Low Rank Approximation of Binary Matrices: Column Subset Selection and Generalizations

Chen Dan[1]
Carnegie Mellon University, Pittsburgh, Pennsylvania, United States
cdan@cs.cmu.edu

Kristoffer Arnsfelt Hansen
Department of Computer Science, Aarhus University, Aarhus, Denmark
arnsfelt@cs.au.dk
https://orcid.org/0000-0002-1155-8072

He Jiang[1]
University of Southern California, Los Angeles, California, United States
jian567@usc.edu
https://orcid.org/0000-0002-4902-2206

Liwei Wang[2]
1. Key Laboratory of Machine Perception, MOE, School of EECS, Peking University;
2. Center for Data Science, Peking University, Beijing Institute of Big Data Research;
Beijing, China
wanglw@cis.pku.edu.cn

Yuchen Zhou[1]
Department of Statistics, University of Wisconsin-Madison, Madison, Wisconsin, United States
yuchenzhou@stat.wisc.edu

Abstract

Low rank approximation of matrices is an important tool in machine learning. Given a data matrix, low rank approximation helps to find factors, patterns, and provides concise representations for the data. Research on low rank approximation usually focuses on real matrices. However, in many applications data are binary (categorical) rather than continuous. This leads to the problem of low rank approximation of binary matrices. Here we are given a $d \times n$ binary matrix $\mathbf{A}$ and a small integer $k < d$. The goal is to find two binary matrices $\mathbf{U}$ and $\mathbf{V}$ of sizes $d \times k$ and $k \times n$ respectively, so that the Frobenius norm of $\mathbf{A} - \mathbf{UV}$ is minimized. There are two models of this problem, depending on the definition of the dot product of binary vectors: The GF(2) model and the Boolean semiring model. Unlike low rank approximation of a real matrix which can be efficiently solved by Singular Value Decomposition, we show that approximation of a binary matrix is NP-hard, even for $k = 1$.

In this paper, our main concern is the problem of Column Subset Selection (CSS), in which the low rank matrix $\mathbf{U}$ must be formed by k columns of the data matrix, and we are interested in the approximation ratio achievable by CSS for binary matrices. For the GF(2) model, we show that CSS has approximation ratio bounded by $\frac{k}{2} + 1 + \frac{k}{2(2^k-1)}$ and this is asymptotically tight. For the Boolean model, it turns out that CSS is no longer sufficient to obtain a bound. We then develop a Generalized CSS (GCSS) procedure in which the columns of $\mathbf{U}$ are generated from Boolean formulas operating bitwise on selected columns of the data matrix. We show that the approximation ratio achieved by GCSS is bounded by $2^{k-1} + 1$, and argue that an exponential dependency on k is seems inherent.

[1] Work done while at Peking University.
[2] Partially supported by National Basic Research Program of China (973 Program) (grant no. 2015CB352502).

43rd International Symposium on Mathematical Foundations of Computer Science (MFCS 2018).
Editors: Igor Potapov, Paul Spirakis, and James Worrell; Article No. 41; pp. 41:1–41:16

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

2012 ACM Subject Classification Theory of computation → Approximation algorithms analysis, Theory of computation → Unsupervised learning and clustering, Computing methodologies → Factorization methods,

Keywords and phrases Approximation Algorithms, Low Rank Approximation, Binary Matrices

Digital Object Identifier 10.4230/LIPIcs.MFCS.2018.41

Related Version A full version of the paper is available at [12], `http://arxiv.org/abs/1511.01699`.

1 Introduction

Low rank approximation of matrices is a classical problem. Given a matrix $\mathbf{A}$ of size $d \times n$, the goal is to find two low rank matrices $\mathbf{U}$ and $\mathbf{V}$, such that $\mathbf{UV}$ approximates $\mathbf{A}$. Formally, the problem is to solve the equation $\min_{\mathbf{U},\mathbf{V}} \|\mathbf{A} - \mathbf{UV}\|_F^2$, where the minimum is over all matrices $\mathbf{U}, \mathbf{V}$ of sizes $d \times k$ and $k \times n$ respectively. The parameter k, typically a small integer, is the desired rank. The error is measured in terms of the Frobenius norm $\|\cdot\|_F$.

In many applications, $\mathbf{A}$ is a *data matrix*: Each column of $\mathbf{A}$ is a d-dimensional data vector, and each row of $\mathbf{A}$ corresponds to an attribute. In the literature, low rank approximation of $\mathbf{A}$ is often called factor analysis or dimensionality reduction: the k columns of the matrix $\mathbf{U}$ are the *factors* or *basis vectors* of the low dimensional space, and each column of $\mathbf{V}$ contains the combination coefficients.

If $\mathbf{A}$, $\mathbf{U}$, $\mathbf{V}$ are real matrices, low rank approximation can be efficiently solved by Singular Value Decomposition (SVD). This problem has been studied for more than a century, and is known as Principal Component Analysis (PCA) [28], Karhunen-Loève Transform [30], to name a few.

In this paper we consider low rank approximation of binary matrices. The motivation is that in many applications data are binary (categorical) rather than continuous. Indeed, nearly half of the data sets in the UCI repository contains categorical features. In the binary case, we require that the data matrix $\mathbf{A}$ *as well as* the rank-k matrices $\mathbf{U}, \mathbf{V}$ are binary. There are two natural formulations of the binary low rank approximation problem, depending on the definition of vector dot product. One formulation will be referred to as the GF(2) model, in which the dot product of two binary vectors $\mathbf{u}, \mathbf{v}$ is defined as $\mathbf{u}^T\mathbf{v} := \oplus_i u_i v_i$. The other formulation will be referred to as the *Boolean* model, in which the dot product is defined as $\mathbf{u}^T\mathbf{v} := \bigvee_i (u_i \wedge v_i)$.

The Boolean model is usually called Boolean Factor Analysis (BFA). It has found numerous applications in machine learning and data mining including latent variable analysis, topic models, association rule mining, clustering, and database tiling [3, 33, 38, 40, 44]. The GF(2) model, while being less studied, has been applied to Independent Component Analysis (ICA) over string data, attracting attention from the signal processing community [25, 35, 48].

Despite of various applications and heuristic algorithms [19, 21, 31, 33], little is known from a theoretical point of view about the binary low rank approximation problem. In fact, previously the only known result is that for the very special case of $k = 1$ (where the GF(2) and the Boolean model are equivalent) there are 2-approximation algorithms (see Section 1.1).

In this paper, we provide the *first* theoretical results for the general binary low rank approximation problem, which is formally stated as follows. Given $\mathbf{A} \in \{0,1\}^{d\times n}$, solve

$$\min_{\mathbf{U}\in\{0,1\}^{d\times k},\mathbf{V}\in\{0,1\}^{k\times n}} \|\mathbf{A}-\mathbf{UV}\|_F^2. \tag{1}$$

where the matrix product $\mathbf{UV}$ is over GF(2) or the Boolean semiring respectively.

Before stating the results, let us first consider the differences between low rank approximation of real matrices and our GF(2) and Boolean models. First, the linear space over GF(2) has a very different structure from the Euclidean space. The dot product over GF(2) is not an inner product and does not induce a norm: there exists $\mathbf{a} \neq \mathbf{0}$ such that $\mathbf{a}^T\mathbf{a} = 0$ over GF(2). An immediate consequence is that for binary matrices of the GF(2) model, there is no Singular Value Decomposition (SVD), which is the basis for low rank approximation of real matrices. The Boolean model is even more different: As it is a semiring rather than a field, we do not even have a linear space (see below for details).

Thus the methodologies from the setting of real matrices do not carry over to the setting of binary matrices. In fact, we will show that finding the exact solution of (1) is NP-hard even for $k = 1$ (see Section 4). This result was obtained independently by Gillis and Vavasis [22].

Another well-studied approach for low rank approximation of matrices is Column Subset Selection (CSS) [20, 32]. The goal of CSS is to find a subset of k columns of $\mathbf{A}$ and form the low rank basis matrix so that the residual is as small as possible. An advantage of CSS is that the result is more interpretable than that of SVD. CSS has been extensively studied for low rank approximation of real matrices [1, 5, 6, 10, 11, 13–16, 24, 36, 43, 46, 47]. Below is a formal definition of CSS over real matrices.

▶ **Definition 1** (CSS for real matrices)**.** Given a matrix $\mathbf{A} \in \mathbb{R}^{d\times n}$ and a positive integer k, pick k columns of $\mathbf{A}$ forming a matrix $\mathbf{P}_A \in \mathbb{R}^{d\times k}$ such that the residual

$$\|\mathbf{A}-\mathbf{P}_A\mathbf{Q}\|_\xi$$

is minimized over all possible $\binom{n}{k}$ choices for the matrix $\mathbf{P}_A$. Here $\mathbf{Q}$ denotes the optimal matrix of size $k \times n$ given $\mathbf{P}_A$, which can be obtained by solving a least squares problem, and $\xi = 2$ or F denotes the spectral norm or Frobenius norm.

The central problem in CSS is to determine the best function $\phi(n, k)$ of n, k satisfying

$$\|\mathbf{A}-\mathbf{P}_A\mathbf{Q}\|_\xi^2 \leq \phi(k,n)\|\mathbf{A}-\mathbf{A}_k\|_\xi^2 \ , \tag{2}$$

where $\mathbf{A}_k$ denotes the best rank-k approximation to the matrix $\mathbf{A}$ as computed with SVD. Two classical results [14, 24] shows that for real matrices we have

$$\|\mathbf{A}-\mathbf{P}_A\mathbf{Q}\|_2^2 \leq (k(n-k)+1)\|\mathbf{A}-\mathbf{A}_k\|_2^2 \ , \tag{3}$$

$$\|\mathbf{A}-\mathbf{P}_A\mathbf{Q}\|_F^2 \leq (k+1)\|\mathbf{A}-\mathbf{A}_k\|_F^2 \ . \tag{4}$$

There is extensive work on developing efficient algorithms for CSS with approximation ratio close to the above bounds, possibly using more than k columns of $\mathbf{A}$. These include methods such as rank revealing QR [36], adaptive sampling [15], subspace sampling (leverage scores) [6, 16], efficient volume sampling [13], projection-cost preserving sketches [11] and greedy CSS [1].

In this work, we study the CSS problem for binary matrices over GF(2) and Boolean semiring respectively. We consider the central problem expressed by Eq. (2) and aiming to determine the best $\phi(k, n)$. We only consider the Frobenius norm, since the spectral norm does not exist in the GF(2) and Boolean models.

The difficulty of the CSS problem for GF(2) and Boolean semiring model is that all methods developed for CSS over real matrices rely on at least one of the following concepts which are intrinsic to the Euclidean space: SVD, volume of a simplex, Euclidean distance, orthogonal projection, and QR decomposition. However, none of these concept exists in the GF(2) or Boolean models.

In this paper, we develop new methods for the CSS problem for GF(2) and Boolean model respectively. For GF(2) model, we show that by picking the best k columns of $\mathbf{A}$ to form $\mathbf{P}_A$, we achieve the bound

$$\|\mathbf{A} - \mathbf{P}_A\mathbf{Q}\|_F^2 \leq \left(\frac{k}{2} + 1 + \frac{k}{2(2^k - 1)}\right) \|\mathbf{A} - \mathbf{A}_k\|_F^2,$$

where $\mathbf{A}_k = \mathbf{UV}$ is the optimal solution of (1). Moreover, we show that the ratio is asymptotically tight.

For Boolean model, it turns out that basic CSS is no longer sufficient for obtaining a bound, simply because the Boolean semiring is not a field. We instead propose a Generalized CSS (GCSS) procedure. In this GCSS framework, we select a larger number of columns of $\mathbf{A}$ and potential basis matrices $\mathbf{P}_A$ are generated from these using carefully designed Boolean formulas operating bitwise on the chosen columns of $\mathbf{A}$. We show that GCSS based on $(2^k - 1)$ columns of $\mathbf{A}$ achieves approximation ratio $(2^{k-1} + 1)$ relative to $\|\mathbf{A} - \mathbf{A}_k\|_F^2$. Moreover, we argue that the exponential dependence in k seems inherent with the Boolean model (see Section 3 for details).

Our work is a first step towards a good understanding of low rank approximation of matrices over GF(2) and the Boolean semiring. While our work gives approximation algorithms for low rank approximation for both the GF(2) and the Boolean model, our work is should mainly by viewed as existence results for (Generalized) CSS for binary matrices, parallel to the classical existence theorems [14, 24] for CSS of real matrices stated in Eq. (3) and (4). Moreover, as SVD does not apply to the GF(2) or Boolean model, CSS is so far the only method that obtains a low rank approximation for binary matrices with theoretical guarantees and deserves an in-depth study. Finally, it is an important future direction to develop efficient algorithms to achieve or approximately achieve the ratios obtained in this paper. We believe this requires new techniques futher exploiting the algebraic structure of GF(2) and the Boolean semiring.

The rest of this paper is organized as follows. In Section 1.1 we discuss existing results on low rank approximation of binary matrices. In Section 2 we present the information-theoretically optimal upper bound for the approximation ratio of CSS over GF(2). In Section 3 we propose the GCSS procedure and give the upper bound for the Boolean semiring model. In Section 4 we show that finding the exaction solution of the low rank binary matrix approximation problem is NP-hard even for $k = 1$. Finally we give our conclusion in Section 5.

1.1 Other Related Works

To the best of our knowledge, all known theoretical results on the low rank approximation problem are about the special case of rank-one, i.e., $k = 1$. In the rank-one case, one looks for binary vectors $\mathbf{u}$, $\mathbf{v}$ such that $\|\mathbf{A} - \mathbf{uv}^T\|_F$ is minimized, and the GF(2) and Boolean models are therefore equivalent.

Shen et al. [39] formulate the rank-one problem as an integer linear program and showed that solving its linear programming relaxation yields a 2-approximation algorithm. They also improved the efficiency by reducing the linear program to a max-flow problem using

a technique developed in [26]. Jiang et al. [29] observed that for the rank-one case, simply choosing the best column from $\mathbf{A}$ yields a 2-approximation algorithm.

In the GF(2) model, low rank approximation is related to the concept of matrix rigidity introduced by Valiant [45], as a method of proving lower bounds for linear circuits. For a matrix $\mathbf{A}$ over GF(2), the rigidity $R_{\mathbf{A}}(k)$ is the smallest number of entries of $\mathbf{A}$ that must be changed in order to bring its rank down to k. Thus for a $d \times n$ matrix $\mathbf{A}$, $R_{\mathbf{A}}(k)$ is *precisely* the minimum approximation error possible by a product of a $d \times k$ matrix $\mathbf{U}$ and a $k \times n$ matrix $\mathbf{V}$. By the results of Valiant, an $n \times n$ matrix $\mathbf{A}$ for which $R_{\mathbf{A}}(k) \geq n^{1+\varepsilon}$, for $k = O(n/\log\log n)$ and for some constant $\varepsilon > 0$ cannot be computed by a linear circuit of size $O(n)$ and depth $O(\log n)$. Such rigid matrices exists in abundance – the challenge is to come up with an explicit construction of a family of rigid matrices. For the low rank approximation problem we are however interested in the setting of $k \ll n$ and we are interested in algorithms rather than explicit matrices.

2 Column Subset Selection for Binary Matrices Over GF(2)

In this section we characterize the best possible approximation ratio of CSS in the GF(2) model. As mentioned in Section 1, the best approximation ratio of CSS for real matrices is $k+1$ under the Frobenius norm. This result is proved by the so-called volume sampling method [14]. Concretely, the volume sampling method randomly samples a set of k columns of $\mathbf{A}$ with probability proportional to the volume of the k-dimensional simplex formed by the k-columns along with the origin. Volume sampling generates an (expected) $k+1$ approximation ratio.

However, the GF(2) model does not have a notion of volume, since the dot product over GF(2) is not an inner product. Nevertheless, we develop a new approach and show the following bound.

▶ **Theorem 2.** *For any binary matrix* $\mathbf{A} \in \{0,1\}^{d\times n}$*, there exist* $\mathbf{P}_A \in \{0,1\}^{d\times k}$ *and* $\mathbf{Q} \in \{0,1\}^{k\times n}$*, where the columns of* $\mathbf{P}_A$ *are chosen from the columns of* $\mathbf{A}$*, such that*

$$\|\mathbf{A} - \mathbf{P}_A\mathbf{Q}\|_F^2 \leq \left(\frac{k}{2} + 1 + \frac{k}{2(2^k-1)}\right) \cdot \mathrm{OPT}_k,$$

where $\mathrm{OPT}_k := \|\mathbf{A} - \mathbf{A}_k\|_F^2$*, and* $\mathbf{A}_k = \mathbf{UV}$ *is the optimal solution of (1). Here all matrix operations are over* GF(2)*.*

Moreover, we show that the approximation ratio $\left(\frac{k}{2} + 1 + \frac{k}{2(2^k-1)}\right)$ is asymptotically tight.

▶ **Theorem 3.** *In the* GF(2) *model, for every* $k \geq 1$ *and every* $\epsilon > 0$*, there exists* $\mathbf{A}$ *such that*

$$\|\mathbf{A} - \mathbf{P}_A\mathbf{Q}\|_F^2 > \left(\frac{k}{2} + 1 + \frac{k}{2(2^k-1)} - \epsilon\right) \cdot \mathrm{OPT}_k,$$

for all $\mathbf{P}_A, \mathbf{Q}$*, where* $\mathbf{P}_A$ *are formed by* k *columns of* $\mathbf{A}$*.*

Below, we give a high level description of the proof of the theorems. Our method uses the structure of GF(2) and is different to the techniques developed for CSS of real matrices.

Consider the problem given by Eq. (1). Throughout this paper, we will call the matrix $\mathbf{U}$ the *basis matrix*, since its column vectors are the basis of the low dimensional space. Likewise we call the right matrix $\mathbf{V}$ the *coefficient matrix*, since its columns contain the linear combination coefficients. Let $\mathbf{U}$ and $\mathbf{V}$ be an optimal solution of Eq. (1), and let $\mathbf{u_1}, \ldots, \mathbf{u_k}$

be the k columns of $\mathbf{U}$. For each column $\mathbf{u}_i$ of the optimal basis matrix $\mathbf{U}$, consider its nearest neighbor among all the columns of $\mathbf{A}$. Let $\mathbf{a}_1, \ldots, \mathbf{a}_n$ be the n columns of $\mathbf{A}$, and denote by $\mathbf{a}_{(\mathbf{u}_i)}$ the nearest neighbor column of $\mathbf{u}_i$ in $\mathbf{A}$. Given an optimal basis matrix $\mathbf{U}$, we thus have a matrix $\mathbf{A}_{(\mathbf{U})} := (\mathbf{a}_{(\mathbf{u}_1)}, \ldots, \mathbf{a}_{(\mathbf{u}_k)})$, consisting of columns of $\mathbf{A}$. Note that the optimal solution of Eq.(1) is not unique. In fact, fixing an optimal basis matrix $\mathbf{U}$, for every matrix $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_k)$, $\mathbf{b}_i \in \{0,1\}^k$, if the rank[3] of $\mathbf{B}$ is k over GF(2), then $(\mathbf{UB}, \mathbf{B}^{-1}\mathbf{V})$ must also be an optimal solution. Each optimal basis matrix $\mathbf{UB}$ induces a nearest neighbor matrix $\mathbf{A}_{(\mathbf{UB})}$. We will show that there must exist a rank k matrix $\mathbf{B}$ such that the induced nearest neighbor matrix $\mathbf{A}_{(\mathbf{UB})}$, which when used as basis matrix, achieves an approximation error at most $(\frac{k}{2} + 1 + \frac{k}{2(2^k-1)})$ times that of the optimal solution $(\mathbf{UB}, \mathbf{B}^{-1}\mathbf{V})$. Let $\mathrm{Err}(\mathbf{b}_1, \ldots, \mathbf{b}_k)$ be the approximation error associated with the basis matrix $\mathbf{A}_{(\mathbf{UB})}$ for $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_k)$. Our goal is to bound the quantity

$$\min_{\mathbf{b}_1,\ldots,\mathbf{b}_k} \mathrm{Err}(\mathbf{b}_1, \ldots, \mathbf{b}_k) \ , \tag{5}$$

where $\mathbf{b}_i \in \{0,1\}^k$ for all $i \in [k]$.

Directly bounding Eq.(5) is prohibitive. The approach we take is to consider a sequence of $k+1$ error minimization problems. For the r-th ($0 \le r \le k$) minimization, we only optimize r vectors among $\mathbf{b}_1, \ldots, \mathbf{b}_k$ and keep the other $k-r$ vectors fixed. Given $\mathbf{b}_1, \ldots, \mathbf{b}_k$, let

$$\mathrm{Err}^{(0)}(\mathbf{b}_1, \ldots, \mathbf{b}_k) := \mathrm{Err}(\mathbf{b}_1, \ldots, \mathbf{b}_k), \tag{6}$$

$$\mathrm{Err}^{(r)}(\mathbf{b}_1, \ldots, \mathbf{b}_{k-r}) := \min_{\mathbf{b}\in\{0,1\}^k} \mathrm{Err}^{(r-1)}(\mathbf{b}_1, \ldots, \mathbf{b}_{k-r}, \mathbf{b}). \tag{7}$$

Note that $\mathrm{Err}^{(k)}()$ is exactly the quantity of Eq.(5).

Although the final goal is to bound the *ratio* between $\mathrm{Err}^{(k)}()$ and the error of the optimal solution of Eq.(1), we instead prove *additive* bounds for $\mathrm{Err}^{(r)}(\mathbf{b}_1, \ldots, \mathbf{b}_{k-r})$ for all $0 \le r \le k$. To be more precise, letting OPT_k be the error of the optimal solution of Eq.(1), we will show that $\mathrm{Err}^{(r)}(\mathbf{b}_1, \ldots, \mathbf{b}_{k-r})$ is bounded by OPT_k plus a term depending on r and $\mathbf{b}_1, \ldots, \mathbf{b}_{k-r}$ (Theorem 5). Then when $r = k$, this additive bound becomes a multiplicative bound with respect to OPT_k and gives the desired ratio. The reason for introducing $\mathrm{Err}^{(0)}, \ldots, \mathrm{Err}^{(k-1)}$ is that we make use of the relation between $\mathrm{Err}^{(r)}$ and $\mathrm{Err}^{(r-1)}$ to prove the bound. More precisely, is the additive bound proved by induction in r.

Although the relation of $\mathrm{Err}^{(r)}$ and $\mathrm{Err}^{(r-1)}$ is

$$\mathrm{Err}^{(r)}(\mathbf{b}_1, \ldots, \mathbf{b}_{k-r}) = \min_{\mathbf{b}} \mathrm{Err}^{(r-1)}(\mathbf{b}_1, \ldots, \mathbf{b}_{k-r}, \mathbf{b}),$$

directly optimizing $\mathbf{b}$ seems very difficult. Our approach is to use *weighted averaging*. Since for each $\mathbf{b} \in \{0,1\}^k$ it holds that,

$$\mathrm{Err}^{(r)}(\mathbf{b}_1, \ldots, \mathbf{b}_{k-r}) \le \mathrm{Err}^{(r-1)}(\mathbf{b}_1, \ldots, \mathbf{b}_{k-r}, \mathbf{b}),$$

we have that for any set of weights $w_{\mathbf{b}}$ such that $w_{\mathbf{b}} \ge 0$ and $\sum_{\mathbf{b}} w_{\mathbf{b}} = 1$,

$$\mathrm{Err}^{(r)}(\mathbf{b}_1, \ldots, \mathbf{b}_{k-r}) \le \sum_{\mathbf{b}} w_{\mathbf{b}} \mathrm{Err}^{(r-1)}(\mathbf{b}_1, \ldots, \mathbf{b}_{k-r}, \mathbf{b}).$$

We carefully choose the weights $w_{\mathbf{b}}$ to get a small upper bound. We perform weighted averaging in two layers. Consider the quotient space $\mathrm{GF}(2)^k/\mathrm{span}(\mathbf{b}_1, \ldots, \mathbf{b}_{k-r})$ and the coset $[\mathbf{b}] := \mathbf{b} + \mathrm{span}(\mathbf{b}_1, \ldots, \mathbf{b}_{k-r})$. In the first layer, we perform weighted averaging within

[3] Throughout this section, matrix inverse and matrix rank are all over GF(2).

each coset $[\mathbf{b}]$, and obtain a bound for $\mathrm{Err}^{(r)}$ depending on the coset. In the second layer we average over all cosets using another set of weights. We need different rules to set the weights in the two layers. Within a coset $[\mathbf{b}]$, we choose the weights as follows. Let $\mathbf{U}, \mathbf{V}$ be the already fixed optimal solution of Eq.(1). For each $\mathbf{c} \in [\mathbf{b}]$, let $n_{\mathbf{c}}$ denote the number of columns of $\mathbf{V}$ that are equal to $\mathbf{c}$. The weight we assign to $\mathbf{c}$ is proportional to $n_{\mathbf{c}}$. For the second layer, let

$$n_{[\mathbf{b}]} := \sum_{\mathbf{c} \in [\mathbf{b}]} n_{\mathbf{c}}$$

be the total number of columns of $\mathbf{V}$ that belong to the coset $[\mathbf{b}]$. We assign the weight to a coset $[\mathbf{b}]$ as follows. If

$$[\mathbf{b}] = \mathrm{span}(\mathbf{b}_1, \ldots, \mathbf{b}_{k-r}),$$

then the weight is set to be zero. Otherwise, we assign the weight to $[\mathbf{b}]$ proportional to

$$\frac{n_{[\mathbf{b}]}}{\sum_{[\mathbf{b}]} n_{[\mathbf{b}]} - \lambda n_{[\mathbf{b}]}},$$

where λ is a constant depending on r. Combining the two layers of averaging we obtain the additive bound and that implies the desired approximation ratio. This finishes the description of the proof of Theorem 2.

The lower bound in Theorem 3 is proved by explicit construction. We construct a matrix which is approximately low rank in the sense that it is the product of two rank-k matrix plus a very sparse matrix. The key ingredient of the proof is the construction of the two rank-k matrices, which have special structures so that the approximation ratio of column subset selection cannot be smaller than $\frac{k}{2} + 1 + \frac{k}{2(2^k-1)}$ significantly.

The additive bounds are stated in Theorem 5, which is technical. Below we first describe the notions that will appear in Theorem 5. These notions will also be frequently used in the proof as well. For clarity, we list the notions in two tables.

▶ **Definition 4.** For $1 \le r \le k$ and linear independent vectors $\mathbf{b}_1, \ldots, \mathbf{b}_r$ in $\{0,1\}^k$:

Table 1 Definitions for vector spans.

Definition	Explanation
$\mathrm{span}^c(\mathbf{b}_1, \ldots, \mathbf{b}_r) := \{0,1\}^k \setminus \mathrm{span}(\mathbf{b}_1, \ldots, \mathbf{b}_r)$	Complement of $\mathrm{span}(\mathbf{b}_1, \ldots, \mathbf{b}_r)$.
$\mathrm{span}^{\setminus i}(\mathbf{b}_1, \ldots, \mathbf{b}_r) := \mathrm{span}(\mathbf{b}_1, \ldots, \mathbf{b}_{i-1}, \mathbf{b}_{i+1}, \ldots, \mathbf{b}_r)$	Span of all vectors except the ith.

Let $\mathbf{A}$ be the matrix to be approximated and $(\mathbf{U}, \mathbf{V})$ be a fixed optimal solution of the problem in Eq.(1). For $\mathbf{u} \in \{0,1\}^d$, $\mathbf{c} \in \{0,1\}^k$, and $\mathcal{X} \subset \{0,1\}^k$:

Table 2 Definitions for errors and nearest neighbors.

Definition	Explanation
$\mathbf{a}_{(\mathbf{u})}$	The nearest neighbor of $\mathbf{u}$ among the columns of $\mathbf{A}$ (If more than one nearest neighbor, choose one arbitrarily.)
$\mathcal{J}_{\mathbf{c}} := \{j \in [n] : \mathbf{V}_j = \mathbf{c}\}$	The set of columns of $\mathbf{V}$ that are equal to vector $\mathbf{c}$.
$n_{\mathbf{c}} := \lvert\mathcal{J}_{\mathbf{c}}\rvert$	The number of columns of $\mathbf{V}$ that are equal to $\mathbf{c}$.
$L_{\mathbf{c}} := \sum_{j \in \mathcal{J}_{\mathbf{c}}} \lvert\mathbf{a}_j - \mathbf{U}\mathbf{c}\rvert$	The total approximation error of those columns in $\mathcal{J}_{\mathbf{c}}$.
$N_{\mathcal{X}} := \sum_{\mathbf{c} \in \mathcal{X}} n_{\mathbf{c}}$	The total number of columns of $\mathbf{V}$ that belong to set $\mathcal{X}$.
$M_{\mathbf{c}} = \begin{cases} \frac{L_{\mathbf{c}}}{n_{\mathbf{c}}} & n_{\mathbf{c}} > 0 \\ d & n_{\mathbf{c}} = 0 \end{cases}$	Upper bound of the average error of the columns in $\mathcal{J}_{\mathbf{c}}$.

Now we can state the additive bounds.

▶ **Theorem 5.** *Let* $\mathbf{b}_1, \ldots, \mathbf{b}_k$ *be* k *linear independent vectors in* $\{0,1\}^k$*. Then for each* $0 \leq r \leq k$*,*

$$\mathrm{Err}^r(\mathbf{b}_1, \ldots, \mathbf{b}_{k-r}) \leq \mathrm{OPT}_k + \lambda_r \cdot \sum_{\mathbf{c} \in \mathrm{span}^c(\mathbf{b}_1, \ldots, \mathbf{b}_{k-r})} L_{\mathbf{c}} + \sum_{i=1}^{k-r} f_i(\mathbf{b}_1, \ldots, \mathbf{b}_{k-r}) M_{\mathbf{b}_i}, \quad (8)$$

where $M_{\mathbf{b}_i}$ *has been defined in Definition 4, and*

$$\lambda_r = \begin{cases} 0 & r = 0 \\ \frac{r}{2}\left(1 + \frac{1}{2^r - 1}\right), & 1 \leq r \leq k \end{cases}$$

and

$$f_i(\mathbf{b}_1, \ldots, \mathbf{b}_{k-r}) = N_{\mathcal{X}} + \frac{1}{2} N_{\mathcal{Y}}, \quad (9)$$

here $\mathcal{X} = \mathbf{b}_i + \mathrm{span}^{\setminus i}(\mathbf{b}_1, \ldots, \mathbf{b}_{k-r})$*, and* $\mathcal{Y} = \mathrm{span}^c(\mathbf{b}_1, \ldots, \mathbf{b}_{k-r})$*.*

The formal proof of Theorem 5 is lengthy and can be found in the full version of the paper [12]. Theorem 2 follows from Theorem 5 immediately.

Proof of Theorem 2. Let $r = k$ in Theorem 5. Then the last term in the RHS of Eq.(8) vanishes. The second term in the RHS of Eq.(8) becomes $\lambda_k \cdot \sum_{\mathbf{c} \in \{0,1\}^k} L_{\mathbf{c}}$. Observe that

$$\sum_{\mathbf{c} \in \{0,1\}^k} L_{\mathbf{c}} = \mathrm{OPT}_k,$$

and

$$1 + \lambda_k = \frac{k}{2} + 1 + \frac{k}{2(2^k - 1)},$$

the theorem follows. ◀

3 Generalized CSS Over Boolean Semiring

It is not difficult to see that the method developed for GF(2) model in the previous section does not apply to the Boolean model, simply because the Boolean semiring does not have a field structure. It turns out that, somewhat surprisingly, CSS is not sufficient to yield a bound relative to the optimal low rank solution in the Boolean model.

Here, we propose a Generalized CSS (GCSS) procedure. In GCSS, instead of using the columns of $\mathbf{A}$ directly to form $\mathbf{P}_A$, we apply carefully designed Boolean formulas (bitwise) to a predefined number of columns of $\mathbf{A}$ to form $\mathbf{P}_A$.

To illustrate the ideas, we first give an informal high level description of GCSS. We can capture our GCSS by the following framework, which we denote as an *oblivious basis generation scheme with advice*. Let $f(k)$ and $g(k)$ be functions of k. An oblivious basis generation scheme with *advice size* $f(k)$ and *column dependence size* $g(k)$ operates as follows. Given as input an advice string $\mathbf{o} \in \{0,1\}^{f(k)}$ the scheme outputs k Boolean formulas $\Phi_1, \ldots, \Phi_k$ each of $g(k)$ bits. Given $g(k)$ columns $\mathbf{a}_{i_1}, \ldots, \mathbf{a}_{i_{g(k)}}$ of the matrix $\mathbf{A}$, the k basis vectors $\mathbf{u}_1, \ldots, \mathbf{u}_k$ of $\mathbf{P}_A$ are constructed as

$$\mathbf{u}_j = \Phi_j(\mathbf{a}_{i_1}, \ldots, \mathbf{a}_{i_{g(k)}}),$$

where the Boolean function Φ_j is applied *entry-wise*. From such a basis generation scheme we immediately obtain an approximation result by iterating over all possible selections of $g(k)$ columns of $\mathbf{A}$ as well as all possible advice strings $\mathbf{o} \in \{0,1\}^{f(k)}$. We stress that the amount of information about $\mathbf{A}$ that can be supplied to the algorithm using the advice string is *independent* of the actual size of $\mathbf{A}$. Our construction of GCSS will have column dependence size $2^k - 1$ and advice size $O(k2^k)$ in which we encode an *ordering* of the given $2^k - 1$ columns. This results in an approximation ratio of $2^{k-1} + 1$.

To give a precise description of GCSS, it is more convenient to use sets instead of vectors as the representation. For a column $\mathbf{a}_i$ of $\mathbf{A}$, let

$$\mathcal{A}_i := \{j \in [d] : (\mathbf{a}_i)_j = 1\},$$

i.e., $\mathbf{a}_i$ is the characteristic vector of $\mathcal{A}_i$. Similarly, for an optimal solution $(\mathbf{U}, \mathbf{V})$ of the Boolean low rank approximation problem, let

$$\mathcal{U}_i := \{j \in [d] : (\mathbf{u}_i)_j = 1\},$$

and

$$\mathcal{V}_i := \{j \in [k] : v_{ij} = 1\}.$$

Thus in this section we will always think of a column of $\mathbf{A}$, $\mathbf{U}$ or $\mathbf{V}$ as a set. Given a set $\mathcal{S} \subset [k]$, let

$$\mathcal{J}_\mathcal{S} := \{j \in [n] : \mathcal{V}_j = \mathcal{S}\},$$

and $n_\mathcal{S} := |\mathcal{J}_\mathcal{S}|$. Using these notions, the Boolean product of $\mathbf{U}$ and a vector which is the characteristic vector of $\mathcal{S}$ will be denoted by $\mathcal{U}_\mathcal{S} := \bigcup_{i\in\mathcal{S}} \mathcal{U}_i$. Abusing the notion slightly, we shall still use $\mathcal{U}_i$ instead of $\mathcal{U}_{\{i\}}$ from now on. Like in the previous section, the nearest neighbor column of $\mathcal{U}_\mathcal{S}$ in $\mathbf{A}$ is defined by $\mathbf{a}_{(\mathcal{U}_\mathcal{S})}$. As we use set representation in this section, for notational simplicity we let $\mathcal{D}_\mathcal{S} \subset [d]$ be the set corresponding to this nearest neighbor column $\mathbf{a}_{(\mathcal{U}_\mathcal{S})}$, i.e.,

$$\mathcal{D}_\mathcal{S} := \{i \in [d] : \mathbf{a}_{(\mathcal{U}_\mathcal{S})_i} = 1\}$$

We are going to construct a rank-k solution $\mathcal{B}_1, \ldots, \mathcal{B}_k$, where $\mathcal{B}_i \subset [d]$ is the set representation of the column of the basis matrix. Once the basis matrix is obtained, the coefficient matrix can be calculated in the same way as in the previous section. The concrete GCSS procedure is described in Algorithm 1.

Now we can state the main result of this section.

▶ **Theorem 6.** *GCSS (as described above) has approximation ratio 2^k relative to the optimal solution of (1) over Boolean semiring.*

We now give the very high level idea of the proof. Fix a bijection π that satisfies $n_{S_1} \leq \cdots \leq n_{S_{2^k-1}}$. By construction the set $\mathcal{D}_{S_\ell}$ is the best approximation to $\mathcal{U}_{\mathcal{S}_\ell}$ given by a column of $\mathbf{A}$. Ideally the sets $\mathcal{B}_1, \ldots, \mathcal{B}_k$ should be such that $\bigcup_{i\in S_\ell} \mathcal{B}_i$ is a comparable substitute for all ℓ. What we instead will be able to achieve is that for all $\ell \in [2^k - 1]$

$$\mathcal{U}_{\mathcal{S}_\ell} \triangle \left(\bigcup_{i\in S_\ell} \mathcal{B}_i \right) \subseteq \left(\bigcup_{\ell' \geq \ell} (\mathcal{U}_{\mathcal{S}_{\ell'}} \triangle \mathcal{D}_{\mathcal{S}_{\ell'}}) \right) \tag{10}$$

where as seen from the algorithm the sets $\mathcal{B}_i$ are Boolean combinations of the sets $\mathcal{D}_{\mathcal{S}_\ell}$. Intuitively, we give more importance to approximating the columns of $\mathbf{A}$ from $\mathcal{J}_{\mathcal{S}_\ell}$ as ℓ increases. As the sizes $n_{\mathcal{S}_\ell}$ of these sets of columns also increase this means that we can account for the extra cost of possible poor approximation of the sets $\mathcal{U}_{\mathcal{S}_\ell}$ for smaller ℓ in terms of the approximation error of the sets $\mathcal{D}_{\mathcal{S}_{\ell'}}$ to $\mathcal{U}_{\mathcal{S}_{\ell'}}$ for larger $\ell' \geq \ell$.

Intuitively we should attempt to approximate all the sets $\mathcal{D}_{\mathcal{S}_\ell}$ simultaneously by $\bigcup_{i\in\mathcal{S}_\ell} \mathcal{B}_i$. But since we work over a semiring we will have to work with under-approximations. So for every ℓ we instead approximate the under-approximation $\bigcup_{i\in\mathcal{S}_\ell} \mathcal{E}_i^\ell$ of $\mathcal{D}_{\mathcal{S}_\ell}$. We do this by initially letting $\mathcal{B}_i = \mathcal{E}_i^1$ and then for each $\ell \in [2^k - 1]$ adding $\left(\bigcup_{i\in\mathcal{S}_\ell} \mathcal{E}_i^\ell\right) \setminus \left(\bigcup_{i\in\mathcal{S}_\ell} \mathcal{E}_i^1\right)$ to $\bigcup_{i\in S_\ell} \mathcal{B}_i$. This last step has to be done carefully piece by piece using the ordering of the sets $\mathcal{S}_1, \ldots, \mathcal{S}_{2^k-1}$. In the algorithm this is done using the sets $\mathcal{F}_i^{\ell_1,\ell_2}$.

The approximation ratio of GCSS over the Boolean semiring is $O(2^k)$, and thus much larger than that of GF(2). However we shall argue that this exponential dependency on k is not an artifact of proof technique, it seems inherent to the model.

Let k be even and let $n = 2^{k/2}$. We define the $n \times n$ matrix $\mathbf{A} = (a_{\alpha,\beta})$ indexed by strings $\alpha, \beta \in \{0,1\}^{k/2}$ by $a_{\alpha,\beta} = 1$ if and only if $\alpha \neq \beta$. Thus $\mathbf{A}$ is just the negation of the $n \times n$ identity matrix. It is well-known that the Boolean rank of $\mathbf{A}$ is equal to k. In particular, we can write $\mathbf{A}$ as the Boolean product of $\mathbf{U}$ and $\mathbf{V}$, where the columns of $\mathbf{U}$ and the rows of $\mathbf{V}$ are indexed by pairs (i, b) where $i \in [k/2]$ and $b \in \{0, 1\}$ and entry $(\alpha, (i, b))$ of $\mathbf{U}$ is 1 if and only if $\alpha_i = b$ and entry $((i, b), \beta)$ of $\mathbf{V}$ is 1 if and only if $\beta_i \neq b$. We note that the columns of $\mathbf{U}$ can be written as Boolean formulas applied entry-wise to (all of) the columns of $\mathbf{A}$. Since we consider approximation algorithms with *multiplicative error*, when supplied with input A and k our algorithm is required to compute an *exact factorization* of $\mathbf{A}$ into $n \times k$ and $k \times n$ matrices $\mathbf{U}$ and $\mathbf{V}$. If the underlying basis generation algorithm receives, say, only half of the columns of $\mathbf{A}$ it does not seem possible to compute such a factorization. It therefore seems that column dependence size at least $2^{k/2-1}$ is necessary, which is about the square-root of the column dependence size of our algorithm.

▶ Remark. Using the technique of weighted averaging developed for the GF(2) model, we can actually improve the approximation ratio to $2^{k-1} + 1$. We omit the details of the proof.

The proof of Theorem 6 can be found in the full version of the paper [12].

4 Hardness of Low Rank Approximation of Binary Matrices

Prior to our work, the computational complexity of the low rank approximation problem was not fully understood. For the rank-1 case, Tan showed that the equivalent problem Maximum Edge Weight Biclique for $\{-1, 1\}$-matrices is NP-hard under *randomized* reductions [42]. In the case when the rank k is unrestricted (i.e. part of the input) deciding whether there exist $\mathbf{U}$ and $\mathbf{V}$ such that $\mathbf{A} = \mathbf{UV}$ in the Boolean semiring model is precisely the NP-complete Minimal Set Basis problem [41], and that immediately implies that the approximation problem is NP-hard to approximate within *any* factor, as noted by Miettinen et al. [34]. On the other hand, this does not imply hardness when $k \ll d, n$. Indeed, the Minimal Set Basis problem is fixed-parameter tractable with parameter k, by a simple kernelization algorithm [17]. Note also that in the GF(2) model, deciding the existence of $\mathbf{U}$ and $\mathbf{V}$ such that $\mathbf{A} = \mathbf{UV}$ is efficiently solvable using Gaussian elimination, regardless of the rank k being unrestricted.

In this section we show the rank-1 Binary Matrix Approximation problem is NP-hard under normal polynomial time reduction. We first define two related problems. Let H be a *complete* bipartite graph with edge weight, and let $\mathbf{W} = (w_{ij})$ be the $d \times n$ matrix consisting

Algorithm 1 Generalized Column Subset Selection.

for all selection of $2^k - 1$ column vectors $\mathcal{A}_{j_1}, \mathcal{A}_{j_2}, \ldots, \mathcal{A}_{j_{2^k-1}}$ in $\mathbf{A}$ **do**
  for all bijections $\pi : [2^k - 1] \to (2^{[k]} \setminus \{\emptyset\})$ **do**
    Let $S_\ell = \pi(\ell)$ for $\ell \in [2^k - 1]$
    for $i \in [k]$ and $\ell \in [2^k - 1]$ **do**
      Compute

$$\mathcal{E}_i^\ell := \bigcap_{\substack{\ell' \geq \ell: \\ i \in \mathcal{S}_{\ell'}}} D_{S_l}$$

      where $D_{\mathcal{S}} = \mathcal{A}_{j_{\pi^{-1}(\mathcal{S})}}$ for $\emptyset \neq \mathcal{S} \subseteq [k]$.
    end for
    for $1 \leq \ell_1 < \ell_2 \leq 2^k - 1$ such that $i \in \mathcal{S}_{\ell_1} \cap \mathcal{S}_{l_2}$ **do**
      Compute

$$\mathcal{F}_i^{\ell_1,\ell_2} := \mathcal{E}_i^{\ell_1+1} \setminus \left[\bigcup_{i' \in \mathcal{S}_{\ell_2}} \mathcal{E}_{i'}^{\ell_1} \right].$$

    end for
    for $i \in [k]$ **do**
      Compute solution vector $\{\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_k\}$by

$$\mathcal{B}_i := \mathcal{E}_i^1 \cup \left(\bigcup_{\substack{\ell_1 < \ell_2: \\ i \in \mathcal{S}_{\ell_1} \cap \mathcal{S}_{\ell_2}}} \mathcal{F}_i^{\ell_1,\ell_2} \right).$$

    end for
  end for
  Compute the approximation error using the solution vector.
  if the approximation error is optimal **then**
    Save $\{\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_k\}$ as the output.
  end if
end for

of these edge weights. The Maximum Edge Weight Biclique problem is to find a biclique subgraph of H with maximizing total edge weight. As an optimization problem: maximize $\mathbf{x}^\mathsf{T}\mathbf{W}\mathbf{y}$, where $\mathbf{x} \in \{0,1\}^d$ and $\mathbf{y} \in \{0,1\}^n$. The Bipartite Max-Cut problem is to find a cut of the vertices of H maximum the total weight of the edges cut. As an optimization problem: maximize $\mathbf{x}^\mathsf{T}\mathbf{W}\mathbf{y}$, where $\mathbf{x} \in \{-1,1\}^d$ and $\mathbf{y} \in \{-1,1\}^n$. Note that these two problems differ only in the domain from which $\mathbf{x}$ and $\mathbf{y}$ are chosen.

Shen, Ji, and Ye [39] observed that the rank-1 Binary Matrix Approximation problem is equivalent to Maximum Edge Weight Biclique when all edge weights are chosen from $\{-1,1\}$. Namely, if $\mathbf{A}$ is a $d \times n$ Boolean matrix, $\mathbf{u} \in \{0,1\}^d$, and $\mathbf{v} \in \{0,1\}^n$, and let $\mathbf{J}_{d,n}$

denote the $d \times n$ all-1 matrix, we have

$$\begin{aligned}\|\mathbf{A} - \mathbf{u}\mathbf{v}^\mathsf{T}\|_F^2 &= \|\mathbf{A}\|_F^2 - 2\mathbf{u}^\mathsf{T}\mathbf{A}\mathbf{v} + \|\mathbf{u}\mathbf{v}^\mathsf{T}\|_F^2 \\ &= \|\mathbf{A}\|_F^2 - \mathbf{u}^\mathsf{T}(2\mathbf{A} - \mathbf{J}_{d,n})\mathbf{v}.\end{aligned}$$

Therefore, minimizing $\|\mathbf{A}-\mathbf{u}\mathbf{v}^\mathsf{T}\|_F^2$ is equivalent of maximizing $\mathbf{u}^\mathsf{T}(2\mathbf{A}-\mathbf{J}_{d,n})\mathbf{v}$. Also note that $(2\mathbf{A} - \mathbf{J}_{d,n})$ is a $\{-1,1\}$-matrix. Thus NP-hardness of Maximum Edge Weight Biclique with $\{-1,1\}$ edge weights implies NP-hardness of rank-1 Binary Matrix Approximation. To show the NP-hardness of Maximum Edge Weight Biclique, we consider reduction from the Bipartite Max-Cut problem.

Roth and Viswanathan showed that Bipartite Max-Cut is NP-hard even when all weights are chosen from the set $\{-1,1\}$ [37]. This is done by first showing NP-hardness when the weights are chosen from $\{-1,0,1\}$ and then reducing to the case of weights from $\{-1,1\}$.

Tan showed that Maximum Edge Weight Biclique is NP-hard [42] when weights are chosen from $\{-1,0,1\}$, and shows NP-hardness under randomized reductions when weights are chosen from $\{-1,1\}$. He leaves it as an open problem to obtain NP-hardness under normal polynomial time reductions. The complexity of this problem was also stated as an open problem by Amit [2]

The reduction from weights chosen from $\{-1,0,1\}$ to $\{-1,1\}$ by Roth and Viswanathan and by Tan is similar. The idea is to transform the $n \times n$ $\{-1,0,1\}$-weight matrix $\mathbf{W}$ into a new $nm \times nm$ $\{-1,1\}$-weight matrix $\mathbf{W}'$, where $\mathbf{W}'$ consists of $m \times m$ blocks corresponding to each entry of $\mathbf{W}$. A (-1)-entry is transformed into the all-(-1) $m \times m$ matrix, and similarly is a 1-entry transformed into the all 1 $m \times m$ matrix. But where Tan transforms a 0-entry to a *random* $m \times m$ $\{-1,1\}$-matrix, Roth and Viswanathan instead transforms a 0-entry into a $m \times m$ *Hadamard* matrix. We will show that this transformation into *Hadamard* matrix also work in the setting of the Maximum Edge Weight Biclique problem, thereby properly establishing its NP-hardness.

▶ **Theorem 7.** *The rank-*1 Binary Matrix Approximation *problem is* NP*-hard.*

We give a polynomial time many-one reduction from Maximum Edge Weight Biclique with weights from $\{-1,0,1\}$ to Maximum Edge Weight Biclique with weights from $\{0,1\}$, thereby showing the theorem. The proof is based on the following three lemmas.

The lemma below is an adaptation of [37, Lemma 4.2] from the $\{-1,1\}$ case to the $\{0,1\}$ case.

▶ **Lemma 8.** *Let* $\mathbf{W}$ *be an* $n \times n$ *matrix and let* $m \geq 1$*, and define* $\mathbf{W}' = \mathbf{W} \otimes \mathbf{J}_m$*, where* $\mathbf{J}_m := \mathbf{J}_{m,m}$*. Then*

$$\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^\mathsf{T}\mathbf{W}'\mathbf{v} = m^2 \cdot \max_{\mathbf{x},\mathbf{y}} \mathbf{x}^\mathsf{T}\mathbf{W}\mathbf{y} \ ,$$

where $\mathbf{u},\mathbf{v} \in \{0,1\}^{mn}$ *and* $\mathbf{x},\mathbf{y} \in \{0,1\}^n$*, respectively. Furthermore, if* $\mathbf{x}$ *and* $\mathbf{y}$ *maximize* $\mathbf{x}^\mathsf{T}\mathbf{W}\mathbf{y}$*, then* $\mathbf{u} = \mathbf{x} \otimes \mathbf{1}_m$ *and* $\mathbf{v} = \mathbf{y} \otimes \mathbf{1}_m$ *maximize* $\mathbf{u}^\mathsf{T}\mathbf{W}'\mathbf{v}$*.*

Proof. Consider first $\mathbf{u} = \mathbf{x} \otimes \mathbf{1}_d$ and $\mathbf{v} = \mathbf{y} \otimes \mathbf{1}_m$. Then

$$\begin{aligned}\mathbf{u}^\mathsf{T}(\mathbf{W} \otimes \mathbf{J}_m)\mathbf{v} &= (\mathbf{x} \otimes \mathbf{1}_m)^\mathsf{T}(\mathbf{W} \otimes \mathbf{J}_m)(\mathbf{y} \otimes \mathbf{1}_m) \\ &= (\mathbf{x}^\mathsf{T}\mathbf{W}\mathbf{y}) \otimes (\mathbf{1}_m^\mathsf{T}\mathbf{J}_m\mathbf{1}_m) = m^2 \cdot (\mathbf{x}^\mathsf{T}\mathbf{W}\mathbf{y}) \ .\end{aligned}$$

Next, take $\mathbf{u}$ and $\mathbf{v}$ maximizing $\mathbf{u}^\mathsf{T}\mathbf{W}'\mathbf{v}$. We show that $\mathbf{u}$ and $\mathbf{v}$ can be brought to the form $\mathbf{u} = \mathbf{x} \otimes \mathbf{1}_m$ and $\mathbf{v} = \mathbf{y} \otimes \mathbf{1}_m$ without decreasing the value of $\mathbf{u}^\mathsf{T}\mathbf{W}'\mathbf{v}$. We first fix $\mathbf{v}$ and bring $\mathbf{u}$ to the desired form, and then similarly bring $\mathbf{v}$ to the desired form.

So fix $\mathbf{v}$, and let $\mathbf{z} = \mathbf{W}'\mathbf{v}$. Note that $\mathbf{u}$ maximizing $\mathbf{u}^\mathsf{T}\mathbf{z}$ must satisfy $\mathbf{u}_i = 1$ when $\mathbf{z}_i > 0$ and $\mathbf{u}_i = 0$ when $\mathbf{z}_i < 0$. Since $\mathbf{W}' = \mathbf{W} \otimes \mathbf{J}_m$ we have that $\mathbf{z}_{jm+1} = \mathbf{z}_{jm+2} = \cdots = \mathbf{z}_{(j+1)m}$ for all $j = 0, 1, \ldots, n-1$. Hence we can choose a maximizing $\mathbf{u}$ satisfying $\mathbf{u}_{jm+1} = \mathbf{u}_{jm+2} = \cdots = \mathbf{u}_{(j+1)m}$ for all $j = 0, 1, \ldots, n-1$ as well, meaning $\mathbf{u} = \mathbf{x} \otimes \mathbf{1}_m$ for suitable $\mathbf{x} \in \{0,1\}^n$. We can now fix $\mathbf{u}$ and in a similar way bring $\mathbf{v}$ to the form $\mathbf{v} = \mathbf{y} \otimes \mathbf{1}_m$ for suitable $\mathbf{y} \in \{0,1\}^n$. ◀

The following lemma, which is the $\{0,1\}$ analogue of [37, Lemma 4.3], is a direct consequence of Lindsey's Lemma. We state the proof for completeness.

▶ **Lemma 9.** *Let $\mathbf{H}$ be a $m \times m$ Hadamard matrix. For every $\mathbf{x}, \mathbf{y} \in \{0,1\}^m$, $|\mathbf{x}^\mathsf{T}\mathbf{H}\mathbf{y}| \leq m^{3/2}$.*

Proof. First note

$$\|\mathbf{H}\mathbf{y}\|^2 = \mathbf{y}^\mathsf{T}(\mathbf{H}^\mathsf{T}\mathbf{H})\mathbf{y} = \mathbf{y}^\mathsf{T}(m\mathbf{I})\mathbf{y} = m \cdot \|\mathbf{y}\|^2 \ .$$

We can then complete the proof by the Cauchy-Schwartz inequality,

$$|\mathbf{x}^\mathsf{T}\mathbf{H}\mathbf{y}| \leq \|\mathbf{x}^\mathsf{T}\| \cdot \|\mathbf{H}\mathbf{y}\| = \sqrt{m} \cdot \|\mathbf{x}\| \cdot \|\mathbf{y}\| \leq m^{3/2} \ . \qquad ◀$$

▶ **Lemma 10.** *Let $\mathbf{W} = (w_{ij})$ be a $n \times n$ $\{-1,0,1\}$-matrix and let $\mathbf{H}$ be a $m \times m$ Hadamard matrix. Define the $(mn) \times (mn)$ $\{-1,1\}$-block matrix $\widetilde{\mathbf{W}} = (\widetilde{\mathbf{W}}_{ij})$, where block $\widetilde{\mathbf{W}}_{ij}$ is given by*

$$\widetilde{\mathbf{W}}_{ij} = \begin{cases} w_{ij}\mathbf{J}_m & \text{if } w_{ij} \neq 0 \\ \mathbf{H} & \text{if } w_{ij} = 0 \end{cases} \ .$$

Let $\mathbf{W}' = \mathbf{W} \otimes \mathbf{J}_m$. Then for all $\mathbf{u}, \mathbf{v} \in \{0,1\}^{mn}$, $\left|\mathbf{u}^\mathsf{T}\widetilde{\mathbf{W}}\mathbf{v} - \mathbf{u}^\mathsf{T}\mathbf{W}'\mathbf{v}\right| \leq n^2 \cdot m^{3/2}$.

Proof. This is by simple estimation.

$$\begin{aligned} \left|\mathbf{u}^\mathsf{T}\widetilde{\mathbf{W}}\mathbf{v} - \mathbf{u}^\mathsf{T}\mathbf{W}'\mathbf{v}\right| &= \left|\mathbf{u}^\mathsf{T}(\widetilde{\mathbf{W}} - \mathbf{W}')\mathbf{v}\right| \\ &\leq n^2 \cdot \max_{\mathbf{x},\mathbf{y} \in \{0,1\}^m} \left|\mathbf{x}^\mathsf{T}\mathbf{H}\mathbf{y}\right| \\ &\leq n^2 \cdot m^{3/2} \ , \end{aligned}$$

where the last inequality follows from Lemma 9. ◀

Proof. of Theorem 7 Suppose now that $\mathbf{W}$ is an $n \times n$ $\{-1,0,1\}$-matrix. Let $m = 2^\ell$ be the smallest power of 2 that is greater than $4n^4$, and let $\mathbf{H}$ be the $m \times m$ Sylvester Hadamard matrix. We then define $\widetilde{\mathbf{W}}$ and $\mathbf{W}'$ as in Lemma 10. Then

$$\begin{aligned} &\left| \max_{\mathbf{u},\mathbf{v} \in \{0,1\}^{mn}} \mathbf{u}^\mathsf{T}\widetilde{\mathbf{W}}\mathbf{v} - m^2 \cdot \max_{\mathbf{x},\mathbf{y} \in \{0,1\}^n} \mathbf{x}^\mathsf{T}\mathbf{W}\mathbf{y} \right| \\ =& \left| \max_{\mathbf{u},\mathbf{v} \in \{0,1\}^{mn}} \mathbf{u}^\mathsf{T}\widetilde{\mathbf{W}}\mathbf{v} - \max_{\mathbf{u},\mathbf{v} \in \{0,1\}^{mn}} \mathbf{u}^\mathsf{T}\mathbf{W}'\mathbf{v} \right| \\ \leq& n^2 \cdot m^{3/2} \leq \frac{m^{1/2}}{2} \cdot m^{3/2} = \frac{m^2}{2} \ , \end{aligned}$$

where the first equality is by Lemma 8 and the first inequality is by Lemma 10.
Since the expression $m^2 \cdot \max_{\mathbf{x},\mathbf{y} \in \{0,1\}^n} \mathbf{x}^\mathsf{T}\mathbf{W}\mathbf{y}$ is an integer multiple of m^2, the value $\max_{\mathbf{u},\mathbf{v} \in \{0,1\}^{mn}} \mathbf{u}^\mathsf{T}\widetilde{\mathbf{W}}\mathbf{v}$ uniquely determines the value $\max_{\mathbf{x},\mathbf{y} \in \{0,1\}^n} \mathbf{x}^\mathsf{T}\mathbf{W}\mathbf{y}$. This then gives the desired reduction. ◀

5 Conclusion

We have studied Column Subset Section (CSS) for low rank binary matrix approximation. CSS is often used as an alternative approach of SVD for low rank approximation of real matrices, where the advantage of CSS is the interpretability of its results. For binary matrices, CSS is so far the only approach yielding theoretical guarantees, as solving the low rank problem exactly is NP-hard. We provide an upper bound on the approximation ratio of CSS for the GF(2) model and show the bound is tight. This is a complete characterization from an information-theoretic point of view. For the Boolean semiring model, we propose a Generalized CSS (GCSS) method, since CSS is not strong enough to yield a bound in this scenario. We also show an upper bound for GCSS.

CSS has been actively studied for nearly three decades and the first work can at least date back to [23], where it was called rank revealing QR in the numerical linear algebra community. The progress on CSS exhibits an interesting trajectory. Early results either gave bounds exponential in k or the running time of the algorithm is $O(n^k)$ [4, 7–9, 18, 27]. After efforts of many researches, there are now polynomial time algorithms that have polynomial bounds for the approximation ratio.

Our understanding of CSS for binary matrices is at the very beginning stage. It is an important problem for future work to develop efficient CSS algorithms that achieves or approximately achieves the bounds of this paper.

References

1 Jason Altschuler, Aditya Bhaskara, Gang Fu, Vahab Mirrokni, Afshin Rostamizadeh, and Morteza Zadimoghaddam. Greedy column subset selection: New bounds and distributed algorithms. *Proceedings of the 33rd International Conference on MachineLearning*, 2016.

2 Noga Amit. The bicluster graph editing problem. M.sc. thesis, Tel Aviv University, 2004.

3 Radim Belohlavek and Vilem Vychodil. Discovery of optimal factors in binary data via a novel method of matrix decomposition. *Journal of Computer and System Sciences*, 76(1):3–20, 2010.

4 Christian H Bischof and Gregorio Quintana-Ortí. Computing rank-revealing qr factorizations of dense matrices. *ACM Transactions on Mathematical Software (TOMS)*, 24(2):226–253, 1998.

5 Christos Boutsidis, Petros Drineas, and Malik Magdon-Ismail. Near-optimal column-based matrix reconstruction. *SIAM Journal on Computing*, 43(2):687–717, 2014.

6 Christos Boutsidis, Michael W Mahoney, and Petros Drineas. An improved approximation algorithm for the column subset selection problem. In *Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 968–977. Society for Industrial and Applied Mathematics, 2009.

7 Tony F Chan. Rank revealing qr factorizations. *Linear algebra and its applications*, 88:67–82, 1987.

8 Tony F Chan and Per Christian Hansen. Low-rank revealing qr factorizations. *Numerical Linear Algebra with Applications*, 1(1):33–44, 1994.

9 Shivkumar Chandrasekaran and Ilse CF Ipsen. On rank-revealing factorisations. *SIAM Journal on Matrix Analysis and Applications*, 15(2):592–622, 1994.

10 Ali Civril and Malik Magdon-Ismail. Column subset selection via sparse approximation of svd. *Theoretical Computer Science*, 421:1–14, 2012.

11 Michael B Cohen, Sam Elder, Cameron Musco, Christopher Musco, and Madalina Persu. Dimensionality reduction for k-means clustering and low rank approximation. In *Proceed-*

ings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, pages 163–172. ACM, 2015.

12 Chen Dan, Kristoffer Arnsfelt Hansen, He Jiang, Liwei Wang, and Yuchen Zhou. On low rank approximation of binary matrices. *CoRR*, abs/1511.01699, 2015.

13 Amit Deshpande and Luis Rademacher. Efficient volume sampling for row/column subset selection. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 329–338. IEEE, 2010.

14 Amit Deshpande, Luis Rademacher, Santosh Vempala, and Grant Wang. Matrix approximation and projective clustering via volume sampling. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pages 1117–1126. Society for Industrial and Applied Mathematics, 2006.

15 Amit Deshpande and Santosh Vempala. Adaptive sampling and fast low-rank matrix approximation. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 292–303. Springer, 2006.

16 Petros Drineas, Michael W Mahoney, and S Muthukrishnan. Relative-error cur matrix decompositions. *SIAM Journal on Matrix Analysis and Applications*, 30(2):844–881, 2008.

17 Herbert Fleischner, Egbert Mujuni, Daniël Paulusma, and Stefan Szeider. Covering graphs with few complete bipartite subgraphs. *Theor. Comput. Sci*, 410(21-23):2045–2053, 2009.

18 Leslie V Foster. Rank and null space calculations using matrix decomposition without column interchanges. *Linear Algebra and its Applications*, 74:47–71, 1986.

19 Mario Frank, Andreas P Streich, David Basin, and Joachim M Buhmann. Multi-assignment clustering for boolean data. *The Journal of Machine Learning Research*, 13(1):459–489, 2012.

20 Alan Frieze, Ravi Kannan, and Santosh Vempala. Fast monte-carlo algorithms for finding low-rank approximations. *Journal of the ACM (JACM)*, 51(6):1025–1041, 2004.

21 Alexander A Frolov, Dusan Husek, Igor P Muraviev, and P Yu Polyakov. Boolean factor analysis by attractor neural network. *Neural Networks, IEEE Transactions on*, 18(3):698–707, 2007.

22 Nicolas Gillis and Stephen A. Vavasis. On the complexity of robust PCA and ℓ_1-norm low-rank matrix approximation. *CoRR*, abs/1509.09236, 2015.

23 Gene Golub. Numerical methods for solving linear least squares problems. *Numerische Mathematik*, 7(3):206–216, 1965.

24 Ming Gu and Stanley C Eisenstat. Efficient algorithms for computing a strong rank-revealing QR factorization. *SIAM Journal on Scientific Computing*, 17(4):848–869, 1996.

25 Harold W Gutch, Peter Gruber, Arie Yeredor, and Fabian J Theis. Ica over finite fields: separability and algorithms. *Signal Processing*, 92(8):1796–1808, 2012.

26 Dorit S Hochbaum and Anu Pathria. Forest harvesting and minimum cuts: a new approach to handling spatial constraints. *Forest Science*, 43(4):544–554, 1997.

27 Yoo Pyo Hong and C-T Pan. Rank-revealing QR factorizations and the singular value decomposition. *Mathematics of Computation*, 58(197):213–232, 1992.

28 Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

29 Peng Jiang, Jiming Peng, Michael Heath, and Rui Yang. A clustering approach to constrained binary matrix factorization. In *Data Mining and Knowledge Discovery for Big Data*, pages 281–303. Springer, 2014.

30 Kari K. Karhunen. über lineare methoden in der wahrscheinlichkeitsrechnung. *Ann. Acad. Sci. Fennicae. Ser. A. I. Math.-Phys.*, 37:1–79, 1947.

31 Claudio Lucchese, Salvatore Orlando, and Raffaele Perego. Mining top-k patterns from binary datasets in presence of noise. In *SDM*, volume 10, pages 165–176, 2010.

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

33 Pauli Miettinen, Taneli Mielikainen, Aristides Gionis, Gautam Das, and Heikki Mannila. The discrete basis problem. *Knowledge and Data Engineering, IEEE Transactions on*, 20(10):1348–1362, 2008.

34 Pauli Miettinen, Taneli Mielikäinen, Aristides Gionis, Gautam Das, and Heikki Mannila. The discrete basis problem. *IEEE Trans. Knowl. Data Eng*, 20(10):1348–1362, 2008.

35 Amichai Painsky, Saharon Rosset, and Meir Feder. Generalized independent component analysis over finite alphabets. *Information Theory, IEEE Transactions on*, 2015.

36 C-T Pan. On the existence and computation of rank-revealing lu factorizations. *Linear Algebra and its Applications*, 316(1-3):199–222, 2000.

37 Ron M. Roth and Krishnamurthy Viswanathan. On the hardness of decoding the gale-berlekamp code. *IEEE Transactions on Information Theory*, 54(3):1050–1060, 2008.

38 Jouni K Seppänen, Ella Bingham, and Heikki Mannila. A simple algorithm for topic identification in 0–1 data. In *Knowledge Discovery in Databases: PKDD 2003*, pages 423–434. Springer, 2003.

39 Bao-Hong Shen, Shuiwang Ji, and Jieping Ye. Mining discrete patterns via binary matrix factorization. In John F. Elder IV, Françoise Fogelman-Soulié, Peter A. Flach, and Mohammed Javeed Zaki, editors, *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 757–766. ACM, 2009.

40 Tomáš Šingliar and Miloš Hauskrecht. Noisy-or component analysis and its application to link analysis. *The Journal of Machine Learning Research*, 7:2189–2213, 2006.

41 Larry Stockmeyer. The minimal set basis problem is NP-complete. IBM Research Report RC-5431, IBM Thomas J. Watson Research Center, 1975.

42 Jinsong Tan. Inapproximability of maximum weighted edge biclique and its applications. In Manindra Agrawal, Ding-Zhu Du, Zhenhua Duan, and Angsheng Li, editors, *TAMC 2008*, volume 4978 of *LNCS*, pages 282–293. Springer, 2008.

43 Joel A Tropp. Column subset selection, matrix factorization, and eigenvalue optimization. In *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 978–986. Society for Industrial and Applied Mathematics, 2009.

44 Jaideep Vaidya, Vijayalakshmi Atluri, and Qi Guo. The role mining problem: finding a minimal descriptive set of roles. In *Proceedings of the 12th ACM symposium on Access control models and technologies*, pages 175–184. ACM, 2007.

45 Leslie G. Valiant. Graph-theoretic arguments in low-level complexity. In Jozef Gruska, editor, *6th Symposium on Mathematical Foundations of Computer Science, MFCS 1977*, volume 53 of *Lecture Notes in Computer Science*, pages 162–176. Springer, 1977.

46 Yining Wang and Aarti Singh. Column subset selection with missing data via active sampling. In *AISTATS*, pages 1033–1041, 2015.

47 Tianbao Yang, Lijun Zhang, Rong Jin, and Shenghuo Zhu. An explicit sampling dependent spectral error bound for column subset selection. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 135–143, 2015.

48 Arie Yeredor. Independent component analysis over galois fields of prime order. *Information Theory, IEEE Transactions on*, 57(8):5342–5359, 2011.