathbb{E}}_{protocol\ \Pi}\left[\|A\circ W\circ(1-W_\Pi)\|_\star+\|L_{opt}\circ W_\Pi\circ(1-W)\|_\star\right]$$
$$\leq \epsilon_1\|A\circ W\|_\star+\epsilon_2\|L_{opt}\circ(1-W)\|_\star.$$

Thus, there is at least one protocol Π (inducing a partition with $\leq 2^{R_{\epsilon_1,\epsilon_2}(f)}$ rectangles) with:

$$\|A\circ W\circ(1-W_\Pi)\|_\star+\|L_{opt}\circ W_\Pi\circ(1-W)\|_\star\leq\epsilon_1\|A\circ W\|_\star+\epsilon_2\|L_{opt}\circ(1-W)\|_\star. \tag{2}$$

Let P_1 be the set of rectangles on which the protocol achieving (2) returns 1 and P_0 be the set on which it returns 0. For any $R\in P_1$ let $L^R=\arg\min_{\text{rank}-k\ \hat{L}}\|A_R\circ W_R-\hat{L}\|_\star$ (note that L^R is the size of R). Let $\bar{L}^R$ be the $n\times n$ matrix equal to L^R on R and 0 elsewhere. Let $\bar{L}=\sum_{R\in P_1}\bar{L}^R$. Note that $\bar{L}$ has rank at most $\sum_{R\in P_1}\text{rank}(\bar{L}^R)\leq k\cdot|P_1|\leq k\cdot 2^{R_{\epsilon_1,\epsilon_2}(f)}$. Thus, by the assumption that L satisfies $\|A\circ W-L\|_\star\leq\min_{rank-k'\ \hat{L}}\|A\circ W-\hat{L}\|_\star+\Delta$:

$$\begin{aligned}\|(A-L)\circ W\|_\star\leq\|A\circ W-L\|_\star&\leq\|A\circ W-\bar{L}\|_\star+\Delta\\&=\|(A\circ W-\bar{L})\circ W_\Pi\|_\star+\|(A\circ W-\bar{L})\circ(1-W_\Pi)\|_\star+\Delta\\&=\|(A\circ W-\bar{L})\circ W_\Pi\|_\star+\|A\circ W\circ(1-W_\Pi)\|_\star+\Delta,\end{aligned}\tag{3}$$

where the third line follows since $\bar{L}$ is 0 outside the support of W_Π (i.e., outside of the rectangles in P_1). Since $\bar{L}$ is equal to the best rank-k approximation to $A_R\circ W_R$ on each rectangle R in P_1, and since these rectangles partition the support of W_Π:

$$\begin{aligned}\|(A\circ W-\bar{L})\circ W_\Pi\|_\star&\leq\|(A\circ W-L_{opt})\circ W_\Pi\|_\star\\&=\|(A-L_{opt})\circ W\circ W_\Pi\|_\star+\|L_{opt}\circ(1-W)\circ W_\Pi\|_\star\\&\leq OPT+\|L_{opt}\circ(1-W)\circ W_\Pi\|_\star.\end{aligned}$$

Plugging back into (3) and applying (2):

$$\begin{aligned}\|(A-L)\circ W\|_\star&\leq OPT+\|L_{opt}\circ(1-W)\circ W_\Pi\|_\star+\|A\circ W\circ(1-W_\Pi)\|_\star+\Delta\\&\leq OPT+\epsilon_1\|A\circ W\|_\star+\epsilon_2\|L_{opt}\circ(1-W)\|_\star+\Delta,\end{aligned}$$

which completes the theorem. ◀

Proof of Theorems 1 and 2. Theorems 1 and 2 follow by applying Theorem 14 with $\epsilon_1=\epsilon_2=\epsilon$ and $\epsilon_1=\epsilon$, $\epsilon_2=0$ respectively, and noting that $\|A\circ W\|_\star\leq\|A\|_\star$ and $\|L_{opt}\circ(1-W)\|_\star\leq\|L_{opt}\|_\star$. When $\|\cdot\|_\star$ is the squared Frobenius norm, L satisfying $\|A\circ W-L\|_\star\leq\min_{rank-k'\ \hat{L}}\|A\circ W-\hat{L}\|_\star+\Delta$ for $\Delta=\epsilon\|(A\circ W)-(A\circ W)_{k'}\|_F^2\leq\epsilon\|A\circ W\|_F^2$ can be computed with high probability in $O(\text{nnz}(A))+n\cdot\text{poly}(k'/\epsilon)$ time. ◀

3.1 Applications of Main Theorem

We can instantiate Theorem 14 for a number of common mask patterns, yielding the results summarized in Table 1. Note that the additive error bounds achieved are stated in terms of $\|A\circ W\|_\star$ and $\|L_{opt}\circ(1-W)\|_\star$, which are only smaller than $\|A\|_\star$ and $\|L_{opt}\|_\star$ respectively.

We start with the case when W is the negation of a diagonal matrix or a block diagonal matrix, corresponding to the Low-Rank Plus Diagonal (LRPD) and Low-Rank Plus Block Diagonal (LRPBD) matrix approximation problems. The argument uses the communication complexity of Equality (EQ). A variant on this problem is also used when W has at most t nonzeros (or nonzero blocks) per row. This corresponds to the Low-Rank Plus Sparse (LRPS) and Low-Rank Plus Block Sparse (LRPBS) approximation problems, which strictly generalize the Low-Rank Plus (Block) Diagonal Problem. Proofs are given in the full paper.

▶ **Corollary 15** (Low-Rank Plus Diagonal Approximation). *Let $W = 1 - I$ where I is the $n \times n$ identity matrix. Then for $k' = O\left(\frac{k}{\epsilon}\right)$ and L with $\|A \circ W - L\|_\star \leq \min_{\text{rank}-k'\ \hat{L}} \|A \circ W - \hat{L}\|_\star + \epsilon\|A \circ W\|_\star$:*

$$\|(A - L) \circ W\|_\star \leq OPT + 2\epsilon\|A \circ W\|_\star.$$

If $\|\cdot\|_\star = \|\cdot\|_F^2$, L can be computed with high probability in $O(\text{nnz}(A)) + n\,\text{poly}(k/\epsilon)$ time.

Proof. The function f corresponding to W is the inequality function NEQ. We have $R_\epsilon^{1-sided}(\neg NEQ) = R_\epsilon^{1-sided}(EQ)$, which by Theorem 10 is bounded by $\log(1/\epsilon) + 5$. Thus $2^{R_\epsilon^{1-sided}(\neg NEQ)} \leq \frac{32}{\epsilon}$. The corollary then follows directly from Theorem 14. ◀

▶ **Corollary 16** (Low-Rank Plus Block Diagonal Approximation). *Consider any partition $B_1 \cup B_2 \cup \ldots \cup B_b = [n]$ and let W be the matrix with $W_{i,j} = 0$ if $i, j \in B_k$ for some k and $W_{i,j} = 1$ otherwise. Then for $k' = O\left(\frac{k}{\epsilon}\right)$ and L with $\|A \circ W - L\|_\star \leq \min_{\text{rank}-k'\ \hat{L}} \|A \circ W - \hat{L}\|_\star + \epsilon\|A \circ W\|_\star$:*

$$\|(A - L) \circ W\|_\star \leq OPT + 2\epsilon\|A \circ W\|_\star.$$

If $\|\cdot\|_\star = \|\cdot\|_F^2$, L can be computed with high probability in $O(\text{nnz}(A)) + n\,\text{poly}(k/\epsilon)$ time.

Proof. The function f corresponding to W is the inequality function NEQ where $x, y \in [n]$ are identified with $j, k \in [b]$ if block B_j contains x and B_k contains y. The randomized communication complexity $\neg f$ is thus bounded by the complexity of equality. By Theorem 10, $R_\epsilon^{1-sided}(EQ) \leq \log(1/\epsilon) + 5$ and so $2^{R_\epsilon^{1-sided}(f)} \leq \frac{32}{\epsilon}$, which gives the corollary. ◀

Beyond equality, a number of common sparsity patterns are related to the communication complexity of Greater-Than (GT), which is bounded by Theorem 13. Since two-sided error is required to give efficient GT protocols, we incur an additional error term depending on L_{opt}. An interesting question is if this is necessary for efficient bicriteria approximation.

▶ **Corollary 17** (Low-Rank Plus Banded Approximation). *For any integer $p \leq n$, let $W \in \{0,1\}^{n \times n}$ be the banded Toeplitz matrix with $W_{i,j} = 0$ iff $|i - j| < p$. Then for $k' = k \cdot \min\left(\frac{p}{\epsilon}, \text{poly}\left(\frac{\log n}{\epsilon}\right)\right)$ and L with $\|A \circ W - L\|_\star \leq \min_{\text{rank}-k'\ \hat{L}} \|A \circ W - \hat{L}\|_\star + \epsilon\|A \circ W\|_\star$:*

$$\|(A - L) \circ W\|_\star \leq OPT + 2\epsilon\|A \circ W\|_\star + \epsilon\|L_{opt} \circ (1 - W)\|_\star.$$

If $\|\cdot\|_\star = \|\cdot\|_F^2$, L can be computed with high probability in $O(\text{nnz}(A)) + n\,\text{poly}(k'/\epsilon)$ time.

Proof. The function f corresponding to W is the negation of the AND of $i + p < j$ and $j + p > i$. Thus, it can be solved with two calls to a protocol for Greater-Than (GT). By Theorem 13, for $\log n$ bit inputs, $R_\epsilon(GT) = O\left(\log\left(\frac{\log n}{\epsilon}\right)\right)$. Thus $R_\epsilon(f) = O\left(\log\left(\frac{\log n}{\epsilon}\right)\right)$ and $2^{R_\epsilon(f)} = \text{poly}\left(\frac{\log n}{\epsilon}\right)$, giving $k' = k \cdot \text{poly}\left(\frac{\log n}{\epsilon}\right)$. When p is small, we can apply our result for W with sparse rows (see full paper), which gives $k' = k \cdot \frac{p}{\epsilon}$, completing the corollary. ◀

In the full paper, we also consider a "multi-dimensional" banded pattern. Here each $i \in \{0,1\}^{\log n}$ corresponds to a point (i_1, i_2) in a $\sqrt{n} \times \sqrt{n}$ grid (i_1 and i_2 are determined by the first $\frac{\log n}{2}$ and last $\frac{\log n}{2}$ bits of i respectively). We focus on the two-dimensional case, achieving rank $k' = k \cdot \text{poly}\left(\frac{\log n}{\epsilon}\right)$ as in the 1-dimensional case. This set up can easily be generalized to higher dimensions. We can also imagine generalizing to different distance measures over the points (i_1, i_2) using efficient sketching methods (which yield efficient communication protocols) for various distances [6, 35].

A similar result holds for low-rank approximation with monotone missing data.

▶ **Corollary 18** (Monotone Missing Data Problem (MMDP)). *Let $W \in \{0,1\}^{n\times n}$ be any matrix where each row of W has a prefix of an arbitrary number of ones, followed by a suffix of zeros. Then for $k' = k \cdot \text{poly}\left(\frac{\log n}{\epsilon}\right)$ and L with $\|A \circ W - L\|_\star \leq \min_{\text{rank}-k'\ \hat{L}} \|A \circ W - \hat{L}\|_\star + \epsilon\|A \circ W\|_\star$:*

$$\|(A - L) \circ W\|_\star \leq OPT + 2\epsilon\|A \circ W\|_\star + \epsilon\|L_{opt} \circ (1 - W)\|_\star.$$

If $\|\cdot\|_\star = \|\cdot\|_F^2$, L can be computed with high probability in $O(\text{nnz}(A)) + n\,\text{poly}(k'/\epsilon)$ time.

Proof. Let p_x be the length of the prefix of ones in the x^{th} row of W. Then the function f corresponding to W is $f(x, y) = 1$ iff $p_x \geq y$. That is, it is just the Greater-Than function where Alice maps her input x to p_x. Thus by Theorem 13, $R_\epsilon(f) \leq R_\epsilon(GT) = O\left(\log\left(\frac{\log n}{\epsilon}\right)\right)$. So $2^{R_\epsilon(f)} = poly\left(\frac{\log n}{\epsilon}\right)$, which gives the corollary. ◀

4 Open Questions

By focusing on bicriteria approximation, we show how to solve masked low-rank approximation in polynomial time using a simple heuristic. A number of open questions remain. It would be very interesting to improve the bicriteria ranks we achieve for common masks (summarized in Table 1). It would also be interesting to give relative error bounds achieving error $(1+\epsilon)\cdot OPT$ instead of our additive error bounds. This is challenging sinces it requires achieving zero error when there is an exact masked low-rank factorization of A.

Relatedly, while we have connected bicriteria masked low-rank approximation to the randomized communication complexity of the mask matrix W (in fact, the public coin partition number of W), it would be very interesting to find a notion of W's complexity that tightly characterizes the bicriteria rank achievable in polynomial time. We make some initial steps via lower bounds in terms of the communication complexity in the full paper, however the question remains mostly unanswered.

Finally, a related problem is *weighted low-rank approximation* – when W is real valued and we seek to minimize $\|W \circ (A - L)\|_F^2$. Approximation algorithms depending exponentially on the rank k, error parameter ϵ, and notions of W's complexity, such as its rank or number of distinct columns are known [54]. However, it would be very interesting to give polynomial time bicriteria approximation algorithms as we have done in the special case of binary W.

References

1 SAS/STAT(R) 9.22 User's guide. See Figure 54.7. URL: `https://support.sas.com/documentation/cdl/en/statug/63347/HTML/defaultviewer.htm#statug_mi_sect017.htm`.

2 Noga Alon. Perturbed identity matrices have high rank: Proof and applications. *Combinatorics, Probability & Computing*, 18(1-2):3, 2009.

3 Yossi Azar, Amos Fiat, Anna Karlin, Frank McSherry, and Jared Saia. Spectral analysis of data. In *Proceedings of the 33rd Annual ACM Symposium on Theory of Computing (STOC)*, 2001.

4 Masoumeh Azghani and Sumei Sun. Low-rank block sparse decomposition algorithm for anomaly detection in networks. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 807–810. IEEE, 2015.

5 Frank Ban, Vijay Bhattiprolu, Karl Bringmann, Pavel Kolev, Euiwoong Lee, and David P Woodruff. A PTAS for ℓ_p-low rank approximation. In *Proceedings of the 30th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 747–766, 2019.

6 Ziv Bar-Yossef, TS Jayram, Robert Krauthgamer, and Ravi Kumar. Approximating edit distance efficiently. In *Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 550–559, 2004.

7 S. Beghelli, R.P. Guidorzi, and U. Soverini. The Frisch scheme in dynamic system identification. *Automatica*, 26(1):171–176, 1990.

8 Peter Benner, Venera Khoromskaia, and Boris N Khoromskij. A reduced basis approach for calculation of the Bethe–Salpeter excitation energies by using low-rank tensor factorisations. *Molecular Physics*, 114(7-8):1148–1161, 2016.

9 Aditya Bhaskara, Silvio Lattanzi, Sergei Vassilvitskii, and Morteza Zadimoghaddam. Residual based sampling for online low rank approximation. In *Proceedings of the 60th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 1596–1614, 2019.

10 Karl Bringmann, Pavel Kolev, and David Woodruff. Approximation algorithms for ℓ_0-low rank approximation. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pages 6648–6659, 2017.

11 Joshua Brody, Amit Chakrabarti, Ranganath Kondapally, David P. Woodruff, and Grigory Yaroslavtsev. Certifying equality with limited interaction. *Algorithmica*, 76(3):796–845, 2016.

12 Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):11, 2011.

13 Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717, 2009.

14 Amit Chakrabarti, Yaoyun Shi, Anthony Wirth, and Andrew Yao. Informational complexity and the direct sum problem for simultaneous message complexity. In *Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 270–278, 2001.

15 Venkat Chandrasekaran, Sujay Sanghavi, Pablo A Parrilo, and Alan S Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

16 Arkadev Chattopadhyay, Nikhil S Mande, and Suhail Sherif. The log-approximate-rank conjecture is false. In *Proceedings of the 51st Annual ACM Symposium on Theory of Computing (STOC)*, pages 42–53, 2019.

17 Flavio Chierichetti, Sreenivas Gollapudi, Ravi Kumar, Silvio Lattanzi, Rina Panigrahy, and David P Woodruff. Algorithms for ℓ_p low-rank approximation. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 806–814, 2017.

18 Kenneth L Clarkson and David P Woodruff. Input sparsity and hardness for robust subspace approximation. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 310–329, 2015. `arXiv:1510.06073`.

19 Kenneth L Clarkson and David P Woodruff. Input sparsity and hardness for robust subspace approximation. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2015.

20 Chen Dan, Kristoffer Arnsfelt Hansen, He Jiang, Liwei Wang, and Yuchen Zhou. On low rank approximation of binary matrices. *arXiv*, 2015. `arXiv:1511.01699`.

21 A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

22 Amit Deshpande and Kasturi R. Varadarajan. Sampling-based dimension reduction for subspace approximation. In *Proceedings of the 39th Annual ACM Symposium on Theory of Computing (STOC)*, pages 641–650, 2007.

23 Dan Feldman, Amos Fiat, Micha Sharir, and Danny Segev. Bi-criteria linear-time approximations for generalized k-mean/median/center. In *Proceedings of the 23rd Annual Symposium on Computational Geometry (SCG)*, pages 19–26, 2007.

24 Fedor V Fomin, Petr A Golovach, and Fahad Panolan. Parameterized low-rank binary matrix approximation. In *Proceedings of the 45th International Colloquium on Automata, Languages and Programming (ICALP)*, 2018.

25 Silvia Gandy, Benjamin Recht, and Isao Yamada. Tensor completion and low-rank tensor recovery via convex optimization. *Inverse Problems*, 27(2):025010, 2011.

26 Nicolas Gillis and Stephen A Vavasis. On the complexity of robust PCA and ℓ_1-norm low-rank matrix approximation. *Mathematics of Operations Research*, 43(4):1072–1084, 2018.

27 Mika Göös, TS Jayram, Toniann Pitassi, and Thomas Watson. Randomized communication vs. partition number. In *Proceedings of the 44th International Colloquium on Automata, Languages and Programming (ICALP)*, 2017.

28 Leslie Greengard and John Strain. The fast Gauss transform. *SIAM Journal on Scientific and Statistical Computing*, 12(1):79–94, 1991.

29 Charles Guyon, Thierry Bouwmans, and El-Hadi Zahzah. Foreground detection based on low-rank and block-sparse matrix decomposition. *Proceedings of the 19th IEEE International Conference on Image Processing*, pages 1225–1228, 2012.

30 Moritz Hardt, Raghu Meka, Prasad Raghavendra, and Benjamin Weitz. Computational limits for matrix completion. In *Proceedings of the 27th Annual Conference on Computational Learning Theory (COLT)*, pages 703–725, 2014.

31 Daniel Hsu, Sham M Kakade, and Tong Zhang. Robust matrix decomposition with sparse corruptions. *IEEE Transactions on Information Theory*, 57(11):7221–7234, 2011.

32 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. ACM, 2013.

33 Rahul Jain and Hartmut Klauck. The partition bound for classical communication complexity and query complexity. In *Proceedings of the 25th Annual IEEE Conference on Computational Complexity (CCC)*, pages 247–258, 2010.

34 Rahul Jain, Troy Lee, and Nisheeth K Vishnoi. A quadratically tight partition bound for classical communication complexity and query complexity. *arXiv*, 2014. `arXiv:1401.4512`.

35 Daniel M Kane, Jelani Nelson, and David P Woodruff. On the exact space complexity of sketching and streaming small norms. In *Proceedings of the 21st Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1161–1178, 2010.

36 Ravi Kannan and Santosh Vempala. Spectral algorithms. *Foundations and Trends in Theoretical Computer Science*, 2009.

37 Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

38 Gillat Kol, Shay Moran, Amir Shpilka, and Amir Yehudayoff. Approximate nonnegative rank is equivalent to the smooth rectangle bound. In *Proceedings of the 41st International Colloquium on Automata, Languages and Programming (ICALP)*, pages 701–712, 2014.

39 Eyal Kushilevitz and Noam Nisan. *Communication complexity*. Cambridge University Press, 1997.

40 Anastasios Kyrillidis and Volkan Cevher. Matrix ALPS: Accelerated low rank and sparse matrix reconstruction. In *2012 IEEE Statistical Signal Processing Workshop (SSP)*, pages 185–188, 2012.

41 Qiuwei Li, Gongguo Tang, and Arye Nehorai. Robust principal component analysis based on low-rank and block-sparse matrix decomposition. *Handbook of Robust Low-Rank and Sparse Matrix Decomposition: Applications in Image and Video Processing*, 2016.

42 Shuangjiang Li, Wei Wang, Hairong Qi, Bulent Ayhan, Chiman Kwan, and Steven Vance. Low-rank tensor decomposition based anomaly detection for hyperspectral imagery. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 4525–4529, 2015.

43 Ji Liu, Przemyslaw Musialski, Peter Wonka, and Jieping Ye. Tensor completion for estimating missing values in visual data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):208–220, 2013.

44 Antoine Liutkus and Kazuyoshi Yoshii. A diagonal plus low-rank covariance model for computationally efficient source separation. In *27th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2017.

45 Canyi Lu, Jiashi Feng, Yudong Chen, Wei Liu, Zhouchen Lin, and Shuicheng Yan. Tensor robust principal component analysis: Exact recovery of corrupted low-rank tensors via convex optimization. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5249–5257, 2016.

46 Haibing Lu, Jaideep Vaidya, and Vijayalakshmi Atluri. Optimal Boolean matrix decomposition: Application to role engineering. In *Proceedings of the 24th IEEE International Conference on Data Engineering*, pages 297–306, 2008.

47 Michael W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2):123–224, 2011.

48 Antonio Valerio Miceli Barone. Low-rank passthrough neural networks. In *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, pages 77–86. Association for Computational Linguistics, 2018.

49 Andrew C Miller, Nicholas J Foti, and Ryan P Adams. Variational boosting: Iteratively refining posterior approximations. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

50 Cun Mu, Bo Huang, John Wright, and Donald Goldfarb. Square deal: Lower bounds and improved relaxations for tensor recovery. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 73–81, 2014.

51 Praneeth Netrapalli, U. N. Niranjan, Sujay Sanghavi, Animashree Anandkumar, and Prateek Jain. Non-convex robust PCA. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 1107–1115, 2014. URL: `http://papers.nips.cc/paper/5430-non-convex-robust-pca`.

52 Noam Nisan. The communication complexity of threshold gates. *Combinatorics, Paul Erdos is Eighty*, 1:301–315, 1993.

53 Anup Rao and Amir Yehudayoff. Communication complexity and applications (early draft), 2019. URL: `https://homes.cs.washington.edu/~anuprao/pubs/book.pdf`.

54 Ilya Razenshteyn, Zhao Song, and David P Woodruff. Weighted low rank approximations with provable guarantees. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing (STOC)*, 2016.

55 Vladimir Rokhlin. Rapid solution of integral equations of classical potential theory. *Journal of Computational Physics*, 60(2):187–207, 1985.

56 Donald B. Rubin and Dorothy T. Thayer. EM algorithms for ML factor analysis. *Psychometrika*, 47(1):69–76, 1982.

57 James Saunderson, Venkat Chandrasekaran, Pablo A Parrilo, and Alan S Willsky. Diagonal and low-rank matrix decompositions, correlation matrices, and ellipsoid fitting. *SIAM Journal on Matrix Analysis and Applications*, 33(4):1395–1416, 2012.

58 Tselil Schramm and Benjamin Weitz. Low-rank matrix completion with adversarial missing entries. *CoRR*, 2015.

59 Bao-Hong Shen, Shuiwang Ji, and Jieping Ye. Mining discrete patterns via binary matrix factorization. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 757–766, 2009.

60 Alexander Sherstov. Lecture notes for CS 289A Communication Complexity, 2012. URL: `http://web.cs.ucla.edu/~sherstov/teaching/2012-winter/docs/lecture05.pdf`.

61 Edward Snelson and Zoubin Ghahramani. Sparse Gaussian processes using pseudo-inputs. *Advances in Neural Information Processing Systems 18 (NeurIPS)*, 18:1257–1264, 2005.

62 Zhao Song, David P Woodruff, and Peilin Zhong. Low rank approximation with entrywise ℓ_1-norm error. In *Proceedings of the 49th Annual ACM Symposium on Theory of Computing (STOC)*, pages 688–701, 2017.

63 Zhao Song, David P Woodruff, and Peilin Zhong. Relative error tensor low rank approximation. In *Proceedings of the 30th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2772–2789. SIAM, 2019.

64 Charles Spearman. "General Intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2):201–292, 1904.

65 Michael L. Stein. Limitations on low-rank approximations for covariance matrices of spatial data. *Spatial Statistics*, 8:1–19, 2014.

66 Jos MF Ten Berge and Henk AL Kiers. A numerical approach to the approximate and the exact minimum rank of a covariance matrix. *Psychometrika*, 56(2):309–315, 1991.

67 Jaideep Vaidya. Boolean matrix decomposition problem: theory, variations and applications to data engineering. In *Proceedings of the 28th IEEE International Conference on Data Engineering*, pages 1222–1224, 2012.

68 Stef van Buuren. *Flexible imputation of missing data*. Chapman and Hall/CRC, 2018. URL: `https://stefvanbuuren.name/fimd/missing-data-pattern.html`.

69 Shusen Wang, Chao Zhang, Hui Qian, and Zhihua Zhang. Improving the modified Nyström method using spectral shifting. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2014.

70 David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1-2):1–157, 2014.

71 John Wright, Arvind Ganesh, Shankar Rao, Yigang Peng, and Yi Ma. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada*. Curran Associates, Inc., 2009.

72 Changjiang Yang, Ramani Duraiswami, Nail A Gumerov, and Larry Davis. Improved fast Gauss transform and efficient kernel density estimation. In *Proceedings of the 9th IEEE International Conference on Computer Vision*, page 464, 2003.

73 Xiao Zhang, Lingxiao Wang, and Quanquan Gu. A unified framework for nonconvex low-rank plus sparse matrix recovery. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 84, 2018.

74 Yong Zhao, Jinyu Li, and Yifan Gong. Low-rank plus diagonal adaptation for deep neural networks. In *Proceedings of the 2016 International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5005–5009, 2016.