Journal of Computer and System Sciences 80 (2014) 849–859

Contents lists available at ScienceDirect

Journal of Computer and System Sciences

www.elsevier.com/locate/jcss

Column Subset Selection Problem is UG-hard

A. Çivril[1]

Meliksah University, Computer Engineering Department, Aksu Sok. No:2 Talas, Kayseri 38280, Turkey

ARTICLE INFO

Article history:
Received 12 June 2012
Received in revised form 26 September 2012
Accepted 9 January 2014
Available online 16 January 2014

Keywords:
Matrices
Subset selection
Hardness
Unique Games Conjecture

ABSTRACT

We address two problems related to selecting an optimal subset of columns from a matrix. In one of these problems, we are given a matrix $A \in \mathbb{R}^{m\times n}$ and a positive integer k, and we want to select a sub-matrix C of k columns to minimize $\|A - \Pi_C A\|_F$, where $\Pi_C = CC^+$ denotes the matrix of projection onto the space spanned by C. In the other problem, we are given $A \in \mathbb{R}^{m\times n}$, positive integers c and r, and we want to select sub-matrices C and R of c columns and r rows of A, respectively, to minimize $\|A - CUR\|_F$, where $U \in \mathbb{R}^{c\times r}$ is the pseudo-inverse of the intersection between C and R. Although there is a plethora of algorithmic results, the complexity of these problems has not been investigated thus far. We show that these two problems are NP-hard assuming UGC.

© 2014 Elsevier Inc. All rights reserved.

1. Introduction

Given a matrix $A \in \mathbb{R}^{m\times n}$ as a set of column/row vectors, we are interested in the problem of obtaining a subset which stands as a good representative of A. The criteria that assess the quality of a subset are generally expressed via unitarily invariant norms of certain matrices related to the chosen subset. In what follows, we state a measure of quality for one of the problems we will consider, namely CSSP. We then continue to state both of the problems.

For the purposes of good reconstruction of a matrix A, in a geometric sense, one may want to select a subset of columns of this matrix so that they are as close as possible to the whole matrix. To formalize this, suppose we are given a column vector u which is not in the selected sub-matrix C. It is well known that the distance of u to the space spanned by a set C of column vectors is $\|u - CC^+u\|_2$ as the projection of u onto C is exactly CC^+u where C^+ is the pseudo-inverse of C. Thus, we can measure the quality of the chosen columns by $\|A - CC^+A\|_F$, which measures the sum of squared distances of all columns of A to the subspace spanned by C. Here, smaller value means a better subset. The importance of selecting such a subset has both theoretical and empirical interest. A good subset can be used to prevent from redundancy. Indeed, subset selection in matrices has been one of the main ingredients in statistical data analysis (e.g. [42]). The problem of selecting a good subset also occurs in various applications such as low-rank approximations to matrices [6,7], sparse solutions to least-squares regression [11], rank deficient least-squares problems [25,37] and sensor selection in a wireless network [38].

The subject matter of this paper is the following two problems:

Problem: Column Subset Selection Problem in Frobenius Norm (CSSP-F)
Instance: A real matrix $A \in \mathbb{R}^{m\times n}$, and $k \in \mathbb{Z}^+$.
Output: A sub-matrix $C \in \mathbb{R}^{m\times k}$ of A such that $\|A - CC^+A\|_F$ is minimum.

E-mail address: acivril@meliksah.edu.tr.

[1] Fax: +90 352 207 7349.

0022-0000/$ – see front matter © 2014 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jcss.2014.01.004

Problem: Column–Row Subset Selection Problem in Frobenius Norm (CRSSP-F)
Instance: A real matrix $A \in \mathbb{R}^{m \times n}$, and $c, r \in \mathbb{Z}^+$.
Output: Sub-matrices $C \in \mathbb{R}^{m \times c}$ and $R \in \mathbb{R}^{r \times n}$ of A such that $\|A - CUR\|_F$ is minimum, where U is the pseudo-inverse of the intersection between C and R.

These problems have received much attention in the past two decades and there is a large body of algorithmic results. This paper provides the first complexity theoretic results regarding these problems. We show that

Theorem 1.1. CSSP-F *and* CRSSP-F *do not have PTAS assuming the Unique Games Conjecture.*

Note that CSSP-F is in fact a special case of CRSSP-F, where one chooses $k = c$, $r = m$ and hence $R = A$. The hardness of CRSSP-F will naturally follow from the hardness of CSSP-F upon choosing the parameters accordingly. Before proceeding further, we would like to elaborate on why Unique Games Conjecture might be necessary to establish NP-hardness of these problems. As stated in the example above, a related but rather easy to analyze measure of quality for a subset is the volume of the parallelepiped they define. The problem stated via this measure is already known to be NP-hard and inapproximable unless $P = NP$ [15,16,41]. NP-hardness of this problem is easily established by a reduction from the Exact Cover by 3-sets Problem (X3C) [16] or from the Set Packing Problem [41]. In either case, one constructs a column vector of dimension n for each set in the set system, where n is the total number of elements in the ground set. One then assigns 1 to the coordinate corresponding to an element contained in the set and 0 otherwise. The volume turns out to be large if there exists a cover, and small in case there is no good covering. Unfortunately, the same strategy cannot be applied to the problems we consider in this paper. Because, they essentially ask to minimize a measure in which we take the whole matrix into account. More precisely, we need to know how far away the whole matrix is from the selected sub-matrix. X3C and Set Packing do not yield any information about how the constructed matrix is conditioned with respect to the selected sub-matrix, but they only provide a means to assess the quality of the selected sub-matrix alone. In other words, for CSSP-F and CRSSP-F one needs to be able to control the "global" information whereas local information is enough for the NP-hardness of the other aforementioned problems.

We give explicit examples in order to show why simple NP-completeness reductions do not work. In order to prove that CSSP-F is NP-hard, one needs to find a reduction from, say X3C to CSSP-F such that if there is an exact cover then there is a set of k column vector subset C of A such that $\|A - CC^+A\|_F \leqslant M$, and if there is no exact cover then for all k column vector subsets we have $\|A - CC^+A\|_F > M$. Note that the reduction should explicitly define k. Consider the following X3C instance where the ground set is $E = \{e_1, e_2, e_3, e_4, e_5, e_6\}$ and the 3-element subsets are $S_i = \{e_1, e_2, e_3\}$ for $1 \leqslant i \leqslant n$, $S_i = \{e_4, e_5, e_6\}$ for $n + 1 \leqslant i \leqslant 2n$, $S_i = \{e_1, e_2, e_4\}$ for $2n + 1 \leqslant i \leqslant 3n$, for some number n. Consider the obvious reduction which creates a $6 \times 3n$ matrix and assigning a 1 to a coordinate of the 6-dimensional column vector whenever the corresponding set contains the element representing that coordinate, and which further selects $k = 2$. Clearly there exists an exact cover for the instance we have defined. However, for n approaching infinity, the total distance of the matrix A to any selected subset is unbounded, i.e. $\|A - CC^+A\|_F$ is unbounded. Hence, one cannot even bound the desired value from above if such a naive reduction is assumed. It is not difficult to construct an X3C instance which does not have an exact cover, but the corresponding CSSP-F instance has bounded value. A similar reduction was offered by one of the referees. We briefly talk about it to show that it does not work for the same reason: One cannot control the global information of the problem. The suggested reduction is from the Edge Dominating Set Problem. It asks whether a graph has an edge dominating of size at most k, a subset of edges that covers all the edges in the graph with respect to adjacency. Suppose there are m vertices and n edges in the graph. Then, the reduction creates an $m \times n$ matrix (n column vectors of size m) by assigning 1 to a coordinate of the column vector whenever the corresponding edge is adjacent to the vertex representing that coordinate. It is not difficult to see that, in this case too, the value $\|A - CC^+A\|_F$ may be unbounded even if there is an edge dominating set of size 1: consider the star graph with arbitrarily large number of vertices. There are also NO instances of this problem which translates into a small $\|A - CC^+A\|_F$ under the reduction. Consider a 4-vertex graph: a, b, c, d and the edge set $\{a, b\}, \{c, d\}$. There is no edge dominating set of size 1 for this instance. However, for the resulting matrix, $\|A - CC^+A\|_F$ is smaller than that of the YES case for large enough n, no matter what k is.

So, one apparently needs to make further assumptions to arrive at any complexity theoretic result for these problems. We extend the work in [15] and show that if Unique Games Conjecture (UGC) is assumed, we can construct a matrix with extra useful structure so that we can quantify the distance of the columns of the whole matrix to the selected sub-matrix. This is possible due to the regular structure of the Label Cover problem. Indeed, by replacing each vertex of the Label Cover problem with a set of column vectors, i.e. by defining a column vector for each label positioned in a suitable manner, we fix their total number. We further control the quantity $\|A - CC^+A\|_F$ by assuming UGC. If UGC is not assumed, then based on the projection function, there might be several labels on the right hand side which is compatible with a single label on the left hand side. But, this creates the same problems we have mentioned above and prevents us from controlling the sum of distances. UGC helps us to guarantee that the label sizes are equal on both sides and no matter how one selects a sub-matrix of size k, the number of vectors that are not included in that sub-matrix and furthermore their total distances to the selected subspace is under control and can be analyzed.

2. Related work

There are various methods to construct concise representations of a matrix including random projections (e.g. [48]) and element-wise sparsification (e.g. [1]). In this section, we only list algorithmic results which construct matrix factorizations via selecting a subset of columns/rows of a matrix, as subset selection is the main focus of the paper. Note that, not all of the results below are directly related to the exact expressions in our problem statements, but they often come in different forms. Particularly, the quality of the solution is frequently compared to the best subspace revealed by the truncated SVD of the matrix (e.g. $\|A - A_k\|_\eta$), rather than the best choice of columns/rows. However, the complexity of the problems as we stated have already been posed as open problems in some of the work.

The theoretical computer science community has investigated the Column Subset Selection Problem (CSSP) by constructing a low-rank matrix approximation which is a k-dimensional subspace that approximates A_k in the spectral or Frobenius norm. The solutions developed thus far have mostly focused on randomized algorithms, and the set of columns selected by these algorithms have usually more than k columns which is proven to contain a k-dimensional subspace arbitrarily close to A_k. The seminal paper of Frieze, Kannan and Vempala [26] gives a randomized algorithm that selects a subset of columns $C \in \mathbb{R}^{m\times c}$ of A such that $\|A - \Pi_C A\|_F \leqslant \|A - A_k\|_F + \epsilon\|A\|_F$, where Π_C is a projection matrix obtained by the truncated SVD of C and c is a polynomial in k and $1/\epsilon$. Subsequent work [20,21] introduced several improvements on the dependence of c on k and $1/\epsilon$ also extending the analysis to the spectral norm, while Rudelson and Vershynin [46,47] provided results of the form $\|A - \Pi_C A\|_2 \leqslant \|A - A_k\|_2 + \epsilon\sqrt{\|A\|_2\|A\|_F}$. Approximations of the form $\|A - \Pi_C A\|_F \leqslant (1+\epsilon)\|A - A_k\|_F$ have also been provided [19,23]. Very recently, deterministic algorithms were given by Çivril and Magdon-Ismail [17], and Guruswami and Sinop [34] with the same performance guarantee. The result by Guruswami and Sinop [34] uses optimal number of columns. Both these results construct a space of dimensionality more than k. A hybrid algorithm for selecting exactly k columns was also proposed by Boutsidis et al. [7]. It selects k columns from a matrix A to approximate A_k, combining the random sampling schemes and the deterministic column pivoting strategies exploited by QR algorithms. Their algorithm provides a performance guarantee of the form $\|A - \Pi_C A\|_F \leqslant O(k\sqrt{\log k})\|A - A_k\|_F$. This paper explicitly poses the complexity of CSSP as an open problem. The result therein was further improved by Deshpande and Rademacher to $\sqrt{k+1}$ approximation by an efficient implementation of volume sampling.

The numerical linear algebra community implicitly provides deterministic solutions for CSSP by approximating A_k in the context of rank revealing QR factorizations, which primarily aim to determine the numerical rank of A. These algorithms select exactly k columns and provide results of the form $\|A - \Pi_C A\|_2 \leqslant p(k,n)\|A - A_k\|_2$ where $p(k,n)$ is a low degree polynomial in k and n [8–10,12,18,31,36,44,49].

Column–Row Subset Selection Problem (CRSSP) was investigated in the theoretical computer science community under the name CUR matrix decomposition. Authors of [22,45] and [24] provide algorithms with guarantees of the form $\|A - CUR\|_F \leqslant \|A - A_k\|_F + \epsilon\|A\|_F$ and $\|A - CUR\|_F \leqslant (1+\epsilon)\|A - A_k\|_F$ respectively. [24] poses the complexity of CRSSP as an open problem.

We would like to note that the algorithms providing RRQR factorizations listed above can also be used to construct CUR decompositions, i.e. the whole matrix can be approximated by the multiplication of three smaller matrices. Other important work which is not cited above includes those of Goreinov, Tyrtyshnikov and Zamarashkin [28–30] under the name pseudoskeleton approximations where they introduced the concept of maximum volume subset. They show that if the matrix $A \in \mathbb{R}^{m\times n}$ is approximated by a rank-k matrix within accuracy ϵ, then one can construct a factorization yielding $\|A - CUR\|_2 \leqslant \epsilon(1 + 2\sqrt{km} + 2\sqrt{kn})$.

As a side note related to our investigation, we would like to point out that most work on randomized matrix algorithms make use of the notion of statistical leverage which quantifies eigenvector localization. Eigenvector localization refers to the situation when most of the mass of an eigenvector is concentrated on a small number of coordinates. This phenomenon occurs in diverse scientific applications and in some of these cases it can be meaningfully interpreted in terms of certain structural irregularities in the data underlined by the existence of high degree nodes, localized small clusters etc (see [43] and references therein). So, it is not improbable on the complexity side, that the results of this paper in fact do hold in a more general setting that encompass the aforementioned notions. However, such an investigation is beyond the scope of this paper. Our main goal is rather to make a single technical contribution on a specific problem.

3. Preliminaries and notation

We introduce some preliminary notation and definitions. Vectors are denoted by small letters. $\|z\|_2$ or simply $\|z\|$ denotes the Euclidean norm of z. Matrices are denoted by capital letters. Given a matrix $A \in \mathbb{R}^{m\times n}$, we override the matrix notation with the set notation and denote the set $A = \{A_1, A_2, \ldots, A_n\}$ where A_i denotes the ith column of A. Similarly, we let $A = \{A_{(1)}, A_{(2)}, \ldots, A_{(m)}\}$ where $A_{(j)}$ denotes the jth row of A. When A is used with two indices, A_{ij} denotes the element of A on the ith row and jth column. $\|A\|_2$ and $\|A\|_F$ denote the spectral and the Frobenius norm of A respectively. Given a sub-matrix $C \in \mathbb{R}^{m\times k}$ of A, we also use the set theoretic notation $C = \{C_1, C_2, \ldots, C_k\}$ where C_i's are the columns of C. Similarly, for a sub-matrix $R \in \mathbb{R}^{k\times n}$ of A, we have $R = \{R_{(1)}, R_{(2)}, \ldots, R_{(k)}\}$. Given a set of column vectors S and a column vector v which is not in S, $\pi_S(v)$ is the orthogonal projection of v onto the space spanned by S. It is well known that $\pi_S(v) = SS^+v$, where S^+ is the pseudo-inverse of S (see for example [27]). We also define $d(v,S)$ to be the distance of v to the space spanned by S, namely $\|v - \pi_S(v)\|_2$.

4. Hardness of Subset Selection Problems

4.1. Label Cover problems and the Unique Games Conjecture

One of the most successful approaches to proving hardness results for combinatorial optimization problems is to create a gap preserving reduction from the Label Cover problem [2], which captures the expressive power of a 2-prover 1-round proof system. The celebrated PCP theorem [3,4] implies the NP-hardness of the Label Cover problem by a canonical reduction from the gap version of MAX-3SAT. In the Label Cover problem, we are given a bipartite graph with vertex sets U and V, and two sets of labels Σ_U and Σ_V for the vertices of U and V, respectively. The edges between U and V are associated with a relation from Σ_U to Σ_V, which are called constraints. The goal is to find a labeling for the vertices such that the number of satisfied edges is maximized, where we say that an edge is satisfied if the label of two incident vertices is an element of the relation associated with that edge.

Label Cover has been successfully used to derive a number of optimal inapproximability results (e.g. [35]) often incorporating tools from Fourier analysis. Some central problems of interest have resisted to yield strong inapproximability results under this general framework until a remarkable conjecture by Khot [39] helped characterize their hardness. Khot conjectured that a special type of Label Cover problem, which we call Unique Label Cover in this paper, is in fact NP-hard and exhibited reductions from this problem to some well known combinatorial optimization problems. To this date, several strong and often optimal hardness results have been established based on this conjecture, the so-called Unique Games Conjecture (UGC for short). One famous example is the hardness of MAX-CUT [40]. The key differences of a Unique Label Cover instance from the usual Label Cover instance are that the label sets are the same for both sides of the graph ($\Sigma_U = \Sigma_V$), and the relations associated with the edges are bijections. That is, the label of a vertex on one side uniquely determines the label of the vertex on the other side, given that the edge between them is satisfied. One further assumes that the gap version of the problem has imperfect completeness to avoid any trivialities. These extra assumptions turn out to provide significant power over the widely accepted $P \neq NP$ assumption. Although whether the conjecture holds is still a major open problem, there is no evidence against it thus far [13,32,50].

In this paper, we say that a problem is *UG-hard* if it is NP-hard assuming UGC. We will also make use of a special Unique Label Cover problem, which we call the *regular Unique Label Cover problem*. In an instance of this problem, both sides have the same number of vertices and each vertex is incident to the same number of edges, i.e. the graph is regular. Formally,

Definition 4.1. A regular Unique Label Cover instance (r-ULC) is a triple $L = (G = (U, V, E), \Sigma, \Pi)$ where

- $G(U, V, E)$ is a regular bipartite graph with vertex sets U and V, and the edge set E. Furthermore G satisfies the equality $|U| = |V| = n$, where each vertex has degree d (we assume $d \geqslant 25$).
- Σ is the label set associated with the vertices in U and V, with $\Sigma = \{1, 2, \ldots, R\}$.
- Π is the collection of constraints on the edge set, where the constraint on an edge e is defined as a bijection $\Pi_e : \Sigma \to \Sigma$.

The assumption $d \geqslant 25$ is particularly important for our reduction and is easily satisfied by considering a constant number of parallel repetitions. More specifically, it is clear that $d \geqslant 2$ since otherwise we trivially have perfect completeness. Hence, applying at most 5 parallel repetitions to the original r-ULC instance, we have $d \geqslant 25$ while blowing up the completeness by at most a constant factor. A labeling is an assignment to the vertices of the graph, $\sigma : \{U, V\} \to \Sigma$. It is said to satisfy an edge $e = (u, v)$ if $\Pi_e(\sigma(u)) = \sigma(v)$. The regular Unique Label Cover problem asks for an assignment σ such that the fraction of the satisfied edges is maximum. This fraction is called the *value* of the instance. The following is implied by the original Unique Games Conjecture. We do not repeat the construction to make the conjecture hold for our regular instance (see Lemma 1.5 of [14] for details):

Conjecture 4.2. *The UGC implies that for every fixed $\epsilon, \delta > 0$, there exists $R = R(\epsilon, \delta)$ such that it is NP-hard to decide whether a regular Unique Label Cover instance has value at least $1 - \epsilon$ or at most δ.*

4.2. The reduction

Before providing formal details, we give the rationale behind our construction and how they are related to the usual practice of proving hardness results. Most combinatorial optimization problems are defined on graphs or Boolean variables. In order to derive hardness results based on the PCP theorem, one replaces a vertex of a Label Cover instance with a Boolean hypercube whose vertices represent the variables or the vertices of the graph. The entities across different hypercubes are then connected in a suitable way to exploit the specific structure of the problem. This general paradigm was first introduced in [5] and has been successfully used (e.g. [35]). In our case, although we do not use the canonical jargon, we implicitly make use of the Boolean hypercube where the vertices of the hypercube represent vectors encoded in the usual sense, namely binary strings. However, we will only use a subset of the vertices on the hypercube. For example, the set of vectors we construct for CSSP-F will be expressed by the Hadamard code. Another practice suggested by the general paradigm is to show the soundness of the reduction by proving the contrapositive. One usually arrives at a contradiction in the Label Cover problem if the assumption that the soundness of the target problem exceeds the desired threshold. This method prevails specifically for constraint satisfaction problems. In contrast, we will follow the direct approach and show that if the Label Cover instance has soundness δ, then the target problem has soundness larger than its completeness, which will yield the desired result. We preferred such an approach as there is apparently no clear way of decoding the vectors with large objective value (i.e. $\|A - CC^+A\|_F$) back to an assignment for the labels.

4.2.1. The reduction for CSSP-F

Given an r-ULC instance L, we first define the matrix A in CSSP-F instance. To this aim, we define a column vector for each vertex-label pair in L, making $2nR$ vectors in total. (Note that $|U| = |V| = n$ and $|\Sigma| = R$.) Each vector is composed of nd "blocks", each of which represent a subspace allocated for an edge. We will use set of block vectors for labels, which can be represented via the rows of a Hadamard matrix. The same construction was used in [15]. Below, we give the lemma and restate the proof for the sake of completeness as the construction is relevant. First, we say that a vector z is a binary vector if its entries take values from the set $\{0, c\}$ for some constant $c > 0$. Let $\bar{z}$ denote the binary vector z with reverse entries, i.e. $\overline{z(i)} = 0$ if $z(i) = c$, and $\overline{z(i)} = c$ if $z(i) = 0$ for all i in the range of the coordinates of z, for some $c > 0$. Then, we have the following:

Lemma 4.3. *(See [15].) For $q \geqslant 2$, there exists a set of vectors $Z = \{z_1, \dots, z_{2^q-1}\}$ of dimension 2^q with binary entries such that the following three conditions hold*:

1. $\|z_i\|_2 = 1$ *for* $2^q - 1 \geqslant i \geqslant 1$.
2. $z_i \cdot \overline{z_j} = \frac{1}{2}$ *for* $2^q - 1 \geqslant i > j \geqslant 1$.
3. $z_i \cdot z_j = \frac{1}{2}$ *for* $2^q - 1 \geqslant i > j \geqslant 1$.

Proof. Consider the Hadamard matrix H of dimension $2^q \times 2^q$ with entries -1 and 1, constructed recursively by Sylvester's method. Let B be the $(2^q - 1) \times 2^q$ matrix consisting of the rows of H for which we replace -1's with 0's, excluding the all 1's row. We claim that the normalized rows of B satisfy the requirements. The first requirement is satisfied trivially. Note also that, for $q \geqslant 2$, two distinct rows of H (excluding the all 1's vector) have exactly 2^{q-2} element-wise dot-products of the following four types: $1 \cdot 1$, $1 \cdot (-1)$, $(-1) \cdot 1$, $(-1) \cdot (-1)$. Considering the construction of B, we have that the dot-product of any two of its rows is 2^{q-2} since all the products in H involving -1 vanishes for B. Similarly the dot-product of a row with the binary complement of another row is 2^{q-2} by symmetry. The entries of the normalized row vectors are $\frac{1}{2^{(q-1)/2}}$. Hence, these translate into dot products of $\frac{1}{2^{q-1}} \cdot 2^{q-2} = \frac{1}{2}$ for the normalized rows. Thus, the second and the third requirement also hold. □

By the lemma, there exists a set of R vectors each of dimension $2^{\lceil \log(R+1) \rceil}$ with the desired properties. Let $A_{u,a}$ be the vector for the vertex label pair $u \in U$ and $a \in \Sigma$. Similarly let $A_{v,b}$ be the vector for the pair $v \in V$ and $b \in \Sigma$. Both of these vectors are $nd2^{\lceil \log(R+1) \rceil}$ dimensional. The block of $A_{u,a}$ corresponding to an edge $e \in E$ is denoted by $A_{u,a}(e)$. The block of $A_{v,b}$ corresponding to an edge $e \in E$ is denoted by $A_{v,b}(e)$. We define

$$A_{u,a}(e) = \begin{cases} \frac{\overline{z_{\Pi_e(a)}}}{\sqrt{d}} & \text{if } e \text{ is incident to } u, \\ \vec{0} & \text{if } e \text{ is not incident to } u, \end{cases} \qquad A_{v,b}(e) = \begin{cases} \frac{z_b}{\sqrt{d}} & \text{if } e \text{ is incident to } v, \\ \vec{0} & \text{if } e \text{ is not incident to } v. \end{cases}$$

The normalization factor $1/\sqrt{d}$ is to make sure that the column vectors have Euclidean norm 1 (note that there are exactly d nonzero blocks of a vector). This construction ensures that if an edge $e = (u, v)$ is satisfied, then there are vectors defined for u and v which are orthogonal on e, resulting in a dot-product of 0. Otherwise, the dot-product is $\frac{1}{2d}$. Altogether, the matrix A is an $M \times N$ matrix with $M = nd2^{\lceil \log(R+1) \rceil}$ and $N = 2nR$, both having polynomial size in n, d and R. We define $k = 2n$ to complete the reduction.

We now informally explain why the gadget in Lemma 4.3 combined with this reduction will provide the hardness result we desire. Consider the YES case, i.e. where almost all the edges are satisfied. The labeling in this case translates into a selection of subset of column vectors C, since we have defined a vector for each vertex-label pair. For this choice of C, the distance of each vector to the subspace spanned by C has two main parts. Suppose the vector is defined for a label of $u \in U$, say $A_{u,a}$. These two parts are the distance of u to the selected vector of u in C (say $A_{u,b}$), and the distance of u to the selected vectors from all the vertices incident to u, which are all in V. Note that $A_{u,a}$ is orthogonal to all the other selected vectors except these. It turns out that if almost all the edges are satisfied, $A_{u,b}$ and the vectors selected from the vertices incident to u are almost orthogonal and we can *almost* apply the Pythagoras Theorem for all the projections of $A_{u,a}$ onto the vectors in C. The squared distance of $A_{u,a}$ to C is then given by {1 – the length of the total projection}. In the NO case, there is no labeling that satisfies more than a tiny fraction of the edges. Then, it turns out that for any choice of C, the selected vectors make acute angles with each other; more precisely, they are a little far away from being orthogonal. Hence, the length of the projection of $A_{u,a}$ onto the subspace spanned by the vectors in C turn out to be smaller than the quantity given by the Pythagoras Theorem. This means that $A_{u,a}$ is a little further away from C than it is in the YES case, which will establish our result. The whole argument crucially depends on the fact that there is exactly one vector defined for a specific vertex-label pair. This allows us to infer that the dot product of $A_{u,a}$ and $A_{u,b}$ is exactly $\frac{1}{2}$. This condition is enforced by the Unique Label Cover instance.

We would like to briefly restate the significance of UGC here. If we reduced from the usual Label Cover instance using the same gadget, one would possibly have several copies of the same vector from a vertex on one side, since in that case, the projection function for an edge is not necessarily a bijection. As explained in the introduction, this prevents us from analyzing the sum of the distances of all the non-selected column vectors to the space spanned by the selected ones. The Unique Label Cover instance has the nice property that, given our reduction, we know the exact positions of all the vectors we construct (e.g. they do not have duplicates) and hence can analyze those distances. Note that, we did not need to assume UGC for the volume maximization problem just because the quantity we try to control depends on the selected column vectors alone and the usual Label Cover instance does not present any obstacles for analyzing it.

4.2.2. The reduction for CRSSP-F

The reduction for CRSSP-F is exactly the same as that of CSSP-F with the following differences: We define the number of selected columns and rows to be $c = 2n$ and $r = M = nd2^{\lceil \log(R+1) \rceil}$. Hence, we essentially define C to be the exact same C in the CSSP-F reduction and $R = A$. As noted, this choice of parameters readily stem from the fact that CRSSP-F is a generalization of CSSP-F. The hardness of CRSSP-F will easily follow by considering a special case, which is essentially CSSP-F.

4.3. Analysis of the reduction for CSSP-F

4.3.1. Completeness

Consider an instance of r-ULC for which there is a labeling that satisfies more than $1-\epsilon$ fraction of the edges, i.e a labeling satisfying more than $(1-\epsilon)nd$ edges. Let C be the set of vectors corresponding to that labeling. That is, consider the set of vectors of the form $A_{u,a}$ and $A_{v,b}$ which satisfy $\sigma(u) = a$ and $\sigma(v) = b$ for all $u \in U$ and $v \in V$, where σ is the labeling function. Note that C contains exactly $2n$ column vectors. We will use the geometric view of the Frobenius norm in order to compute $\|A - CC^+A\|_F$; namely, we will compute the distances of all the column vectors not in C to the subspace defined by C. There are $2nR - 2n$ such column vectors. To this aim, we first define $\epsilon(u)$ to be the number of unsatisfied edges incident to u, so that we have $\sum_{u\in U} \epsilon(u) = \sum_{v \in V} \epsilon(v) \leqslant \epsilon nd$. Take a specific $A_{u,a} \in C$. We are interested in $d(A_{u,b}, C) = \|A_{u,b} - \pi_C(A_{u,b})\|_2 = \|A_{u,b} - CC^+A_{u,b}\|_2$, the distance of $A_{u,b}$ to C for all $b \neq a$. Clearly, the norm of the orthogonal projection of $A_{u,b}$ onto $A_{u,a}$ is $\|\pi_{A_{u,a}}(A_{u,b})\|_2 = A_{u,a} \cdot A_{u,b} = \frac{1}{2}$. Let $v_1, v_2, \dots, v_d$ be the neighbors of u and let $e_1, e_2, \dots, e_n$ be the corresponding edges between $A_{u,a}$ and v_i's. Let $A_{v_i,a_i} \in C$ for $i = 1, \dots, d$. Note that these vectors are all orthogonal to each other. If $A_{u,a}$ and A_{v_i,a_i} was orthogonal for all i, we would easily compute $d(A_{u,b}, C)$ and use the Pythagoras Theorem. However, some e_i might be unsatisfied. Hence, we consider the vector $A_{u,a'}$ which is the orthogonal part of $A_{u,a}$ with respect to A_{v_i,a_i}'s. Noting that $A_{u,a} \cdot A_{v_i,a_i} = 0$ if e_i is satisfied, $A_{u,a} \cdot A_{v_i,a_i} = \frac{1}{2d}$ if e_i is not satisfied and using the Gram–Schmidt procedure, we have

$$A_{u,a'} = A_{u,a} - \sum_{i=1}^{d} \frac{(A_{u,a} \cdot A_{v_i,a_i})A_{v_i,a_i}}{\|A_{v_i,a_i}\|_2^2} = A_{u,a} - \frac{1}{2d}\sum_{j=1}^{\epsilon(u)} A_{v_{i_j},a_{i_j}}$$

where v_{i_j} for $j = 1, \dots, \epsilon(u)$ in the last expression are the set of vertices for which e_{i_j} is not satisfied. Thus,

$$A_{u,a'} \cdot A_{u,b} = A_{u,a} \cdot A_{u,b} - \frac{1}{2d}\sum_{j=1}^{\epsilon(u)} A_{v_{i_j},a_{i_j}} \cdot A_{u,b} \geqslant \frac{1}{2} - \frac{1}{2d} \cdot \frac{1}{2d} \cdot \epsilon(u) = \frac{1}{2} - \frac{\epsilon(u)}{4d^2}.$$

The inequality follows since $A_{v_{i_j},a_{i_j}} \cdot A_{u,b} \leqslant \frac{1}{2d}$. Because, if e_{i_j} is satisfied $A_{v_{i_j},a_{i_j}} \cdot A_{u,b} = \frac{1}{2d}$. However, if e_{i_j} is not satisfied, that dot product *might* be 0. This observation also yields

$$\sum_{i=1}^{d} (A_{v_i,a_i} \cdot A_{u,b})^2 \geqslant \frac{1}{4d^2} \cdot \big(d - \epsilon(u)\big).$$

We gather the last two calculated quantity as

$$\begin{aligned}
\|CC^+A_{u,b}\|_2^2 &= (A_{u,a'} \cdot A_{u,b})^2 + \sum_{i=1}^{d} (A_{u,b} \cdot A_{v_i,a_i})^2 && (1)\\
&\geqslant \left(\frac{1}{2} - \frac{\epsilon(u)}{4d^2}\right)^2 + \frac{d - \epsilon(u)}{4d^2} && (2)\\
&= \frac{1}{4} + \frac{1}{4d} - \frac{\epsilon(u)}{2d^2} + \frac{(\epsilon(u))^2}{16d^4}. && (3)
\end{aligned}$$

Now, we will start computing our target value by summing over all the column vectors. First, calculate the vectors defined for u:

$$\sum_{\substack{b\in[R]\\ b\neq a}} \left\|A_{u,b}-CC^{+}A_{u,b}\right\|_2^2 \leqslant (R-1)\left(\frac{3}{4}-\frac{1}{4d}+\frac{\epsilon(u)}{2d^2}-\frac{(\epsilon(u))^2}{16d^4}\right).$$

Note that the sum of this expression for all the vertices in U is the same as the sum for V as the two sides are symmetric. Then, we get

$$\begin{aligned}
\left\|A-CC^{+}A\right\|_F^2 &= 2\sum_{u\in U}\sum_{\substack{b\in[R]\\ b\neq a}} \left\|A_{u,b}-CC^{+}A_{u,b}\right\|_2^2\\
&\leqslant 2n(R-1)\left(\frac{3}{4}-\frac{1}{4d}\right)+2(R-1)\sum_{u\in U}\left(\frac{\epsilon(u)}{2d^2}-\frac{(\epsilon(u))^2}{16d^4}\right)\\
&= 2n(R-1)\left(\frac{3}{4}-\frac{1}{4d}\right)+\frac{(R-1)}{d^2}\sum_{u\in U}\epsilon(u)-\frac{(R-1)}{8d^4}\sum_{u\in U}\bigl(\epsilon(u)\bigr)^2\\
&\leqslant 2n(R-1)\left(\frac{3}{4}-\frac{1}{4d}\right)+\frac{(R-1)}{d^2}\cdot\epsilon nd.
\end{aligned}$$

The first inequality directly follows from our previous calculation. In establishing the second inequality, we appealed to the fact that $\sum_{u\in U}\epsilon(u)\leqslant \epsilon nd$ and $\epsilon(u)\geqslant 0$ for any u. Rearranging the last expression, we get that there exists a sub-matrix C which satisfies

$$\left\|A-CC^{+}A\right\|_F^2 \leqslant 2n(R-1)\left(\frac{3}{4}-\frac{1}{4d}\right)+n(R-1)\left(\frac{\epsilon}{d}\right). \tag{4}$$

4.3.2. Soundness

Our goal is to show that, if there is no labeling that satisfies more than some δ fraction of the edges in the r-ULC instance, then for all sub-matrices C, $\|A-CC^{+}A\|_F^2$ is strictly larger than the right hand side of (4). As usual, proving the soundness is more involved. In our case, we rule out the possibility of a good sub-matrix via an argument which carefully considers all possible choices. First, we introduce some notation and terminology. Let $A_w=\bigcup_{a\in[R]}A_{w,a}$ for all $w\in U\cup V$. For a choice of sub-matrix C, let $C_w=A_w\cap C$. Given $|C_w|=t\geqslant 0$, we say that C “selects” t vectors from w. We have the following two cases with respect to the choice of C: (1) C selects more than one vector from a vertex. (2) C selects exactly one vector from each vertex.

Case 1: We will show that, in this case, the value $\|A-CC^{+}A\|_F$ turns out to be at least as large as than the case where C selects exactly one vector from each vertex. Then, the analysis reduces to the second case. Suppose that C selects ℓ vectors from a vertex u for some $\ell\geqslant 2$. Then, there exists at least $\ell-1$ vertices from which no vectors are selected. Let these vertices be $S=\{v_1,v_2,\dots,v_{\ell-1}\}$. Let $w_1,w_2,\dots,w_d$ be the neighbors of some $v\in S$. We know that the norm of the projection of $A_{v,a}$ for some $a\in[R]$ onto the t_i vectors that C selects from w_i can at most be $\frac{1}{d}$ by construction. Thus, summing the squares of $A_{v,a}$'s projections onto the vectors selected by C, we get $\|CC^{+}A_{v,a}\|_2^2\leqslant\frac{1}{d^2}\cdot d=\frac{1}{d}$, which means $\|A_{v,a}-CC^{+}A_{v,a}\|_2^2\geqslant 1-\frac{1}{d}$. Summing over all $a\in[R]$ and all v in S, we have

$$\sum_{\substack{v\in S\\ a\in[R]}} \left\|A_{v,a}-CC^{+}A_{v,a}\right\|_2^2 \geqslant (\ell-1)R\left(1-\frac{1}{d}\right)=\ell R-\frac{\ell R}{d}-R+\frac{R}{d}. \tag{5}$$

Hence, we have calculated the distances of all vectors that belong to vertices from which no vectors are selected from. Ideally, one also has to compute this value for the vectors of the vertex from which ℓ vectors are selected, sum these two quantities and compare it with the second case. However, to make our calculations more manageable, we will first show that if $\ell\geqslant 5$ the quantity in (5) already exceeds the value in the second case. Hence, we will find an upper bound for the sum of distances of all the non-selected vectors from ℓ vertices in the case where exactly one vector is selected from these ℓ vertices (thus, for a total of $\ell(R-1)$ vectors). Let w be a vertex among these ℓ vertices, say forming the set T, and let $A_{w,a}$ be the selected vertex from w. Let $\epsilon(w)$ be the number of unsatisfied edges incident to w. Then for a label $b\neq a$, we have already computed that (from (1))

$$\left\|CC^{+}A_{w,b}\right\|_2^2 \geqslant \frac{1}{4}+\frac{1}{4d}-\frac{\epsilon(u)}{2d^2}+\frac{(\epsilon(u))^2}{16d^4}\geqslant\frac{1}{4}+\frac{1}{4d}-\frac{\epsilon(u)}{2d^2}\geqslant\frac{1}{4}+\frac{1}{4d}-\frac{1}{2d}=\frac{1}{4}-\frac{1}{4d}$$

where the third inequality follows due to the fact that $d\geqslant\epsilon(u)$. Thus, summing over all the non-selected vectors and all the vertices in T, we get

$$\sum_{\substack{w\in T\\ b\in[R]-\{a\}}} \left\|A_{w,b}-CC^{+}A_{w,b}\right\|_2^2 \leqslant \ell(R-1)\left(\frac{3}{4}+\frac{1}{4d}\right)=\frac{3\ell R}{4}+\frac{\ell R}{4d}-\frac{3\ell}{4}-\frac{\ell}{4d}. \tag{6}$$

Subtracting (5) from (6), we have

$$\frac{\ell R}{4}-\frac{5\ell R}{4d}-R+\frac{R}{d}+\frac{3\ell}{4}+\frac{\ell}{4d}.$$

For $d \geqslant 25$ and $\ell \geqslant 5$, it is easy to verify that $\frac{\ell R}{4} \geqslant \frac{5\ell R}{4d}+R$. Thus, it remains to analyze the case $\ell \leqslant 4$, which requires the computations of the distances of the non-selected vectors of the vertex from which ℓ vectors are selected. We will then add this value to (5) and compare it again with (6). To underline our argument, we would like to note this computation is rather cumbersome for general ℓ and this is why we first ruled out the case $\ell \geqslant 5$.

Now, we have that C selects ℓ vectors from a vertex u for some $4 \geqslant \ell \geqslant 2$. We will consider each value of ℓ separately.

1. For $\ell = 2$, suppose that the vectors selected from u are $A_{u,a}$ and $A_{u,b}$. Let $w_1, w_2, \ldots, w_d$ be the neighbors of u. We will find a lower bound for the distance of a vector $A_{u,c}$ to the space spanned by C. It is clear that, as previously argued, the norm of the projection of $A_{u,c}$ onto the t_i vectors selected from w_i can be at most $1/d$. It is also clear that $A_{u,c}\cdot A_{u,a}=\frac{1}{2}$ by construction. Orthogonalizing $A_{u,b}$ against $A_{u,a}$ and then normalizing, we get $A'_{u,b}=(A_{u,b}-\frac{1}{2}A_{u,a})/(\frac{\sqrt{3}}{2})$. Here, we have used the fact that $\|(A_{u,b}-\frac{1}{2}A_{u,a})\|_2=\frac{\sqrt{3}}{2}$, which can be easily verified by looking at two of the vectors provided by Lemma 4.3. These are essentially the normalized rows of a 0–1 Hadamard matrix excluding the all 1's vector. As an example, consider the case $q = 2$ and the following vectors whose existence is guaranteed by the lemma:

   $$x=\left(\tfrac{1}{\sqrt{2}}\ \ 0\ \ \tfrac{1}{\sqrt{2}}\ \ 0\right),\qquad y=\left(\tfrac{1}{\sqrt{2}}\ \ \tfrac{1}{\sqrt{2}}\ \ 0\ \ 0\right),\qquad z=\left(0\ \ \tfrac{1}{\sqrt{2}}\ \ \tfrac{1}{\sqrt{2}}\ \ 0\right),$$

   where $x-\frac{1}{2}y=\left(\frac{1}{2\sqrt{2}}\ -\frac{1}{2\sqrt{2}}\ \frac{1}{\sqrt{2}}\ 0\right)$ with $\|x-\frac{1}{2}y\|_2=\frac{\sqrt{3}}{2}$. Hence, $A_{u,c}\cdot A'_{u,b}=\frac{2}{\sqrt{3}}(A_{u,c}\cdot A_{u,b}-\frac{1}{2}A_{u,c}\cdot A_{u,a})=\frac{2}{\sqrt{3}}(\frac{1}{2}-\frac{1}{4})=\frac{1}{2\sqrt{3}}$. Noting that all the dot products are positive between the vectors from u and w_i's (that they are forming acute angles) and using the Pythagoras Theorem, we have

   $$\left\|CC^{+}A_{u,c}\right\|_2^2 \leqslant (A_{u,c}\cdot A_{u,a})^2+\left(A_{u,c}\cdot A'_{u,b}\right)^2+\frac{1}{d^2}\cdot d=\frac{1}{4}+\frac{1}{12}+\frac{1}{d}=\frac{1}{3}+\frac{1}{d}.$$

   Summing the distances over all the $(R-2)$ vectors, we get

   $$\sum_{c\in[R]-\{a,b\}} \left\|A-CC^{+}A_{u,c}\right\|_2^2 \geqslant (R-2)\left(\frac{2}{3}-\frac{1}{d}\right)$$

   Summing this value with (5) and subtracting (6) for $\ell = 2$, we get

   $$\frac{R}{6}-\frac{5R}{2d}+\frac{5}{2d}+\frac{1}{6}.$$

   It is easy to see that this value is strictly greater than 0 for $d \geqslant 25$.
2. For $\ell = 3$, we don't need to reiterate Gram–Schmidt procedure and compute the coordinates in detail. The dot product $A_{u,c}\cdot A'_{u,b}=\frac{1}{2\sqrt{3}}$ computed in the case $\ell = 2$ is clearly an upper bound on the other dot products. More specifically for $\ell = 3$, suppose that the vectors selected from u are $A_{u,a}$, $A_{u,b}$ and $A_{u,c}$. Let $A_{u,a}$, $A'_{u,b}$ and $A'_{u,c}$ form an orthonormal set. For $A_{u,d}$, we have

   $$\left\|CC^{+}A_{u,d}\right\|_2^2 \leqslant (A_{u,d}\cdot A_{u,a})^2+\left(A_{u,d}\cdot A'_{u,c}\right)^2+\left(A_{u,d}\cdot A'_{u,b}\right)^2+\frac{1}{d^2}\cdot d=\frac{1}{4}+\frac{1}{12}+\frac{1}{12}+\frac{1}{d}=\frac{5}{12}+\frac{1}{d}.$$

   Summing the distances over all the $(R-2)$ vectors, we get

   $$\sum_{d\in[R]-\{a,b,c\}} \left\|A-CC^{+}A_{u,d}\right\|_2^2 \geqslant (R-2)\left(\frac{7}{12}-\frac{1}{d}\right)$$

   Summing this value with (5) and subtracting (6) for $\ell = 3$, we get

   $$\frac{R}{3}-\frac{15R}{4d}+\frac{15}{4d}+\frac{1}{2}.$$

   For $d \geqslant 25$, this value is strictly greater than 0.

3. Similarly, for $\ell = 4$, the dot product $A_{u,c} \cdot A'_{u,b} = \frac{1}{2\sqrt{3}}$ computed in the case $\ell = 2$ is an upper bound on the other dot products. Suppose that the vectors selected from u are $A_{u,a}$, $A_{u,b}$, $A_{u,c}$ and $A_{u,d}$. Let $A_{u,a}$, $A'_{u,b}$, $A'_{u,c}$ and $A'_{u,d}$ form an orthonormal set. For $A_{u,f}$, we have

$$\left\|CC^{+}A_{u,f}\right\|_2^2 \leqslant (A_{u,d} \cdot A_{u,a})^2 + 3\left(A_{u,d} \cdot A'_{u,d}\right)^2 + \frac{1}{d^2} \cdot d = \frac{1}{4} + \frac{1}{4} + \frac{1}{d} = \frac{1}{2} + \frac{1}{d}.$$

Summing the distances over all the $(R-2)$ vectors, we get

$$\sum_{f \in [R]-\{a,b,c,d\}} \left\|A - CC^{+}A_{u,f}\right\|_2^2 \geqslant (R-2)\left(\frac{1}{2} - \frac{1}{d}\right)$$

Summing this value with (5) and subtracting (6) for $\ell = 4$, we get

$$\frac{R}{2} - \frac{5R}{d} + \frac{5}{d} + 1,$$

which is again strictly greater than 0 for $d \geqslant 25$.

Case 2: In this case, C selects exactly one vector from each vertex. For a vertex u, let $A_{u,a}$ be the vector selected by C. We will compute the distances of all other vectors from u to the space spanned by C. Let $A_{u,b}$ such a vector and let $v_1, v_2, \ldots, v_d$ be the neighbors of u from which C selects $A_{v_1,a_1}, A_{v_2,a_2}, \ldots, A_{v_3,a_3}$ and let $e_1, e_2, \ldots, e_n$ be the corresponding edges between $A_{u,a}$ and v_i's. Noting that $A_{u,a} \cdot A_{v_i,a_i} = 0$ if (u, v_i) is satisfied and $A_{u,a} \cdot A_{v_i,a_i} = \frac{1}{2d}$ otherwise and denoting the number of satisfied edges adjacent to u by $\delta(u)$, the orthogonal part of $A_{u,a}$ against A_{v_i,a_i}'s is

$$A'_{u,a} = A_{u,a} - \sum_{i=1}^{d} \frac{(A_{u,a} \cdot A_{v_i,a_i})A_{v_i,a_i}}{\|A_{v_i,a_i}\|_2^2} = A_{u,a} - \frac{1}{2d} \sum_{j=1}^{d-\delta(u)} A_{v_{i_j},a_{i_j}},$$

where v_{i_j} for $j = 1, \ldots, d - \delta(u)$ are the set of vertices for which e_{i_j} is not satisfied. Thus,

$$A_{u,a'} \cdot A_{u,b} = A_{u,a} \cdot A_{u,b} - \frac{1}{2d} \sum_{j=1}^{d-\delta(u)} A_{v_{i_j},a_{i_j}} \cdot A_{u,b} = \frac{1}{2} - \frac{1}{2d} \cdot \frac{1}{2d} \cdot \left(d - \delta(u)\right) = \frac{1}{2} - \frac{(d - \delta(u))}{4d^2}.$$

Note that $\sum_{i=1}^{d} (A_{u,b} \cdot A_{v_i,a_i})^2 \leqslant \frac{1}{4d^2} \cdot d = \frac{1}{4d}$. So, the norm of the projection of $A_{u,b}$ onto C is

$$\begin{aligned}\left\|CC^{+}A_{u,b}\right\|_2^2 &= (A_{u,a'} \cdot A_{u,b})^2 + \sum_{i=1}^{d} (A_{u,b} \cdot A_{v_i,a_i})^2 \leqslant \frac{1}{4d} + \left(\frac{1}{2} - \frac{d - \delta(u)}{4d^2}\right)^2 \\ &= \frac{1}{4} + \frac{1}{4d} + \frac{1}{16d^4}\left(d^2 + \left(\delta(u)\right)^2 - 2d\delta(u)\right) - \frac{1}{4d^2}\left(d - \delta(u)\right) = \frac{1}{4} + \frac{1}{16d^2} + \frac{\delta(u)}{4d^2} - \frac{\delta(u)}{8d^3} + \frac{(\delta(u))^2}{16d^4}.\end{aligned}$$

Computing over all the vectors from u:

$$\sum_{\substack{b \in [R] \\ b \neq a}} \left\|A_{u,b} - CC^{+}A_{u,b}\right\|_2^2 \geqslant (R-1)\left(\frac{3}{4} - \frac{1}{16d^2} - \frac{\delta(u)}{4d^2} + \frac{\delta(u)}{8d^3} - \frac{(\delta(u))^2}{16d^4}\right).$$

This in turn yields

$$\begin{aligned}\left\|A - CC^{+}A\right\|_F^2 &= 2 \sum_{u \in U} \sum_{\substack{b \in [R] \\ b \neq a}} \left\|A_{u,b} - CC^{+}A_{u,b}\right\|_2^2 \\ &\geqslant 2n(R-1)\left(\frac{3}{4} - \frac{1}{16d^2}\right) + 2(R-1) \sum_{u \in U} \left(\frac{\delta(u)}{4d^3} - \frac{\delta(u)}{2d^2} - \frac{(\delta(u))^2}{8d^4}\right) \\ &\geqslant 2n(R-1)\left(\frac{3}{4} - \frac{1}{16d^2}\right) - (R-1) \sum_{u \in U} \frac{\delta(u)}{d^2} - (R-1) \sum_{u \in U} \frac{(\delta(u))^2}{4d^4} \\ &\geqslant 2n(R-1)\left(\frac{3}{4} - \frac{1}{16d^2}\right) - (R-1)\frac{\delta nd}{d^2} - (R-1)\frac{nd^2}{4d^4}\end{aligned}$$

$$
\begin{aligned}
&= 2n(R-1)\left(\frac{3}{4} - \frac{1}{4d}\right) + 2n(R-1)\left(\frac{1}{4d} - \frac{1}{16d^2}\right) - n(R-1)\left(\frac{\delta}{d} + \frac{1}{4d^2}\right)\\
&= 2n(R-1)\left(\frac{3}{4} - \frac{1}{4d}\right) + n(R-1)\left(\frac{1}{2d} - \frac{1}{8d^2} - \frac{\delta}{d} - \frac{1}{4d^2}\right)\\
&= 2n(R-1)\left(\frac{3}{4} - \frac{1}{4d}\right) + n(R-1)\left(\frac{1-2\delta}{2d} - \frac{3}{8d^2}\right),
\end{aligned}
$$

where in the third inequality we have used the fact that $\sum_{u\in U}\delta(u) \leqslant \delta nd$ and $\delta(u) \leqslant d$ for any u. Compare this expression with the one we found for the completeness, i.e. the expression (4):

$$
\left\|A - CC^{+}A\right\|_F^2 \leqslant 2n(R-1)\left(\frac{3}{4} - \frac{1}{4d}\right) + n(R-1)\left(\frac{\epsilon}{d}\right).
$$

It is clear that $(\frac{1-2\delta}{2d} - \frac{3}{8d^2}) > \frac{\epsilon}{d}$ in the limit of ϵ and δ approaching 0 and given the fact that $d \geqslant 25$. Hence, the soundness exceeds the completeness. In other words, the value returned by any subset of vectors in the NO case is larger than the value returned by a specific subset in the YES case. Recall that the hardness of CRSSP-F easily follows from the hardness of CSSP-F as it is a generalization. This completes the proof of Theorem 1.1. In fact, noting that d is a constant, the ratio of the values for the NO case and YES case tells us that there cannot be any PTAS for these problems assuming UGC.

5. Final remarks

We believe that our reduction implies UG-hardness of the problems we consider in spectral norm, too. However, we have not been able to show so. Indeed, Frobenius norm of a matrix has a natural geometric interpretation as the sum of squared norms or distances, whereas spectral norm is much more complicated to analyze. Another obvious question is whether one can find unconditional hardness results for these problems. We suspect it is possible by using the techniques of [33]. On the algorithmic side, one might also be interested in finding approximation algorithms for these problems comparing the value of the solution found by the algorithm $\|A - CC^{+}A\|$ with the optimal value $\|A - C_{opt}C_{opt}^{+}A\|$, rather than the value $\|A - A_k\|$. A constant factor approximation algorithm would be interesting. Similarly, in almost all the algorithmic results, the quality of the solution is compared to the best subspace revealed by the truncated SVD of the matrix (e.g. $\|A - A_k\|_F$), hence our result ruling out PTAS are not comparable with them.

Acknowledgments

We would like to thank all the referees for their insightful comments which improved the presentation of the paper.

References

[1] D. Achlioptas, F. Mcsherry, Fast computation of low-rank matrix approximations, J. ACM 54 (2007).
[2] S. Arora, C. Lund, Hardness of Approximations, PWS Publishing Company, 1997, pp. 399–446.
[3] S. Arora, C. Lund, R. Motwani, M. Sudan, M. Szegedy, Proof verification and the hardness of approximation problems, J. ACM 45 (3) (1998) 501–555.
[4] S. Arora, S. Safra, Probabilistic checking of proofs: a new characterization of NP, J. ACM 45 (1) (1998) 70–122.
[5] M. Bellare, O. Goldreich, M. Sudan, Free bits, pcps, and nonapproximability—towards tight results, SIAM J. Comput. 27 (1998) 804–915.
[6] C. Boutsidis, P. Drineas, M. Magdon-Ismail, Near optimal column-based matrix reconstruction, in: Proceedings of the 52nd Annual IEEE Foundations of Computer Science, 2011, pp. 305–314.
[7] C. Boutsidis, M.W. Mahoney, P. Drineas, An improved approximation algorithm for the column subset selection problem, in: Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms, 2009, pp. 968–977.
[8] T.F. Chan, Rank revealing QR factorizations, Linear Algebra Appl. 88/89 (1987) 67–82.
[9] T.F. Chan, P. Hansen, Low-rank revealing QR factorizations, Numer. Linear Algebra Appl. (1) (1994) 33–44.
[10] T.F. Chan, P.C. Hansen, Computing truncated singular value decomposition least squares solutions by rank revealing QR-factorizations, SIAM J. Sci. Stat. Comput. 11 (3) (1990) 519–530.
[11] T.F. Chan, P.C. Hansen, Some applications of the rank revealing QR factorization, SIAM J. Sci. Stat. Comput. 13 (3) (1992) 727–741.
[12] S. Chandrasekaran, I.C.F. Ipsen, On rank-revealing factorizations, SIAM J. Matrix Anal. Appl. 15 (1994) 592–622.
[13] M. Charikar, K. Makarychev, Y. Makarychev, Near-optimal algorithms for unique games, in: Proceedings of the 38th annual ACM Symposium on Theory of Computing, 2006, pp. 205–214.
[14] S. Chawla, R. Krauthgamer, R. Kumar, Y. Rabani, D. Sivakumar, On the hardness of approximating multicut and sparsest-cut, Comput. Complex. 15 (2006) 94–114.
[15] A. Çivril, M. Magdon-Ismail, Exponential inapproximability of selecting a maximum volume sub-matrix, Algorithmica 65 (1) (2013) 159–176.
[16] A. Çivril, M. Magdon-Ismail, On selecting a maximum volume sub-matrix of a matrix and related problems, Theor. Comput. Sci. 410 (47–49) (2009) 4801–4811.
[17] A. Çivril, M. Magdon-Ismail, Column subset selection via sparse approximation of SVD, Theor. Comput. Sci. 421 (2012) 1–14.
[18] F.R. de Hoog, R.M.M. Mattheijb, Subset selection for matrices, Linear Algebra Appl. (422) (2007) 349–359.
[19] A. Deshpande, S. Vempala, Adaptive sampling and fast low-rank matrix approximation, in: 10th International Workshop on Randomization and Computation, 2006, pp. 292–303.
[20] P. Drineas, A. Frieze, R. Kannan, S. Vempala, V. Vinay, Clustering in large graphs and matrices, in: Proceedings of the 10th Annual ACM-SIAM Symposium on Discrete Algorithms, 1999, pp. 291–299.

[21] P. Drineas, R. Kannan, M.W. Mahoney, Fast Monte Carlo algorithms for matrices II: Computing a low-rank approximation to a matrix, SIAM J. Comput. 36 (1) (2006) 158–183.

[22] P. Drineas, M.W. Mahoney, On the Nyström method for approximating a gram matrix for improved kernel-based learning, J. Mach. Learn. Res. 6 (2005) 2005.

[23] P. Drineas, M.W. Mahoney, S. Muthukrishnan, Subspace sampling and relative-error matrix approximation: Column-based methods, in: 10th International Workshop on Randomization and Computation, 2006, pp. 316–326.

[24] P. Drineas, M.W. Mahoney, S. Muthukrishnan, Relative-error CUR matrix decompositions, SIAM J. Matrix Anal. Appl. 30 (2008) 844–881.

[25] L. Foster, R. Kommu, Algorithm 853: An efficient algorithm for solving rank-deficient least squares problems, ACM Trans. Math. Softw. 32 (2006) 157–165.

[26] A. Frieze, R. Kannan, S. Vempala, Fast monte-carlo algorithms for finding low-rank approximations, J. ACM 51 (6) (2004) 1025–1041.

[27] G.H. Golub, C.V. Loan, Matrix Computations, Johns Hopkins Univ. Press, 1996.

[28] S.A. Goreinov, E.E. Tyrtyshnikov, A theory of pseudoskeleton approximations, Linear Algebra Appl. (261) (1997) 1–21.

[29] S.A. Goreinov, E.E. Tyrtyshnikov, The maximal-volume concept in approximation by low-rank matrices, in: Contemp. Math., vol. 280, AMS, 2001, pp. 47–51.

[30] S.A. Goreinov, N.L. Zamarashkin, E.E. Tyrtyshnikov, Pseudo-skeleton approximations by matrices of maximal volume, Mat. Zametki 62 (1997) 619–623.

[31] M. Gu, S.C. Eisenstat, Efficient algorithms for computing a strong rank-revealing QR factorization, SIAM J. Sci. Comput. 17 (4) (1996) 848–869.

[32] A. Gupta, K. Talwar, Approximating unique games, in: Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms, 2006, pp. 99–106.

[33] V. Guruswami, P. Raghavendra, R. Saket, Y. Wu, Bypassing ugc from some optimal geometric inapproximability results, in: Proceedings of the 23rd Annual ACM-SIAM Symposium on Discrete Algorithms, 2012, pp. 699–717.

[34] V. Guruswami, A.K. Sinop, Optimal column-based low-rank matrix reconstruction, in: Proceedings of the 23rd Annual ACM-SIAM Symposium on Discrete Algorithms, 2012, pp. 1207–1214.

[35] J. Håstad, Some optimal inapproximability results, J. ACM 48 (2001) 798–859.

[36] Y.P. Hong, C.T. Pan, Rank-revealing QR factorizations and the singular value decomposition, Math. Comput. 58 (1992) 213–232.

[37] I.C.F. Ipsen, C.T. Kelley, S.R. Pope, Rank-deficient nonlinear least squares problems and subset selection, SIAM J. Numer. Anal. 49 (2011) 1244–1266.

[38] S. Joshi, S. Boyd, Sensor selection via convex optimization, IEEE Trans. Signal Process. 57 (2009) 451–462.

[39] S. Khot, On the power of unique 2-prover 1-round games, in: Proceedings of the 34th annual ACM Symposium on Theory of Computing, 2002.

[40] S. Khot, G. Kindler, E. Mossel, R. O'Donnell, Optimal inapproximability results for max-cut and other 2-variable CSPs?, SIAM J. Comput. 37 (2007) 319–357.

[41] I. Koutis, Parameterized complexity and improved inapproximability for computing the largest j-simplex in a V-polytope, Inf. Process. Lett. 100 (2006) 8–13.

[42] F.G. Kuruvilla, P.J. Park, S.L. Schreiber, Vector algebra in the analysis of genome-wide expression data, Genome Biol. 3 (3) (2002).

[43] M.W. Mahoney, Randomize Algorithms for Matrices and Data, Found. Trends Mach. Learn., vol. 3, NOW, 2011.

[44] C.T. Pan, P.T.P. Tang, Bounds on singular values revealed by QR factorizations, BIT Numer. Math. 39 (1999) 740–756.

[45] P. Drineas, R. Kannan, M.W. Mahoney, Fast Monte Carlo algorithms for matrices III: Computing a compressed approximate matrix decomposition, SIAM J. Comput. 36 (1) (2006) 184–206.

[46] M. Rudelson, Random vectors in the isotropic position, J. Funct. Anal. 164 (1) (1999) 60–72.

[47] M. Rudelson, R. Vershynin, Sampling from large matrices: An approach through geometric functional analysis, J. ACM 54 (4) (2007).

[48] T. Sarlos, Improved approximation algorithms for large matrices via random projections, in: Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science, 2006, pp. 143–152.

[49] G.W. Stewart, G.W. Stewart, Four algorithms for the efficient computation of truncated pivoted qr approximations to a sparse matrix, Numer. Math. 83 (1998) 313–323.

[50] L. Trevisan, Approximation algorithms for unique games, Theory Comput. 4 (1) (2008) 111–128.