Electronic Colloquium on Computational Complexity, Revision 3 of Report No. 56 (2013)

Common information and unique disjointness

Gábor Braun and Sebastian Pokutta

ISyE, Georgia Institute of Technology, Atlanta, GA 30332, USA. *Email:* {gabor.braun, sebastian.pokutta}@isye.gatech.edu

September 30, 2015

Abstract

We provide an information-theoretic framework for establishing strong lower bounds on the nonnegative rank of matrices by means of common information, a notion previously introduced in Wyner [1975]. The framework is a generalization of the one in Braverman and Moitra [2012] for the shifted UDISJ (uniqe disjointness) matrix to arbitrary nonnegative matrices. Common information is a natural lower bound for the nonnegative rank of a matrix and by combining it with Hellinger distance estimations we can compute the (almost) exact common information of UDISJ (unique disjointness) partial matrix. The bounds are obtained very naturally. We also establish robustness of this estimation under various perturbations of the UDISJ partial matrix, where rows and columns are randomly or adversarially removed or where entries are randomly or adversarially altered. This robustness translates, via a variant of Yannakakis's Factorization Theorem, to lower bounds on the average case and adversarial approximate extension complexity of perturbations. We present the first family of polytopes, the hard pair introduced in Braun et al. [2012] related to the CLIQUE problem, with high average case and adversarial approximate extension complexity of perturbations. The framework relies on a strengthened version of the link between information theory and Hellinger distance from Bar-Yossef et al. [2004]. We also provide an information theoretic variant of the fooling set method that allows us to extend fooling set lower bounds from extension complexity to approximate extension complexity.

1 Introduction

Nonnegative matrix factorization plays a crucial role in many disciplines of theoretical computer science and mathematics, such as machine learning, data mining and data analysis, quantum mechanics, probability theory, communication complexity, convex geometry, polyhedral combinatorics, and many more. Nonnegative factorizations have also been studied very early on in information theory, however reinterpreting them as probability distributions, and the notion of common information introduced in Wyner [1975] provides a very natural information theoretic lower bound on the nonnegative rank. Despite its many applications in different disciplines, our analysis is conducted with approximate extended formulations in mind. In fact due to Yannakakis's factorization theorem (see Yannakakis [1988, 1991]) and the equivalence to a communication model given in Faenza et al. [2012] and Zhang [2012], it turns out that many open problems regarding the size of an optimal (exact or approximate) linear representation of a combinatorial optimization problem *are equivalent to* questions about the nonnegative rank of certain matrices and the related communication problems. Only recently, two major open problems in the theory of extended formulations were solved. In Rothvoß [2014] it was proven that every linear programming formulation of the matching polytope has exponential size. Later it was also established in Braun and Pokutta [2014] that for all fixed $0 < \varepsilon < 1$, even every linear program approximating the matching polytope by a factor of $(1 + \varepsilon/n)$ must have exponential size, where n is the number of nodes. In Chan et al. [2013] it was shown that MAXCUT cannot

ISSN 1433-8092

be approximated better than $1/2 - \varepsilon$ via *any* linear program of polynomial size. Other important problems include whether a generic polygon needs a linear number of inequalities in any linear representation in the worst case; for the latter for example the best-known lower bound is $\Omega(\sqrt{n})$ by Fiorini et al. [2012b] and the best-known upper bound is $\frac{6}{7}n$ by Shitov [2014]. We refer the interested reader to the excellent surveys Conforti et al. [2010] and Kaibel [2011] for an introduction.

A typical approach to lower bound the nonnegative rank is via techniques from communication complexity, see e.g., Faenza et al. [2012] and Zhang [2012] interpreting nonnegative rank as communication in expectation. Whereas in communication complexity only the support of the matrix matters, for nonnegative rank the actual value of the entries matter too, especially if the matrix has full support, i.e., all entries are positive. This requires improved arguments for lower bounding the nonnegative rank, like the hyperplane separation bound used for maximum matching in Rothvoß [2014]. Here we use the analogue of information cost in communication complexity, inspired by the recent work of Braverman and Moitra [2012], where a $2^{O(n^{\varepsilon})}$ lower bound on the nonnegative rank of the unique disjointness (partial) matrix (UDISJ) was obtained. Therefore with the elegance of a direct sum argument as in Bar-Yossef et al. [2004], we establish strong lower bounds on perturbations of UDISJ by means of information theory and common information in a generalized framework, which has no direct connection to communication. These bounds translate into lower bounds for the average case and the adversarial approximate extension complexity of perturbations. Although this work does not provide lower bounds for matrices significantly differing from the original UDISJ matrix, our approach can be applied to other matrices (see e.g., Braun and Pokutta [2014] for the case of the matching slack matrix).

Related work

While nonnegative factorizations have a huge variety of applications we will focus on the particular link between nonnegative matrix factorization, information theory and communication complexity, as well as extended formulations. Especially for the latter, nonnegative matrix factorizations and lower bounds for those are the main (arguably even the only) strong tools to establish lower bounds on the extension complexity. As mentioned above, the relation to extended formulations and nonnegative factorizations is established by the fundamental factorization theorem of Yannakakis (see Yannakakis [1988, 1991]). Given a polytope $P = \operatorname{conv}(v_1, \dots, v_n) = \{x : Ax \leq b\}$, a slack matrix of a polytope is given by the matrix $S_{ij} = b_i - A_i v_j$ for all i, j. Yannakakis's theorem establishes that the extension complexity $\operatorname{xc}(P)$ of a polytope P, that is the minimum number of linear inequalities needed in any linear programming formulation so that its feasible region linearly projects to the given polytope P, is equal to the nonnegative rank of any of the polytope's slack matrices S, i.e., $\operatorname{xc}(P) = \operatorname{rank}_+(S)$. Using this link, a super-polynomial lower bound have been established in Fiorini et al. [2012a] on the extension complexity of the correlation polytope, the cut polytope, the stable set polytope, and the TSP polytope. A crucial part of the proof is a strong lower bound on the nondeterministic communication complexity of the unique disjointness (partial) matrix (UDISJ), which was initially obtained by Wolf [2003] using Razborov [1992]. An existence proof of a polytope with high extension complexity, or equivalently of a slack matrix with high nonnegative rank, was given in Rothvoß [2011] via a beautiful counting argument. By means of a reduction mechanism, lower bounds have been also obtained for various other polytopes (see Avis and Tiwary [2013], Pokutta and Van Vyve [2013]) using the lower bound in Fiorini et al. [2012a].

The notion of extended formulations can be generalized to approximate extended formulations, giving rise to the notion of the ρ-approximate extension complexity where ρ is the performance guarantee. Here one considers polyhedral extensions that approximately project to a given polytope (e.g., relaxations as often used in approximation algorithms) and in Braun et al. [2012] it was shown that (a natural linear encoding of) the CLIQUE problem cannot be approximated within a factor better than $n^{1/2-\varepsilon}$ with a linear program using a polynomial number of inequalities. A similar inapproximability result was obtained for a certain spectrahedron (of small size) showing that SDPs have indeed much more expressive power than LPs. Subsequently, these bounds were improved to $n^{1-\varepsilon}$ in Braverman and Moitra [2012], matching the algorithmic inapproximability result of Håstad [1999] for CLIQUE. Recently, using our Corollary 5.10, strong lower bounds on the average case polyhedral complexity of the stable set problem have been established in Braun et al. [2013a].

Algorithms for nonnegative matrix factorizations have been considered, e.g., in Arora et al. [2012], Moitra [2013], Gillis [2012]. Regarding lower bounding techniques, works closely related to ours are the original work of Razborov [1992] and its generalization in Braun et al. [2012], the information theoretic approaches of Bar-Yossef et al. [2004] and Braverman and Moitra [2012], as well as the notion of common information introduced in Wyner [1975], the notion of correlation complexity from Zhang [2012], and the analysis of common information in terms of distribution approximation in Jain et al. [2013]. We combine crucial insights from these works linking them more closely together. As a follow-up to our work, in Braun et al. [2013b], it was shown that the amortized log nonnegative rank of a matrix is equal to the common information and sharp bounds on the convergence behavior have been established.

Contribution

Our main contributions can be separated into three parts. (A) We introduce a new framework for lower bounding the rank of a matrix by means of common information and Hellinger distance. (B) In this framework we can simplify and extend previous results for the lower bound on the nonnegative rank of the UDISJ matrix and (C) we sketch applications and implications for (approximate) extended formulations.

While the framework is geared towards lower bounding the nonnegative rank of a matrix, it is conceivable that some insights translate to communication complexity.

(A) *A generalized framework for lower bounding the nonnegative rank.* We interpret a nonnegative matrix factorization as *compression of correlation*, which leads to the well-known notion of common information introduced in Wyner [1975]: We regard a nonnegative matrix $M \in \mathbb{R}_+^{m \times n}$ (after scaling) as a joint probability distribution over rows and columns. A nonnegative factorization decomposes it to a sum of product distributions, i.e., making the row and column conditionally independent. The 'correlation complexity' of the distribution induced by M has to be captured by the random variable Π choosing the summand in the factorization. This is a straightforward generalization of the sampling procedure yielding a uniform distribution for the UDISJ matrix in [Braverman and Moitra, 2012, Algorithm 1], with the new element that the target distribution is now M.

The constructed probability space enables the use of information theoretic tools, as in Braverman and Moitra [2012], but here we explicitly use mutual information, (and hence common information), together with the direct sum property as in Bar-Yossef et al. [2004]. We derive a strengthened *cut-and-paste* property of the Hellinger distance, improving over the communication version given in Bar-Yossef et al. [2004], which is at the core of many proofs for establishing lower bounds later.

Our main tools for the analysis are of an information theoretic nature, inspired by the recent work Braverman et al. [2012a,b], and Hellinger distance from Bar-Yossef et al. [2004].

(B) *(Almost) Optimal lower bounds for (perturbations of) UDISJ.* We use the new paradigm to replicate and extend previous results for lower bounding the nonnegative rank of the UDISJ matrix in a very concise and consistent way. The UDISJ matrix is a partial matrix $M \subseteq \mathbb{R}_+^{[n] \times [n]}$ with $M(a, b) = 1$ if $a \cap b = \emptyset$ and $M(a, b) = 0$ if $|a \cap b| = 1$ with $a, b \subseteq [n]$. We first analyze (shifts of) the UDISJ matrix as those are of particular importance in the study of (approximate) extended formulations (see Braun et al. [2012] for details). To this end, we strengthen the core estimation of Bar-Yossef et al. [2004] via the new cut-and-paste property in Theorem 4.1. The obtained bounds for the common information of the UDISJ pattern (on n bit strings) of $\frac{6-3\log 3}{4}n$ when conditioning on disjoint strings and $\frac{2}{3}n$ in general are optimal for the case without shift and we provide matching factorizations of (a completion of) UDISJ realizing these bounds. The lower bounds on common information lead to lower bounds on the nonnegative rank optimal up to a small linear factor. In the case with shifts, the bounds on the common information are optimal up to a factor of $1/\ln 2$. Using the same framework we analyze various perturbations of the UDISJ pattern which are crucial for analyzing the average case and adversarial approximate extension complexity of a family of polytopes (see (C) below).

We obtain lower bounds as indicated in the table. In the following $[n] := \{1, \dots, n\}$. *Shifts* refer to adding a constant to each entry of the UDISJ pattern (i.e., for pairs $a, b \subseteq [n]$ with $|a \cap b| \in \{0, 1\}$), and *flipping* refers to replacing the entry in the matrix of a position a, b with $|a \cap b| = 0$ with one for $|a \cap b| = 1$ and vice versa.

Perturbation	$\log \operatorname{rank}_+ \geq$	Remarks
(0) UDISJ	$\frac{6-3\log 3}{4}n$	Optimal estimation
(1) Shifts of UDISJ	$\frac{1}{8\rho}n$	$(\rho - 1)$-shift
(2) Sets of fixed size $\frac{n}{4} + O(n^{1-\varepsilon})$	$\frac{n}{8\rho} - O(n^{1-\varepsilon})$	
Removing a fraction of rows and columns from UDISJ (remaining dimension indicated)		
(3) Random $2^{(1-\alpha)n} \times 2^{(1-\beta)n}$	$(\frac{1}{8\rho} - \alpha - \beta)n$	in expectation
(4) Adversarial $(1-\alpha)2^n \times (1-\beta)2^n$	$(\frac{1}{8\rho} - \alpha - \beta)n - \log 3$	removal of fractions per size
Flipping of a fraction τ of DISJ entries and NDISJ entries of (1)		
(5) Random	$\frac{1-2\tau}{8(\rho-\tau)}n - O(1)$	with high probability
(6) Adversarial	$\frac{\rho(1-10\tau)}{8(\rho-\tau)^2}n - O(1)$	with mild restrictions

The precise statements for (1) and (2) are to be found in Section 4 in Theorem 4.1, Proposition 4.3, and Theorem 4.4, (3) and (4) in Section 5.1 in Corollary 5.3, Corollary 5.7, and Corollary 5.10, and (5) and (6) in Section 5.2 in Theorem 5.11 and Theorem 5.13. Whereas cases (3) and (4) give rise to lower bounds on the average and adversarial extension complexity of the hard pair in Braun et al. [2012], cases (5) and (6) show that the UDISJ pattern is very rigid, i.e., even changing a large fraction of entries does not reduce the nonnegative rank significantly. In a polyhedral context these could be understood as moving vertices (changing a whole column of the slack matrix) which is captured by this model.

(C) *Applications to approximate extended formulations.* We provide the first example of a family of polytopes with high *average case approximate extension complexity* and *adversarial approximate extension complexity*. The considered family is the hard pair from Braun et al. [2012] and it is closely related to the CLIQUE problem; a more formal definition of the pair as well as approximate extension complexity is to be found in Section 2.2. The associated slack matrix has rows indexed by cliques and columns indexed by graphs (where we confine ourselves to stable sets only) and the entries denote the difference between the size of the clique and the largest clique in the graph. For the subsets of graphs that corresponds to stable sets the resulting matrix contains the UDISJ pattern as submatrix; we refer the interested reader to Braun et al. [2012] for more details. The variants studied in (B) correspond now to the removal of cliques or stable sets and translate to lower bounds on the *average case* and *adversarial approximate extension complexity* of the hard pair. More precisely, in Corollary 5.14 we show that when restricting to cliques and stable sets of a given size k, then the extension complexity remains high even for a ρ-approximate extended formulation. In Corollary 5.15 we show that even when an adversary can remove an α-fraction of cliques for each size k and a β-fraction of stable sets for each size k, then the ρ-approximate extension complexity remains high. We then combine, in

Corollary 5.16, both results and show that even if we only consider cliques and stable sets of a fixed size k and an adversary can remove a large fraction of cliques and stable sets, then the ρ-approximate extension complexity remains high. These new bounds are at the core of the uniform average case model in Braun et al. [2013a], showing that the maximum stable set problem has super-polynomial complexity for any random class of graphs, and hence the hardness is spread out among the graphs and not concentrated to a small fraction.

Finally, we obtain a new *information theoretic fooling set method*, in Corollary 6.1, from our framework that allows for obtaining lower bounds from 'approximate' fooling sets. In this context, a fooling set for a matrix M is a set of indices (a,b) so that $M(a,b) \neq 0$ however for any two distinct pairs $(a_1,b_1),(a_2,b_2)$, either $M(a_1,b_2) = 0$ or $M(a_2,b_1) = 0$ and the size of a fooling set is a lower bound on the nonnegative rank due to the rectangle property. We can relax this condition to only require that $M(a_1,b_2)$ and $M(a_2,b_1)$ need to be reasonably small. By doing so, any lower bound for the extension complexity of a polytope from a fooling set immediately gives rise to a lower bound for ρ-approximate extension complexity of that polytope where ρ is some small approximation factor.

Outline

In Section 2 we recall notions from information theory, including Hellinger distance of distributions, and prove basic lower bounds used later. We also provide a brief overview of approximate extended formulations. We then present the general framework for nonnegative factorizations in Section 3. In Section 4 we apply the framework to the UDISJ matrix where we obtain a very compact proof for its high nonnegative rank and we provide (almost) matching upper bounds. We then proceed with considering various perturbations of the UDISJ matrix, such as random and adversarial removal of rows and columns, and flipping of bits in Section 5. In Section 6 we introduce the approximate fooling set method and we conclude with some final remarks in Section 7. At the end of each section we provide implications for approximate extended formulations.

2 Preliminaries

2.1 Information theory and distance of distributions

We will now briefly recall basic notions from information theory. For a detailed introduction see Cover and Thomas [2006]. In the following, capital letters will represent random variables; we will slightly abuse notation sometimes and also use capital letters for events. Further, $\log(.)$ denotes the logarithm to base 2 and $\ln(.)$ is the natural logarithm. Let $\mathbb{H}[A] := \sum_{a \in \mathrm{range}(A)} \mathbb{P}[A = a] \log(1/\mathbb{P}[A = a])$ denote the *entropy* of a discrete random variable A and for $0 \leq p \leq 1$ let $\mathbb{H}[p] = p \log 1/p + (1-p) \log[1/(1-p)]$ be the entropy of a coin with bias p. This definition extends to *conditional entropy* $\mathbb{H}[A \mid B]$ by using the respective conditional distribution, but note that expectation is automatically taken: i.e., $\mathbb{H}[A \mid B] = \sum_b \mathbb{P}[B = b] \, \mathbb{H}[A \mid B = b]$.

Fact 2.1 (Properties of entropy)**.**

Obvious bounds $0 \leq \mathbb{H}[A] \leq \log|\mathrm{range}(A)|$;

Monotonicity $\mathbb{H}[A] \geq \mathbb{H}[A \mid B]$;

Chain rule $\mathbb{H}[A,B] = \mathbb{H}[A] + \mathbb{H}[B \mid A]$.

A central notion is the *mutual information* $\mathbb{I}[A;B] := \mathbb{H}[A] - \mathbb{H}[A \mid B]$ of two random variables A and B, which captures how much information about A is leaked by considering B instead. Formally, A and B can also be a collection of variables considered as one variable: a comma is used to separate the components of A or B, and a semicolon to separate A and B themselves: e.g., $\mathbb{I}[A_1, A_2; B] = \mathbb{I}[(A_1, A_2); B]$.

Mutual information is symmetric and extends to *conditional mutual information* $\mathbb{I}[A;B\,|\,C]$ by using the respective conditional distributions where C is a random variable. Expectation is also automatically taken here. Note that entropy is a special case: $\mathbb{H}[A] = \mathbb{I}[A;A]$.

We will condition on both *events and random variables* with the usual automatic expectation convention, as explained above. However, conditioning on a random variable Π, the conditional probability $\mathbb{P}[A = a\,|\,\Pi]$ is a function $\mathbb{P}[A = a\,|\,\Pi = \pi]$ in π, as customary.

Fact 2.2 (Properties of mutual information)**.**

Obvious bounds If A is a discrete variable, then $0 \leq \mathbb{I}[A;B\,|\,C] \leq \mathbb{H}[A] \leq \log|\mathrm{range}(A)|$

Chain rule $\mathbb{I}[A_1,A_2;B] = \mathbb{I}[A_1;B] + \mathbb{I}[A_2;B|A_1]$.

Symmetry

1. $\mathbb{I}[A;B] = \mathbb{I}[B;A]$
2. $\mathbb{I}[A;B] - \mathbb{I}[A;B\,|\,C]$ is symmetric in A, B, C.

Monotonicity $\mathbb{I}[A;B] \geq \mathbb{I}[A;B\,|\,C]$ if $\mathbb{I}[A;C\,|\,B] = 0$ or $\mathbb{I}[B;C\,|\,A] = 0$

Independent variables

1. If A and B are independent, then $\mathbb{I}[A;B] = 0$.
2. If $A_1, \dots, A_n$ are mutually independent, then $\mathbb{I}[A_1,\dots,A_n;C] \geq \sum_{i\in[n]} \mathbb{I}[A_i;C]$

In the proof of our main theorem we will rely on the (squared) Hellinger distance of two distributions.

Definition 2.3 (Hellinger distance)**.** Let μ_A, μ_B be discrete distributions over the same space. Then their *squared Hellinger distance* is

$$h^2(\mu_A;\mu_B) := 1 - \sum_{\pi} \sqrt{\mu_A(\pi)\mu_B(\pi)} = \frac{1}{2}\left\|\sqrt{\mu_A} - \sqrt{\mu_B}\right\|_2^2 \geq 0,$$

where $\mu_A(\pi)$ and $\mu_B(\pi)$ are the probabilities of the element π under μ_A and μ_B, respectively and $\sqrt{\mu_X}$ with $X \in \{A, B\}$ is to be understood coordinate-wise.

We will apply the following relation between Hellinger distance, entropy, and mutual information. The second part of Lemma 2.4 is well known and was already proven in [Bar-Yossef et al., 2004, Lemma 6.2].

Lemma 2.4. *Let A be a (generalized) binary random variable with values a_1, a_2, and Π an arbitrary random variable. Then*

$$\mathbb{H}[A\,|\,\Pi] \leq 2\sqrt{\mathbb{P}[A = a_1] \cdot \mathbb{P}[A = a_2]}(1 - h^2(\Pi|A = a_1;\Pi|A = a_2)).$$

In particular, if A is uniformly distributed then $\mathbb{I}[A;\Pi] \geq h^2(\Pi|A = a_1;\Pi|A = a_2)$.

Proof. By [Lin, 1991, Theorem 8], when $\mathbb{P}[\Pi = \pi] \neq 0$:

$$\begin{aligned}\mathbb{H}[A\,|\,\Pi = \pi] &\leq 2\sqrt{\mathbb{P}[A = a_1\,|\,\Pi = \pi] \cdot \mathbb{P}[A = a_2\,|\,\Pi = \pi]} \\ &= \frac{2\sqrt{\mathbb{P}[A = a_1] \cdot \mathbb{P}[A = a_2]}}{\mathbb{P}[\Pi = \pi]} \cdot \sqrt{\mathbb{P}[\Pi = \pi\,|\,A = a_1] \cdot \mathbb{P}[\Pi = \pi\,|\,A = a_2]}.\end{aligned}$$

Taking expectation proves the first claim. The second claim obviously follows from the first one. □

We will now provide a generalization to uniform variables with more values.

Lemma 2.5. *Let Z be a uniform random variable on n values, and Π another random variable. Then*

$$\mathbb{I}\left[Z;\Pi\right] \geq \log n - \frac{1}{n} \sum_{\substack{z_1,z_2\in\operatorname{range}(Z)\\ z_1\neq z_2}} \left(1 - h^2(\Pi|Z=z_1;\Pi|Z=z_2)\right). \tag{1}$$

Proof. First, we prove by induction on n that for all probability distributions $p_1,\dots,p_n$:

$$\mathbb{H}\left[p_1,\dots,p_n\right] \leq 2 \sum_{1\leq i<j\leq n} \sqrt{p_i p_j}. \tag{2}$$

The case $n = 1$ is clear, and $n = 2$ is [Lin, 1991, Theorem 8], see also [Bar-Yossef et al., 2004, Lemma 6.2]. For $n > 2$, let us choose an integer $1 < k < n$. Let X be a random variable with range $[n]$ and distribution $\mathbb{P}\left[X = i\right] = p_i$. Let $I_{X\leq k}$ be the indicator of the event $X \leq k$. Applying the induction hypothesis:

$$\begin{aligned}
\mathbb{H}\left[p_1,\dots,p_n\right] &= \mathbb{H}\left[X\right] = \mathbb{H}\left[X, I_{X\leq k}\right] = \mathbb{H}\left[I_{X\leq k}\right] + \mathbb{H}\left[X \,\middle|\, I_{X\leq k}\right] \\
&\leq 2\sqrt{(p_1+\dots+p_k)(p_{k+1}+\dots+p_n)} + (p_1+\dots+p_k) \sum_{1\leq i<j\leq k} 2\sqrt{\frac{p_i}{p_1+\dots+p_k}\cdot\frac{p_j}{p_1+\dots+p_k}} \\
&\qquad + (p_{k+1}+\dots+p_n) \sum_{k+1\leq i<j\leq n} 2\sqrt{\frac{p_i}{p_{k+1}+\dots+p_n}\cdot\frac{p_j}{p_{k+1}+\dots+p_n}} \\
&\qquad \leq 2\sum_{1\leq i\leq k}\sqrt{p_i}\cdot\sum_{k+1\leq j\leq n}\sqrt{p_j} + 2\sum_{1\leq i<j\leq k}\sqrt{p_i p_j} + 2\sum_{k+1\leq i<j\leq n}\sqrt{p_i p_j} = 2\sum_{i<j}\sqrt{p_i p_j}.
\end{aligned}$$

Now we turn to the proof of the lemma. For simplicity, let us assume that the range of Z is $[n]$ and we introduce the shorthand $p_i(\pi) = \mathbb{P}\left[Z = i \,|\, \Pi = \pi\right]$. Applying (2) to the distribution of Z conditioned on Π:

$$\mathbb{H}\left[Z \,|\, \Pi = \pi\right] \leq \sum_{i\neq j} \sqrt{p_i(\pi) p_j(\pi)}.$$

We take now expectation of both sides. Because (if Π is non-discrete, the left-hand side below should be the Radon–Nikodym derivative $\mathrm{d}(\Pi|Z=i)/\mathrm{d}\Pi$)

$$\frac{\mathbb{P}\left[\Pi = \pi \,|\, Z = i\right]}{\mathbb{P}\left[\Pi = \pi\right]} = \frac{\mathbb{P}\left[Z = i \,|\, \Pi = \pi\right]}{\mathbb{P}\left[Z = i\right]} = n p_i(\pi),$$

we obtain

$$\mathbb{H}\left[Z \,|\, \Pi\right] \leq \frac{1}{n} \sum_{i\neq j} \left(1 - h^2(\Pi|Z=i;\Pi|Z=j)\right).$$

As $\mathbb{I}\left[Z;\Pi\right] = \mathbb{H}\left[Z\right] - \mathbb{H}\left[Z \,|\, \Pi\right]$, the result follows. □

2.2 Approximate extended formulations

We will now briefly introduce the necessary notions and results from (approximate) extended formulations. For a more complete overview we refer the interested reader to the excellent surveys Conforti et al. [2010] and Kaibel [2011] as well as Pashkovich [2012], Braun et al. [2012].

The *approximate extended formulation* model is based on a pair of polyhedra $P \subseteq Q$. The facets of the outer polyhedron Q correspond to the (generators of) objective functions of interest, and the vertices of the inner polytope P correspond to feasible solutions. The *extension complexity* $\operatorname{xc}(P,Q)$ *of the pair* P, Q is defined to be the minimum number of facets of a polyhedron K having an affine image sandwiched between P and Q, i.e., there is an affine map $\operatorname{proj}\colon K \to Q$ with $P \subseteq \operatorname{proj} K$. We might want to think of (the projection of) K as being a relaxation of P which we only require to be exact for the objective functions generated over Q. The polyhedron Q is typically given by the inequalities of the form $cx \leq \max_{x \in P} cx$ where c is a linear objective function of interest. In order to study the size of such relaxations, we need the concept of a slack matrix.

Definition 2.6 (Slack matrix of a pair of polyhedra)**.** Given a polytope $P = \operatorname{conv}(v_1, \ldots, v_n)$ and a polyhedron $Q = \{x : Ax \leq b\}$, the *slack matrix of the pair* P, Q is given by the matrix $S_{ij} = b_i - A_i v_j$ for all i, j.

It turns out that Yannakakis's factorization theorem (see Yannakakis [1988, 1991]) extends to this case. Recall that the nonnegative rank $\operatorname{rank}_+ S$ of a nonnegative matrix S is the smallest nonnegative integer r such that $S = \sum_{i \in [r]} S_i$ is a sum of nonnegative matrices S_i of rank 1.

Theorem 2.7 (Pashkovich [2012], Braun et al. [2012])**.** *Let P, Q be a polyhedral pair and let S be any of its slack matrices. Then* $\operatorname{rank}_+ S - 1 \leq xc(P, Q) \leq \operatorname{rank}_+ S$. *(Equality holds if P, Q are polytopes)*

We obtain the notion of the ρ-approximate extension complexity of the pair P, Q

Definition 2.8 (ρ-approximate extension complexity)**.** Let P, Q be a polyhedral pair and let $\rho \geq 1$. The *ρ-approximate extension complexity of* P, Q is defined as $\operatorname{xc}(P, \rho Q)$, where ρQ is the ρ-dilate of Q (and we assume $P \subseteq \rho Q$).

This notion corresponds precisely to the minimum number of facets in any polyhedral relaxation of P so that for any linear objective function generated from Q (as positive combination of the facets) the maximum over the relaxation is within a factor of at most ρ compared to the maximum over P. This coincides with the standard notion of an approximation factor.

If S is a slack matrix for the pair $P = \operatorname{conv}(v_1, \ldots, v_n)$ and $Q = \{x : Ax \leq b\}$, then a slack matrix $\tilde{S}$ for the pair $P, \rho Q$ is obtained simply as $\tilde{S}_{ij} = S_{ij} + (\rho - 1) b_i$ with the above definition, i.e., we shift the slack matrix by adding positive entries. We obtain

Corollary 2.9. *Let $P = \operatorname{conv}(v_1, \ldots, v_n)$, $Q = \{x : Ax \leq b\}$ be a polyhedral pair and let S be the associated slack matrix. Then*

$$\operatorname{rank}_+(S + (\rho - 1)B) - 1 \leq \operatorname{xc}(P, \rho Q) \leq \operatorname{rank}_+(S + (\rho - 1)B),$$

where $B_{ij} = b_i$ for all i, j. (Equality holds if P, Q are polytopes)

We are mainly interested in the pair $P = \mathrm{COR}(n) := \operatorname{conv}\left(\{bb^T : b \in \{0,1\}^n\}\right)$, $Q = Q(n) = \left\{x \in \mathbb{R}_+^{n \times n} : \langle 2 \operatorname{diag}(a) - aa^T, x\rangle \leq 1, a \in \{0,1\}^n\right\}$ from Braun et al. [2012], where both the vertices of P and edges of Q are indexed by subsets of $[n]$, and the slack matrix M of the pair is $M(a, b) = (1 - |a \cap b|)^2$ which is an extension of the unique disjointness matrix. The vertices of P are considered as possible cliques, and the facets of Q are discrete subgraphs, i.e., stable sets.

3 Lower bounds via common information

Our approach is a combination of the sampling framework introduced in Braverman and Moitra [2012], previous lower bounding techniques given in Bar-Yossef et al. [2004] relying on the Hellinger distance, and common information approach of Wyner [1975]. We will use the following link between the nonnegative rank of a matrix and a latent random variable Π choosing rank-1 matrices in the decomposition, which is equivalent to the framework in Wyner [1975]. Recall that random variables are denoted by capital letters.

Definition 3.1 (Common information)**.** Let A, B be random variables, and Z an event, a random variable or a mixture of both. The *common information* of A, B given Z is the quantity

$$\mathbb{C}\left[A; B \,\middle|\, Z\right] := \inf_{\substack{\Pi : A \perp B | \Pi \\ \Pi \perp Z | A, B}} \mathbb{I}\left[A, B; \Pi \,\middle|\, Z\right], \tag{3}$$

where the infimum is taken over all random variables Π in all extensions of the probability space making

1. A and B conditionally independent given Π,
2. Z and Π conditionally independent given A and B.

The Π satisfying the above conditions will be called *seed*.

Remark 3.2. Common information was introduced in [Wyner, 1975, Eq (1.10)]. We extended this definition in the obvious way to the conditional version, which will play a crucial role later. The term *seed* for Π was adopted from Jain et al. [2013].

The conditional independence of Z and Π formulates the natural requirement to forbid Π making use of the external condition Z.

It is worthwhile to observe that the *partitions* in Razborov [1992] serve a similar purpose, i.e., making Alice and Bob conditionally independent.

Definition 3.3 (Induced distribution)**.** Let M be a nonnegative matrix. Its *induced distribution* consists of a random row A of M, and a random column B of M with probabilities

$$\mathbb{P}\left[A = a, B = b\right] = \frac{M(a,b)}{\sum_{x,y} M(x,y)}$$

for every row a and column b. We define the *common information* of M conditioned on Z as

$$\mathbb{C}\left[M \,|\, Z\right] \coloneqq \mathbb{C}\left[A; B \,|\, Z\right]. \tag{4}$$

Common information is a continuous measure of information contained in a factorization and it is easily seen that it lower bounds (the log of) the nonnegative rank as it bounds $\mathbb{H}\left[\Pi\right]$ from below. Note however that we consider the *common information conditioned on* Z to fine-tune the distribution of A, B for better bounds; we also need to condition as we will work with partial matrices and equivalently only partially-defined distributions.

Lemma 3.4. *Every nonnegative factorization of a nonnegative matrix M induces a seed with range of size of the number of summands in the factorization. In particular,* $\log \operatorname{rank}_+ M \geq \mathbb{C}\left[M \,|\, Z\right]$ *for any condition Z.*

Proof. Let a factorization of M be given by

$$M(a,b) = \sum_{\pi} \alpha_\pi(a)\beta_\pi(b).$$

We introduce a fresh random variable Π running through the index π in the factorization, therefore having the same number of values as the number of summands. Given A, B, the value of Π is chosen with private probabilities (in particular independent of Z given A, B)

$$\mathbb{P}\left[\Pi = \pi \,|\, A = a, B = b\right] = \frac{\alpha_\pi(a)\beta_\pi(b)}{M(a,b)}.$$

(When $M(a,b) = 0$, i.e., $\mathbb{P}\left[A = a, B = b\right] = 0$, then the distribution of Π can be chosen arbitrarily.) It readily follows that

$$\mathbb{P}\left[\Pi = \pi\right] = \frac{\sum_{x,y} \alpha_\pi(x)\beta_\pi(y)}{\sum_{x,y} M(x,y)},$$
$$\mathbb{P}\left[A = a, B = b \,|\, \Pi = \pi\right] = \frac{\alpha_\pi(a)\beta_\pi(b)}{\sum_{x,y} \alpha_\pi(x)\beta_\pi(y)}.$$

The right-hand side of the last formula is a product with every term depending only on either a, π or b, π, verifying the conditional independence of A and B given Π.

Finally, by choosing a minimal factorization, the range of Π has size $\operatorname{rank}_+ M$:

$$\log \operatorname{rank}_+ M \geq \mathbb{H}\left[\Pi\right] \geq \mathbb{I}\left[A, B; \Pi \,|\, Z\right] \geq \mathbb{C}\left[M \,|\, Z\right]. \qquad \square$$

The following lemma formulates the cut-and-paste property for correlation complexity, which is stronger than the communication version. It will be useful to lower bound common information.

Lemma 3.5 (Cut & paste). *Let A, B be discrete random variables conditionally independent given a third variable Π. Let $\Pi_{a,b}$ denote the distribution of Π conditioned on $A = a$ and $B = b$. Then*

$$\begin{aligned}\sqrt{\mathbb{P}\left[A = a_1, B = b_1\right]}\sqrt{\mathbb{P}\left[A = a_2, B = b_2\right]} \cdot \left(1 - h^2(\Pi_{a_1,b_1}; \Pi_{a_2,b_2})\right) \\ = \sqrt{\mathbb{P}\left[A = a_1, B = b_2\right]}\sqrt{\mathbb{P}\left[A = a_2, B = b_1\right]} \cdot \left(1 - h^2(\Pi_{a_1,b_2}; \Pi_{a_2,b_1})\right)\end{aligned} \tag{5}$$

for all values a_1, a_2 of A and values b_1, b_2 of B. As a consequence, for nonzero $\mathbb{P}\left[A = a_1, B = b_1\right]$ and $\mathbb{P}\left[A = a_2, B = b_2\right]$:

$$h^2(\Pi_{a_1,b_1}; \Pi_{a_2,b_2}) \geq 1 - \sqrt{\frac{\mathbb{P}\left[A = a_1, B = b_2\right]\mathbb{P}\left[A = a_2, B = b_1\right]}{\mathbb{P}\left[A = a_1, B = b_1\right]\mathbb{P}\left[A = a_2, B = b_2\right]}}. \tag{6}$$

As a special case when A, B are the random row and column of the induced distribution of a nonnegative matrix M, and $M(a_1, b_1)$ and $M(a_2, b_2)$ are nonzero,

$$h^2(\Pi_{a_1,b_1}; \Pi_{a_2,b_2}) \geq 1 - \sqrt{\frac{M(a_1, b_2)M(a_2, b_1)}{M(a_1, b_1)M(a_2, b_2)}}. \tag{7}$$

Remark 3.6. If $\mathbb{P}\left[A = a, B = b\right] = 0$ then $\Pi_{a,b}$ is undetermined but for the statement above, it can be chosen arbitrarily.

Proof. By independence,

$$\mathbb{P}\left[A = a_1, B = b_1 \mid \Pi\right] \cdot \mathbb{P}\left[A = a_2, B = b_2 \mid \Pi\right] = \mathbb{P}\left[A = a_1, B = b_2 \mid \Pi\right] \cdot \mathbb{P}\left[A = a_2, B = b_1 \mid \Pi\right].$$

This can be written (if Π is discrete) via multiplying with $\mathbb{P}\left[\Pi = \pi\right]^2$ as

$$\begin{aligned}&\mathbb{P}\left[A = a_1, B = b_1\right]\mathbb{P}\left[\Pi = \pi \mid A = a_1, B = b_1\right] \cdot \mathbb{P}\left[A = a_2, B = b_2\right]\mathbb{P}\left[\Pi = \pi \mid A = a_2, B = b_2\right] \\ &= \mathbb{P}\left[A = a_1, B = b_2\right]\mathbb{P}\left[\Pi = \pi \mid A = a_1, B = b_2\right] \cdot \mathbb{P}\left[A = a_2, B = b_1\right]\mathbb{P}\left[\Pi = \pi \mid A = a_2, B = b_1\right], \\ &\qquad \pi \in \operatorname{range}(\Pi).\end{aligned}$$

(In general, instead of $\mathbb{P}\left[\Pi = \pi \mid A = a, B = b\right]$ one should write the Radon–Nikodym derivative $\mathrm{d}\Pi_{a,b}/\mathrm{d}\Pi$ above, and integrate instead of summing up below.) Taking square root and summing up

$$\begin{aligned}&\sqrt{\mathbb{P}\left[A = a_1, B = b_1\right]\mathbb{P}\left[A = a_2, B = b_2\right]} \cdot \left(1 - h^2(\Pi_{a_1,b_1}; \Pi_{a_2,b_2})\right) \\ &= \sqrt{\mathbb{P}\left[A = a_1, B = b_2\right]\mathbb{P}\left[A = a_2, B = b_1\right]} \cdot \left(1 - h^2(\Pi_{a_1,b_2}; \Pi_{a_2,b_1})\right) \\ &\leq \sqrt{\mathbb{P}\left[A = a_1, B = b_2\right]\mathbb{P}\left[A = a_2, B = b_1\right]}.\end{aligned}$$

Using the latter inequality, it also follows

$$h^2(\Pi_{a_1,b_1}; \Pi_{a_2,b_2}) \geq 1 - \sqrt{\frac{\mathbb{P}\left[A = a_1, B = b_2\right]\mathbb{P}\left[A = a_2, B = b_1\right]}{\mathbb{P}\left[A = a_1, B = b_1\right]\mathbb{P}\left[A = a_2, B = b_2\right]}}. \qquad \square$$

4 An almost tight lower bound for (shifts of) UDISJ

In this section we will estimate the common information of matrices containing (shifts of) the UDISJ patterns. We prove the following main theorem.

Theorem 4.1. *Let M be a nonnegative matrix with rows and columns indexed by all subsets of $[n]$ satisfying*

$$M(a,b) = \begin{cases} 1 & \text{if } a \cap b = \emptyset \\ 1-\varepsilon & \text{if } |a \cap b| = 1 \end{cases} \tag{8}$$

for all $a, b \subseteq [n]$. (The other entries can be arbitrary nonnegative numbers.) Let $C = (C_1, \dots, C_n)$ be a collection of n fair coins independent of the induced distribution of M. If C_i is heads, then let D_i be the indicator of i belonging to the subset A indexing the random row of M. If C_i is tails, let D_i be the indicator of i belonging to the subset B indexing the random column. Let $D = (D_1, D_2, \dots, D_n)$ be the collection of the D_i; as a shorthand let $D = 0$ denote $D_1 = 0, \dots, D_n = 0$. Then

$$\mathbb{C}\left[M \mid D = 0, C\right] \geq \frac{\varepsilon n}{8}. \tag{9}$$

Proof. Let A_i and B_i be the indicator of $i \in A$ and $i \in B$, respectively. Let Π be a seed for A, B. We reduce the analysis to the case $n = 1$. Observe that $\mathbb{P}\left[A = a, B = b, C = c \mid D = 0\right] = \mathbb{P}\left[A = a, B = b, C = c\right] / \mathbb{P}\left[D = 0\right] = 1/4^n$ provided the values a, b, c imply $D = 0$. In other words, A, B, C are jointly uniformly distributed conditioned on $D = 0$, hence the pairs $\{(A_j, B_j) : j \in [n]\}$ are independent given $D = 0, C$, so that

$$\mathbb{I}\left[A, B; \Pi \mid D = 0, C\right] \geq \sum_{j \in [n]} \mathbb{I}\left[A_j, B_j; \Pi \middle| D = 0, C\right]$$

Now observe that the distribution of A_j, B_j, Π, D_j, C_j given $D_i = 0$ and C_i for all $i \neq j$ satisfies the assumptions for the case $n = 1$. This can be seen by computing the probabilities via summing up the ones of A, B. Therefore the case $n = 1$ provides $\mathbb{I}\left[A_j, B_j; \Pi \middle| D = 0, C\right] \geq \varepsilon/8$, which concludes the proof as $\sum_{j \in [n]} \mathbb{I}\left[A_j, B_j; \Pi \middle| D = 0, C\right] \geq \frac{\varepsilon n}{8}$ follows.

It remains to prove the case $n = 1$. Suggestively, we write $\mathbb{A}$ for heads and $\mathbb{B}$ for tails, so e.g., $D_1 = A_1$ if $C_1 = \mathbb{A}$. In a first step we identify the terms that need to be estimated:

$$\begin{aligned} \mathbb{I}\left[A, B; \Pi \mid D = 0, C\right] &= \mathbb{I}\left[A_1, B_1; \Pi \mid D_1 = 0, C_1\right] \\ &= \frac{\mathbb{I}\left[A_1, B_1; \Pi \mid D_1 = 0, C_1 = \mathbb{A}\right] + \mathbb{I}\left[A_1, B_1; \Pi \mid D_1 = 0, C_1 = \mathbb{B}\right]}{2} \end{aligned}$$

Now the event $D_1 = 0, C_1 = \mathbb{A}$ is the same as $A_1 = 0, C_1 = \mathbb{A}$. As C_1 is independent of A_1, B_1, Π (recall that Π is a seed), we obtain

$$\mathbb{I}\left[A_1, B_1; \Pi \mid D_1 = 0, C_1 = \mathbb{A}\right] = \mathbb{I}\left[A_1, B_1; \Pi \mid A_1 = 0\right].$$

Let Π_{ab} denote the distribution of Π given $A_1 = a$ and $B_1 = b$. As A_1, B_1 is a uniform binary variable given $A_1 = 0$ by Equation (8), Lemma 2.4 applies:

$$\mathbb{I}\left[A_1, B_1; \Pi \mid A_1 = 0\right] \geq h^2(\Pi_{00}; \Pi_{01}).$$

All in all, we obtain

$$\mathbb{I}\left[A_1, B_1; \Pi \mid D_1 = 0, C_1 = \mathbb{A}\right] \geq h^2(\Pi_{00}; \Pi_{01}).$$

Similarly,

$$\mathbb{I}\left[A_1, B_1; \Pi \mid D_1 = 0, C_1 = \mathbb{B}\right] \geq h^2(\Pi_{00}; \Pi_{10}).$$

Thus

$$\begin{aligned}\mathbb{I}\left[A,B;\Pi\,|\,D=0,C\right] &\geq \frac{h^2(\Pi_{00};\Pi_{01})+h^2(\Pi_{00};\Pi_{10})}{2} \geq \frac{\left(h(\Pi_{00};\Pi_{01})+h(\Pi_{00};\Pi_{10})\right)^2}{4}\\ &\geq \frac{h^2(\Pi_{01};\Pi_{10})}{4} \geq \frac{\varepsilon}{8},\end{aligned}$$

where the second inequality follows with Cauchy-Schwarz and the third one is the triangle inequality. The last inequality follows from Lemma 3.5 by the independence of A_1 and B_1 given Π:

$$h^2(\Pi_{01};\Pi_{10}) \geq 1-\sqrt{\frac{\mathbb{P}\left[A_1=0,B_1=0\right]\cdot\mathbb{P}\left[A_1=1,B_1=1\right]}{\mathbb{P}\left[A_1=0,B_1=1\right]\cdot\mathbb{P}\left[A_1=1,B_1=0\right]}} = 1-\sqrt{1-\varepsilon} \geq \frac{\varepsilon}{2}. \tag{10}$$

□

Remark 4.2. The proof of Theorem 4.1 is mostly identical to the one in Bar-Yossef et al. [2004], except the better cut-and-paste relation. This is due to considering *correlation compression* instead of a *protocol* manifesting in $\mathbb{P}\left[\Pi\,|\,A=a,B=b\right]$, and not $\mathbb{P}\left[A=a,B=b\,|\,\Pi\right]$, decomposing into a product $\alpha_\pi(a)\beta_\pi(b)$ (see Lemma 3.5).

We will now provide an upper bound on the common information of the matrices occurring in Theorem 4.1, showing that our estimation is tight up to a factor of $1/\ln 2$. The factor stems from estimating mutual information by the squared Hellinger distance via Lemma 2.4. We will also derive the exact common information for the case $\varepsilon = 1$.

Proposition 4.3 (Common Information of UDISJ). *With C, D as in Theorem 4.1:*

1. *The lower bound of Theorem 4.1 is optimal up to a factor of* $1/\ln 2$ *for small ε: There is an extension M of UDISJ with*
$$\mathbb{C}\left[M\,|\,D=0,C\right] \leq \left(\frac{\varepsilon}{8\ln 2}+O(\varepsilon^2)\right)n.$$

2. *For $\varepsilon = 1$, we have for all extensions M*
$$\log\operatorname{rank}_+(M) \geq \mathbb{C}\left[M\,|\,D=0,C\right] \geq \frac{6-3\log 3}{4}\cdot n \approx 0.3113\cdot n,$$
and there is an M realizing this bound.

Proof. We establish the case $n = 1$ and then generalize to all n by a simple tensor argument. We consider the explicit decomposition of

$$M = \begin{pmatrix}1 & 1\\ 1 & 1-\varepsilon\end{pmatrix} = \begin{pmatrix}\frac{1}{2} & \frac{1-\sqrt{\varepsilon}}{2}\\ \frac{1+\sqrt{\varepsilon}}{2} & \frac{1-\varepsilon}{2}\end{pmatrix} + \begin{pmatrix}\frac{1}{2} & \frac{1+\sqrt{\varepsilon}}{2}\\ \frac{1-\sqrt{\varepsilon}}{2} & \frac{1-\varepsilon}{2}\end{pmatrix}$$

into nonnegative rank-1 matrices.

Even though the mutual information with the induced seed Π can be computed directly, we present a short calculation explaining the terms of the formula. Let $I(A = B)$ denote the indicator of the event $A = B$, which is independent of Π. Also note that A, B and Π are independent given $A = B$. These independences imply

$$\begin{aligned}\mathbb{I}\left[A,B;\Pi\right] &= \mathbb{I}\left[A,B;\Pi\,|\,I(A=B)\right]\\ &= \mathbb{P}\left[A=B\right]\cdot\underbrace{\mathbb{I}\left[A,B;\Pi\,|\,A=B\right]}_{0}+\mathbb{P}\left[A\neq B\right]\cdot\mathbb{I}\left[A,B;\Pi\,|\,A\neq B\right]\\ &= \mathbb{P}\left[A\neq B\right]\left(\mathbb{H}\left[A,B\,|\,A\neq B\right]-\mathbb{H}\left[A,B\,|\,A\neq B,\Pi\right]\right)\\ &= \frac{2}{4-\varepsilon}\left(1-\mathbb{H}\left[\frac{1+\sqrt{\varepsilon}}{2}\right]\right) = \frac{\varepsilon}{4\ln 2}+O(\varepsilon^2)\end{aligned}$$

For the conditional mutual information, we follow a straightforward approach:

$$\mathbb{I}\left[B;\Pi \,|\, A=0\right] = \mathbb{H}\left[B \,|\, A=0\right] - \mathbb{H}\left[B \,|\, A=0,\Pi\right] = 1 - \frac{1}{2}\mathbb{H}\left[\frac{1}{2+\sqrt{\varepsilon}}\right] - \frac{1}{2}\mathbb{H}\left[\frac{1}{2-\sqrt{\varepsilon}}\right],$$

and similarly with A, B exchanged. Hence

$$\begin{aligned}\mathbb{I}\left[A,B;\Pi \,|\, D=0,C\right] &= \frac{\mathbb{I}\left[A,B;\Pi \,|\, A=0\right] + \mathbb{I}\left[A,B;\Pi \,|\, B=0\right]}{2} \\ &= 1 - \frac{1}{2}\mathbb{H}\left[\frac{1}{2+\sqrt{\varepsilon}}\right] - \frac{1}{2}\mathbb{H}\left[\frac{1}{2-\sqrt{\varepsilon}}\right] = \frac{\varepsilon}{8\ln 2} + O(\varepsilon^2).\end{aligned}$$

For $\varepsilon = 1$ we can say more: The decomposition is optimal in both the conditional and unconditional case, and we can explicitly determine the common information. For the unconditional case $\mathbb{C}\left[M\right] = \frac{2}{3}$ as proved in [Witsenhausen, 1976, Theorem 7].

The conditional case follows by a functional relationship between the conditional and unconditional mutual information for every seed Π:

$$\begin{aligned}\mathbb{H}\left[A,B \,|\, \Pi\right] &= \mathbb{H}\left[A \,|\, \Pi\right] + \mathbb{H}\left[B \,|\, \Pi\right] \\ &= \frac{2}{3}\mathbb{H}\left[A \,|\, \Pi, B=0\right] + \frac{2}{3}\mathbb{H}\left[B \,|\, \Pi, A=0\right] = \frac{4}{3}\mathbb{H}\left[A,B \,|\, \Pi, D=0,C\right],\end{aligned}$$

where

$$\begin{aligned}\mathbb{H}\left[A \,|\, \Pi\right] = \mathbb{H}\left[A \,|\, \Pi, B\right] &= \mathbb{P}\left[B=0\right]\mathbb{H}\left[A \,|\, \Pi, B=0\right] + \mathbb{P}\left[B=1\right]\mathbb{H}\left[A \,|\, \Pi, B=1\right] \\ &= \frac{2}{3}\mathbb{H}\left[A \,|\, \Pi, B=0\right],\end{aligned}$$

as $\mathbb{H}\left[A \,|\, \Pi, B=1\right] = 0$.

We conclude that

$$\mathbb{I}\left[A,B;\Pi\right] = \frac{4}{3}\mathbb{I}\left[A,B;\Pi \,|\, D=0,C\right] + \mathbb{H}\left[A,B\right] - \frac{4}{3}\mathbb{H}\left[A,B \,|\, D=0,C\right].$$

The entropies are easily calculated as

$$\begin{aligned}\mathbb{H}\left[A,B\right] &= \log 3, \\ \mathbb{H}\left[A,B \,|\, D=0,C\right] &= 1.\end{aligned}$$

Finally, taking infimum leads to

$$\mathbb{C}\left[M\right] = \frac{4}{3}\mathbb{C}\left[M \,|\, D=0,C\right] + \mathbb{H}\left[A,B\right] - \frac{4}{3}\mathbb{H}\left[A,B \,|\, D=0,C\right]. \tag{11}$$

In particular, $\mathbb{C}\left[M \,|\, D=0,C\right] = (6 - 3\log 3)/4 \approx 0.3113$. Thus, with the lower bound from Theorem 4.1, the claims follow for $n = 1$.

To generalize the above decomposition to all n, we take n independent copies $(A_i, B_i, \Pi_i : i \in [n])$ of the variables. This distribution is obviously induced by the matrix $M(a,b) = (1-\varepsilon)^{|a \cap b|}$ extending UDISJ, given Π the variables A and B are independent, and $\mathbb{I}\left[A,B;\Pi \,|\, D=0,C\right] = \sum_{i\in[n]} \mathbb{I}\left[A_i,B_i;\Pi_i \,|\, D_i=0,C_i\right]$. Finally, the lower bound for the case $n = 1$ and $\varepsilon = 1$ generalizes by the same argument as in Theorem 4.1. □

Equation (11) highlights the importance of conditioning: the strength of $\log \operatorname{rank}_+(M) \geq \mathbb{C}\left[M \,|\, Z\right]$ depends on the choice of the condition Z. Given the optimality of the estimation by means of common information, further improvements on the lower bound will only be possible by considering a different condition.

Observe that Proposition 4.3 also establishes that in general it will not be possible to lower bound the common information (or the entropy of Π for that matter) by means of the nonnegative rank from below. This rules out a tight characterization of the nonnegative rank of a matrix M in terms of the entropy of Π or the mutual information of Π and (A, B) whose infimum is the common information of M. At its core the gap between common information and the nonnegative rank stems from the fact that the mutual information is a continuous measure whereas the nonnegative rank is not. However, when relaxing the notion of nonnegative rank to $\operatorname{rank}_+^{(\delta)}(M) := \min \left\{\operatorname{rank}_+(M') \mid \|M - M'\|_1 \leq \delta\right\}$ where $\|.\|_1$ is the total variance, then

$$\log \operatorname{rank}_+^{(\delta)}(M) \leq O(\mathbb{C}\left[M\right] + 1)/\delta$$

as shown in [Jain et al., 2013, Corollary 1.1]. Moreover, it turns out that common information is equal to the amortized log nonnegative rank; see Braun et al. [2013b] for the quantitative statement including convergence rates and Wyner [1975] for the qualitative one.

A ρ-shift $\tilde{M}$ of a (partial) matrix M is obtained by adding ρ to each entry of the (partial) matrix. Such shifts are at the core of the study of the complexity of approximate extended formulations (see Braun et al. [2012]), for which it is more natural to write the shift in the form $\rho - 1$ instead of ρ. We obtain the following theorem, slightly improving over Braverman and Moitra [2012] in terms of the explicit constants:

Theorem 4.4 (Nonnegative rank of shifted UDISJ). *Let $M \in \mathbb{R}_+^{2^n \times 2^n}$ be a $(\rho - 1)$-shift of the unique disjointness matrix UDISJ, i.e.,*

$$M(a, b) := \begin{cases} \rho & \text{if } a \cap b = \emptyset, \\ \rho - 1 & \text{if } |a \cap b| = 1 \\ \geq 0 & \text{otherwise} \end{cases}$$

for some $\rho \geq 1$. Then $\operatorname{rank}_+(M) \geq 2^{n/8\rho}$. If $\rho = 1$, then $\operatorname{rank}_+(M) \geq 2^{\frac{6-3\log 3}{4} \cdot n} \geq 2^{0.3113 \cdot n}$.

Proof. Applying Theorem 4.1 with $\varepsilon = 1/\rho$, we obtain $\mathbb{C}\left[M \mid D = 0, C\right] = \mathbb{C}\left[M/\rho \mid D = 0, C\right] \geq n/8\rho$. Note that multiplying the matrix by a positive scalar does not change its common information. Hence by Lemma 3.4, we obtain $\operatorname{rank}_+ M \geq 2^{n/8\rho}$. The second claim follows similarly from Proposition 4.3. □

4.1 Application to (approximate) extended formulations

We immediately obtain strengthened versions of [Braun et al., 2012, Theorems 7 and 8] as proven in [Braverman and Moitra, 2012, Section 4] by plugging in the improved lower bound on the nonnegative rank of M:

Theorem 4.5 (Inapproximability of CLIQUE). *Let $\rho \geqslant 1$, let n be a positive integer and let $P = \mathrm{COR}(n)$, $Q = Q(n)$ be as in Braun et al. [2012]. Then $\mathrm{xc}(P, \rho Q) = 2^{n/8\rho}$. In particular if $\rho = n^{1-\varepsilon}$ for some constant $\varepsilon < 1$, then $\mathrm{xc}(P, \rho Q) = 2^{n^\varepsilon/8}$. Therefore for the linear encoding defined in Braun et al. [2012], every $n^{1-\varepsilon}$-approximate EF of CLIQUE has size $2^{n^\varepsilon/8}$, for all $0 < \varepsilon < 1$.*

The latter lower bound for CLIQUE matches the algorithmic inapproximability of Håstad [1999].

5 Robustness of the UDISJ matrix

We will now show that the above lower bound on the nonnegative rank of UDISJ is robust with respect to random and adversarial removal of rows and columns as well as random and adversarial change of entries in the matrix. To this end we will first formulate the case $n = 1$ of Theorem 4.1 for general distribution to incorporate noise as well as adversarial flips of bits.

Lemma 5.1 (Information from noised-up submatrices). *Let* $(A,B) \in \{0,1\}^2$ *with distribution*

	$B=0$	$B=1$
$A=0$	α	γ
$A=1$	β	δ

with $\alpha+\beta+\gamma+\delta = 1$. *Then*

$$\mathbb{H}\left[A,B \,|\, D=0,C,\Pi\right] \leq \frac{3\sqrt{\alpha \max(\beta,\gamma)} + \alpha\sqrt{\delta/\min(\beta,\gamma)}}{2\alpha+\beta+\gamma},$$

where Π *is a seed for* A,B *with condition* $D=0,C$, *and* D,C *are as in Theorem 4.1.*

Proof. We estimate

$$\begin{aligned}&\mathbb{H}\left[A,B \,|\, D=0,C,\Pi\right]\\ &\qquad= \frac{\alpha+\gamma}{2\alpha+\beta+\gamma}\mathbb{H}\left[A,B \,|\, D=0,C=\mathbb{A},\Pi\right] + \frac{\alpha+\beta}{2\alpha+\beta+\gamma}\mathbb{H}\left[A,B \,|\, D=0,C=\mathbb{B},\Pi\right]\end{aligned}$$

Now the event $D=0,C=\mathbb{A}$ is the same as $A=0,C=\mathbb{A}$. As C is independent of A,B,Π, we obtain

$$\mathbb{H}\left[A,B \,|\, D=0,C=\mathbb{A},\Pi\right] = \mathbb{H}\left[A,B \,|\, A=0,\Pi\right].$$

Let Π_{ab} denote the distribution of Π given $A=a$ and $B=b$. Lemma 2.4 applies:

$$\mathbb{H}\left[A,B \,|\, A=0,\Pi\right] \leq \frac{2\sqrt{\alpha\gamma}}{\alpha+\gamma}\cdot(1-h^2(\Pi_{00};\Pi_{01})).$$

All in all, we obtain

$$\mathbb{H}\left[A,B \,|\, D=0,C=\mathbb{A},\Pi\right] \leq \frac{2\sqrt{\alpha\gamma}}{\alpha+\gamma}\cdot(1-h^2(\Pi_{00};\Pi_{01})).$$

Similarly,

$$\mathbb{H}\left[A,B \,|\, D=0,C=\mathbb{B},\Pi\right] \leq \frac{2\sqrt{\alpha\beta}}{\alpha+\beta}\cdot(1-h^2(\Pi_{00};\Pi_{10})).$$

Thus

$$\mathbb{H}\left[A,B \,|\, D=0,C,\Pi\right] \leq \frac{2\sqrt{\alpha\max(\beta,\gamma)}}{2\alpha+\beta+\gamma}\cdot(2-h^2(\Pi_{00};\Pi_{01})-h^2(\Pi_{00};\Pi_{10})).$$

Finally we estimate the Hellinger distances:

$$\begin{aligned}h^2(\Pi_{00};\Pi_{01})+h^2(\Pi_{00};\Pi_{10}) &\geq \frac{(h(\Pi_{00};\Pi_{01})+h(\Pi_{00};\Pi_{10}))^2}{2}\\ &\geq \frac{h^2(\Pi_{01};\Pi_{10})}{2} \geq \frac{1-\sqrt{\alpha\delta/(\beta\gamma)}}{2}.\end{aligned}$$

The last inequality follows from Lemma 3.5 by the independence of A and B given Π. Combining the estimates finishes the proof:

$$\begin{aligned}\mathbb{H}\left[A,B \,|\, D=0,C,\Pi\right] &\leq \frac{2\sqrt{\alpha\max(\beta,\gamma)}}{2\alpha+\beta+\gamma}\cdot\left(2-\frac{1-\sqrt{\alpha\delta/(\beta\gamma)}}{2}\right)\\ &= \frac{3\sqrt{\alpha\max(\beta,\gamma)}+\alpha\sqrt{\delta\max(\beta,\gamma)/(\beta\gamma)}}{2\alpha+\beta+\gamma} = \frac{3\sqrt{\alpha\max(\beta,\gamma)}+\alpha\sqrt{\delta/\min(\beta,\gamma)}}{2\alpha+\beta+\gamma}.\end{aligned}$$ □

5.1 Adversarial and random removal of rows and columns

In the random setting we choose a submatrix S randomly, and we are bounding the expectation of the nonnegative rank. The rough idea is summing over minimal factorizations for submatrices to obtain a factorization of the original UDISJ matrix. Even though this factorization has a huge number of summands, these are expected to be very similar, and thus reveal little information.

The distribution of the submatrix S will not be uniform in general: the probability of a given submatrix will be proportional to the sum of its entries.

We shall denote by $x \in S$ the event that the row, column or entry x is contained in S.

Theorem 5.2. *Let M be a nonnegative matrix, and $\mathcal{S}$ be a family of submatrices of M with every entry of M contained in exactly a γ-fraction of the members. Let $S \in \mathcal{S}$ be a random submatrix with distribution $\mathbb{P}[S = s] = \sum_{(a,b)\in s} M(a,b)/\gamma\,|\mathcal{S}| \sum_{a,b} M(a,b)$. Then*

$$\mathbb{E}[\mathbb{C}[S \,|\, Z]] \geq \mathbb{C}[M \,|\, Z] + \log\gamma. \tag{12}$$

Proof. The key to the proof is to construct the right probability space for comparing $\mathbb{C}[S \,|\, Z]$ and $\mathbb{C}[M \,|\, Z]$.

Let A, B be the random row-column pair of the induced distribution of M. Given A, B let $S \in \mathcal{S}$ be chosen uniformly with the restriction $A, B \in S$:

$$\mathbb{P}[S = s \,|\, A = a, B = b] = \frac{1}{\gamma\,|\mathcal{S}|}, \quad (a,b) \in s. \tag{13}$$

This induces the same distribution on S as given in the theorem:

$$\begin{aligned}\mathbb{P}[S = s] = \sum_{(a,b)\in s} \mathbb{P}[A = a, B = b]\,\mathbb{P}[S = s \,|\, A = a, B = b] \\ = \sum_{(a,b)\in s} \frac{M(a,b)}{\sum_{x,y} M(x,y)} \frac{1}{\gamma\,|\mathcal{S}|} = \frac{\sum_{(a,b)\in s} M(a,b)}{\sum_{x,y} M(x,y) \cdot \gamma\,|\mathcal{S}|}.\end{aligned}$$

Note that given S, the distribution of A, B is the one induced by S, i.e., for $(a,b) \in s$:

$$\mathbb{P}[A = a, B = b \,|\, S = s] = \mathbb{P}[S = s \,|\, A = a, B = b] \frac{\mathbb{P}[A = a, B = b]}{\mathbb{P}[S = s]} = \frac{M(a,b)}{\sum_{(x,y)\in s} M(x,y)}.$$

This leads to the following interpretation of $\mathbb{E}[\mathbb{C}[S \,|\, Z]]$: For every $s \in \mathcal{S}$ consider the probability space conditioned on $S = s$. Introduce seeds Π_s for A and B. For every such collection $\{\Pi_s : s \in \mathcal{S}\}$, glue it together to a random variable Π, i.e., $\Pi = \Pi_s$ given $S = s$. (Here the ranges of Π_s are considered pairwise disjoint without loss of generality. In particular, Π determines S.) Then

$$\mathbb{E}[\mathbb{C}[S \,|\, Z]] = \inf_{(\Pi_s : s \in S)\text{ seeds}} \mathbb{I}[A, B; \Pi \,|\, Z, S].$$

By construction, Π is a seed for M, therefore

$$\mathbb{I}[A, B; \Pi \,|\, Z] \geq \mathbb{C}[M \,|\, Z].$$

Now by the chain rule

$$\begin{aligned}\mathbb{I}[A, B; \Pi \,|\, Z] - \mathbb{I}[A, B; \Pi \,|\, Z, S] = \mathbb{I}[A, B; S \,|\, Z] \\ = \mathbb{H}[S \,|\, Z] - \mathbb{H}[S \,|\, Z, A, B] \leq \log|\mathcal{S}| - \log\gamma\,|\mathcal{S}| = -\log\gamma,\end{aligned}$$

as S is uniform given A, B, Z (because S is both uniform and independent of Z given A, B by construction). Rearranging provides

$$\mathbb{I}[A, B; \Pi \,|\, Z, S] \geq \mathbb{I}[A, B; \Pi \,|\, Z] + \log\gamma \geq \mathbb{C}[M \,|\, Z] + \log\gamma,$$

and taking infimum over Π produces the result. □

We obtain the following corollary.

Corollary 5.3. *Let $0 < \alpha, \beta < 1$ and let S be a random $2^{(1-\alpha)n} \times 2^{(1-\beta)n}$ submatrix of UDISJ with distribution $\mathbb{P}[S = s] = c \sum_{(a,b)\in s} s(a,b)$ for a suitable $c > 0$. Then $\mathbb{E}[\mathrm{rank}_+ S] \geq 2^{(\varepsilon/8-\alpha-\beta)n}$.*

Proof. By Jensen's inequality and Theorems 4.1 and 5.2 with Z being $C, D = 0$ the conditional from Theorem 4.1

$$\mathbb{E}[\mathrm{rank}_+ S] \geq \mathbb{E}\left[2^{\mathbb{C}[S \mid Z]}\right] \geq 2^{\mathbb{E}[\mathbb{C}[S \mid Z]]} \geq 2^{(\varepsilon/8-\alpha-\beta)n}. \qquad \square$$

Remark 5.4. It is worthwhile to compare the bounds from Corollary 5.3 to the special case of forbidding A and B to contain certain elements $i \in [n]$. Say, A must not contain the first αn elements, and B must not contain the last βn elements. Provided $\alpha + \beta < 1$, this means that actually only the $(1 - \alpha - \beta)n$ elements in the middle count, hence we obtain the significantly larger lower bound $\mathrm{rank}_+ S \geq 2^{\varepsilon/8\cdot(1-\alpha-\beta)n}$. However, this is not unexpected: the removal of rows and columns by forbidding elements is rather homogeneous, whereas a random removal could potentially remove much more information.

We will now switch our attention to adversarial removal of rows and columns. The following observation is useful to understand what type of bounds we can expect.

Observation 5.5 (Existence of a large subset with no disjoint pairs). With $\alpha = \beta = 1/2$ the adversary can choose

$$S = \{A \mid A \subseteq [n], 1 \in A\} \times \{B \mid B \subseteq [n], 1 \in B\}.$$

Clearly, $|S| = 2^{2(n-1)}$ and hence $1/4$ of all pairs, however all pairs intersect, hence the partial matrix S is 0.

This is the largest submatrix with no disjoint pairs, as can be seen as follows. Let S_A be the set of rows, and S_B be the set of columns of S. We identify rows and columns with the subsets of $[n]$ indexing them. If S has no disjoint pairs, then for all $X \subseteq [n]$ it is impossible that $X \in S_A$ and $[n] \setminus X \in S_B$). Thus

$$|S_A| + |S_B| = \sum_{X \subseteq [n]} \left(I_{X \in S_A} + I_{[n] \setminus X \in S_B}\right) \leq 2^n.$$

Comparing the geometric mean and the arithmetic mean, we obtain

$$|S_A| \cdot |S_B| \leq \left(\frac{|S_A| + |S_B|}{2}\right)^2 = 2^{2n}/4 = 2^{2(n-1)}.$$

The main tool for the adversarial case is the following insight.

Theorem 5.6. *Let M be a nonnegative matrix, and S be a submatrix of M. Let Z be a condition, which is a mixture of an event Z_{event} and random variables. Furthermore, let A, B be the random row-column pair of the induced distribution of M, and I be the three-valued indicator of whether $A, B \in S$, $A \notin S$ or $A \in S$ but $B \notin S$. Then*

$$\begin{aligned} \mathbb{P}[A, B \in S \mid Z_{\mathrm{event}}]\, \mathbb{C}[S \mid Z, I] \geq \mathbb{C}[M \mid Z] - \mathbb{P}[A \notin S \mid Z_{\mathrm{event}}]\, \mathbb{H}[A \mid Z] \\ - \mathbb{P}[B \notin S \mid Z_{\mathrm{event}}]\, \mathbb{H}[B \mid Z] - \log 3. \end{aligned} \tag{14}$$

Proof. Given $A, B \in S$ and Z, we choose an arbitrary seed Π_S of A, B for S. We define a seed Π_M for M via

$$\Pi_M := \begin{cases} \Pi_S & \text{if } A, B \in S, \\ A & \text{if } A \notin S, \\ B & \text{if } A \in S \text{ but } B \notin S \end{cases}$$

with the values of Π_M being pairwise distinct in the three cases. It follows

$$\begin{aligned}\mathbb{C}\left[M \mid Z\right] \leq \mathbb{I}\left[A,B;\Pi_M \mid Z\right] &= \mathbb{I}\left[A,B;\Pi_M \mid Z,I\right] + \mathbb{I}\left[A,B;I \mid Z\right] \\ &\leq \mathbb{P}\left[A,B \in S \mid Z_{\text{event}}\right] \mathbb{I}\left[A,B;\Pi_S \mid Z\right] + \mathbb{P}\left[A \notin S \mid Z_{\text{event}}\right] \mathbb{I}\left[A,B;A \mid Z\right] \\ &\quad + \mathbb{P}\left[A \in S, B \notin S \mid Z_{\text{event}}\right] \mathbb{I}\left[A,B;B \mid Z\right] + \mathbb{I}\left[A,B;I \mid Z\right] \\ &\leq \mathbb{P}\left[A,B \in S \mid Z_{\text{event}}\right] \mathbb{I}\left[A,B;\Pi_S \mid Z\right] + \mathbb{P}\left[A \notin S \mid Z_{\text{event}}\right] \mathbb{H}\left[A \mid Z\right] \\ &\quad + \mathbb{P}\left[B \notin S \mid Z_{\text{event}}\right] \mathbb{H}\left[B \mid Z\right] + \log 3\end{aligned}$$

Taking the infimum over Π_S, the result follows. □

We will now show that when the adversary is restricted as to only remove up to an α-fraction per *each potential size* of subsets, then the resulting matrix has still high nonnegative rank.

Corollary 5.7 (Homogeneous, adversarial removal of rows and columns). *Let $0 \leq \alpha, \beta < 1$. Let S be any submatrix of UDISJ M obtained as follows. For every $0 \leq k \leq n$, we select the rows and columns indexed by subsets of size k, and delete at most an α-fraction of these rows and a β-fraction of these columns. Then*

$$\operatorname{rank}_+ S \geq 2^{\left(\frac{1}{8\rho} - (\alpha+\beta)\right)n - \log 3}$$

Proof. We use the variables C, D from Theorem 4.1. Note that because of symmetry, the marginal distributions of A and B are uniform even given $D = 0$ when the size of the sets are fixed, hence by assumption

$$\begin{aligned}\mathbb{P}\left[A \notin S \mid D = 0, |A|\right] &\leq \alpha, \\ \mathbb{P}\left[B \notin S \mid D = 0, |B|\right] &\leq \beta.\end{aligned}$$

Taking expectation it follows

$$\begin{aligned}\mathbb{P}\left[A \notin S \mid D = 0\right] &\leq \alpha, \\ \mathbb{P}\left[B \notin S \mid D = 0\right] &\leq \beta.\end{aligned}$$

Applying Theorem 5.6 and Theorem 4.1:

$$\begin{aligned}\mathbb{C}\left[S \mid D = 0, C\right] &\geq \mathbb{P}\left[A,B \in S \mid D = 0\right] \mathbb{C}\left[S \mid D = 0, C\right] \\ &\geq \mathbb{C}\left[M \mid D = 0, C\right] - \alpha \mathbb{H}\left[A \mid D = 0, C\right] - \beta \mathbb{H}\left[B \mid D = 0, C\right] - \log 3 \\ &= \mathbb{C}\left[M \mid D = 0, C\right] - (\alpha + \beta)n \geq \left(\frac{1}{8\rho} - (\alpha + \beta)\right) n - \log 3.\end{aligned}$$

Now the estimation on the nonnegative rank is immediate. □

The following lemma establishes that even restricting to subsets of fixed size close to $n/4$, the common information of UDISJ does not decrease significantly. This construction significantly improves over the simple trick of splitting $[n]$ into 3 disjoint sets, arguing on the first set via the unrestricted argument, and use the other two for padding to ensure a fixed size.

Lemma 5.8 (Restriction to fixed size subsets). *For the UDISJ matrix M, let M_k be the submatrix for sets of size k. Then for $0 < \varepsilon \leq 1$*

$$\mathbb{C}\left[M_k \mid A \cap B = \emptyset\right] \geq \frac{n}{8\rho} - O(n^{1-\varepsilon}) \qquad \textit{for } k = n/4 + O(n^{1-\varepsilon}).$$

Proof. The proof is similar to the one in Theorem 4.1, however now we have to account for loss of common information due to dependence of the pairs A_i, B_i.

First we replace the condition $A \cap B = \emptyset$ with $D = 0, C$ with C, D from Theorem 4.1. Any seed Π for A, B given $A \cap B = \emptyset$ can be introduced to be independent of C given A, B. Therefore it will also be a seed given $D = 0, C$. By symmetry, given either $D = 0$ or $A \cap B = \emptyset$, the distribution of A, B will be uniform. In particular, as these conditions are independent of Π given A, B, the variables A, B, Π have the same joint distribution given either $D = 0$ or $A \cap B = \emptyset$. Hence (recall that C is part of the probability space)

$$\mathbb{I}\left[A, B; \Pi \middle| A \cap B = \emptyset\right] = \mathbb{I}\left[A, B; \Pi \middle| D = 0\right] \geq \mathbb{I}\left[A, B; \Pi \middle| D = 0, C\right],$$

where the inequality follows from the independence of C and Π given A, B.

To estimate the loss due to dependence of the A_i, B_i observe that

$$\mathbb{H}\left[A, B \middle| D = 0, C, \Pi\right] \leq \sum_{i \in [n]} \mathbb{H}\left[A_i, B_i \middle| D = 0, C, \Pi\right].$$

Combining it with $\mathbb{H}\left[A, B \middle| D = 0, C, \Pi\right] = \mathbb{H}\left[A, B \middle| D = 0, C\right] - \mathbb{I}\left[A, B; \Pi \middle| D = 0, C\right]$ we obtain

$$\mathbb{I}\left[A, B; \Pi \middle| D = 0, C\right] \geq \mathbb{H}\left[A, B \middle| D = 0, C\right] - \sum_{i \in [n]} \mathbb{H}\left[A_i, B_i \middle| D = 0, C, \Pi\right]. \tag{15}$$

It therefore suffices to estimate both terms separately. First, we estimate $\mathbb{H}\left[A, B \middle| D = 0, C\right]$. Note that A, B, C is uniformly distributed given $D = 0$, with $n! / k!^2\,(n - 2k)!$ possible ways of choosing A, B (two disjoint subsets of size k), and for each A, B there are 2^{n-2k} choices for C.

$$\begin{aligned}\mathbb{H}\left[A, B \middle| D = 0, C\right] = \mathbb{H}\left[A, B, C \middle| D = 0\right] - \mathbb{H}\left[C \middle| D = 0\right] &\geq \log \frac{n!}{k!^2\,(n - 2k)!} 2^{n-2k} - n \\ &= \log \frac{n!}{k!^2\,(n - 2k)!} - 2k = n\mathbb{H}\left[2k/n\right] + O(\log n) = n + O(\log n).\end{aligned}$$

Second, we estimate $\mathbb{H}\left[A_i, B_i \middle| D = 0, C, \Pi\right]$. Given $D_j = 0$ for all $j \neq i$, which we denote by $\mathcal{D}_j$ in the following, the distribution of A_i, B_i is the following with a normalizing constant K:

$$\begin{aligned}&\mathbb{P}\left[A_i = x, B_i = y \middle| \mathcal{D}_j\right] \\ &\qquad = K \frac{(n-1)!}{(k-x)!\,(k-y)!\,(n-1-2k+x+y)!} \frac{1}{2^{2k-x-y}} \cdot \begin{cases} \rho - 1 & \text{if } (x, y) \neq (1, 1), \\ \rho & \text{if } (x, y) = (1, 1), \end{cases}\end{aligned}$$

where $\frac{(n-1)!}{(k-x)!(k-y)!(n-1-2k+x+y)!}$ is the number of values of A and B with $A_i = x$ and $B_i = y$, and $\frac{1}{2^{2k-x-y}}$ is the probability of $D_j = 0$ for all $j \neq i$ given such an A, B.

Separating common factors for a better overview:

$$\begin{aligned}&\mathbb{P}\left[A_i = x, B_i = y \middle| \mathcal{D}_j\right] \\ &\qquad = \frac{K(n-1)!}{k!^2\,(n-1-2k)!\,2^{2k}} \cdot \begin{cases} \rho, & \text{if } (x, y) = (0, 0), \\ \rho \frac{2k}{n-2k}, & \text{if } (x, y) = (1, 0) \text{ or } (x, y) = (0, 1), \\ (\rho - 1) \frac{(2k)^2}{(n-2k)(n-2k+1)}, & \text{if } (x, y) = (1, 1). \end{cases}\end{aligned}$$

The assumption $k = n/4 + O(n^{1-\varepsilon})$ provides $2k/(n-2k) = 1 + O(n^{-\varepsilon})$ and $(2k)^2/(n-2k)(n-2k+1) = 1 + O(n^{-\varepsilon})$. Thus the probabilities are (with $\tilde{K}$ a common factor)

$$\begin{aligned}\alpha &= \mathbb{P}\left[A_i = 0, B_i = 0 \middle| \mathcal{D}_j\right] = \tilde{K}\rho, \\ \beta &= \mathbb{P}\left[A_i = 1, B_i = 0 \middle| \mathcal{D}_j\right] = \tilde{K}\rho(1 - O(n^{-\varepsilon})), \\ \gamma &= \mathbb{P}\left[A_i = 0, B_i = 0 \middle| \mathcal{D}_j\right] = \tilde{K}\rho(1 - O(n^{-\varepsilon})), \\ \delta &= \mathbb{P}\left[A_i = 1, B_i = 1 \middle| \mathcal{D}_j\right] = \tilde{K}(\rho - 1)(1 - O(n^{-\varepsilon})).\end{aligned}$$

Lemma 5.1 applies to the conditional distribution $D_j = 0$ for all $j \neq i$ with Π replaced by $(\Pi, C_j : j \neq i)$, and adding a subscript i to the other variables A, B, C, D in the lemma:

$$\begin{aligned}
\mathbb{H}\left[A_i, B_i \middle| D = 0, C, \Pi\right] &\leq \frac{3\sqrt{\alpha \max(\beta,\gamma)} + \alpha\sqrt{\delta / \min(\beta,\gamma)}}{2\alpha + \beta + \gamma} \\
&= \frac{3\rho(1 - O(n^{-\varepsilon})) + \rho(1 - O(n^{-\varepsilon}))\sqrt{\frac{(\rho-1)(1-O(n^{-\varepsilon}))}{\rho(1-O(n^{-\varepsilon}))}}}{4\rho(1 - O(n^{-\varepsilon}))} \\
&\leq \frac{3\rho(1 - O(n^{-\varepsilon})) + \rho(1 - O(n^{-\varepsilon}))\left(1 - \frac{1+O(\rho n^{-\varepsilon})}{2\rho(1-O(n^{-\varepsilon}))}\right)}{4\rho(1 - O(n^{-\varepsilon}))} \\
&= 1 - \frac{1}{8\rho} + O(n^{-\varepsilon})
\end{aligned} \tag{16}$$

with the constant factor in the error term independent of i.

Finally, combining the estimates we obtain

$$\begin{aligned}
\mathbb{I}\left[A, B; \Pi \middle| A \cap B = \emptyset\right] &\geq \mathbb{H}\left[A, B \middle| D = 0, C\right] - \sum_{i \in [n]} \mathbb{H}\left[A_i, B_i \middle| D = 0, C, \Pi\right] \\
&\geq n + O(\log n) - n\left(1 - \frac{1}{8\rho} + O(n^{-\varepsilon})\right) = \frac{n}{8\rho} - O(n^{1-\varepsilon}). \qquad \square
\end{aligned}$$

We will now briefly show that the estimation in (15) is really the same as in Theorem 4.1, however accounting for the loss of independence.

Remark 5.9 (Estimating entropy instead of mutual information). In order to establish the link to the estimation in Theorem 4.1, observe that

$$\begin{aligned}
\mathbb{H}\left[A, B \middle| D = 0, C, M, \Pi\right] &\leq \sum_{i \in [n]} \mathbb{H}\left[A_i, B_i \middle| D = 0, C, M, \Pi\right] \\
\mathbb{H}\left[A, B \middle| D = 0, C, M\right] - \mathbb{I}\left[A, B; \Pi \middle| D = 0, C, M\right] &\leq \sum_{i \in [n]} \mathbb{H}\left[A_i, B_i \middle| D = 0, C, M\right] \\
&\quad - \mathbb{I}\left[A_i, B_i; \Pi \middle| D = 0, C, M\right]
\end{aligned}$$

and hence

$$\begin{aligned}
\mathbb{I}\left[A, B; \Pi \middle| D = 0, C, M\right] \geq{}& \sum_{i \in [n]} \mathbb{I}\left[A_i, B_i; \Pi \middle| D = 0, C, M\right] \\
&+ \mathbb{H}\left[A, B \middle| D = 0, C, M\right] - \sum_{i \in [n]} \mathbb{H}\left[A_i, B_i \middle| D = 0, C, M\right].
\end{aligned}$$

Finally, we combine adversarial removal and fixed size subsets:

Corollary 5.10. *For the UDISJ matrix M, let M_k be the submatrix for sets of size k. Let S be any submatrix of M_k obtained by deleting at most an α-fraction of rows and at most a β-fraction of columns for some $0 \leq \alpha, \beta < 1$. Then for $0 < \varepsilon \leq 1$*

$$\operatorname{rank}_+ S \geq 2^{(1/8\rho - (\alpha+\beta)\mathbb{H}[1/4])n - O(n^{1-\varepsilon})} \qquad \textit{for } k = n/4 + O(n^{1-\varepsilon}).$$

Proof. Note that because of symmetry, the marginal distributions of A and B are uniform given $A \cap B = \emptyset$, hence

$$\begin{aligned}
\mathbb{P}\left[A \notin S \middle| A \cap B = \emptyset\right] &= \alpha, \\
\mathbb{P}\left[B \notin S \middle| A \cap B = \emptyset\right] &= \beta.
\end{aligned}$$

Applying Theorem 5.6 and Lemma 5.8:

$$\begin{aligned}\mathbb{C}\left[S \mid A \cap B = \emptyset\right] &\geq \mathbb{P}\left[A, B \in S \mid A \cap B = \emptyset\right] \mathbb{C}\left[S \mid A \cap B = \emptyset\right] \\ &\geq \mathbb{C}\left[M \mid A \cap B = \emptyset\right] - \alpha \mathbb{H}\left[A \mid A \cap B = \emptyset\right] - \beta \mathbb{H}\left[B \mid A \cap B = \emptyset\right] - \log 3 \\ &= \mathbb{C}\left[M \mid A \cap B = \emptyset\right] - (\alpha + \beta) \log \binom{n}{k} \geq \frac{n}{8\rho} - (\alpha + \beta) n \mathbb{H}\left[1/4\right] - O(n^{1-\varepsilon}). \square\end{aligned}$$

5.2 Adversarial and random flipping of bits

We will now analyze the behavior of the nonnegative rank of the UDISJ pattern provided we allow for changing a large fraction of entries. Using Lemma 5.1 we can establish the following lower bounds:

Theorem 5.11 (Random flipping of bits). *Let $0 \leq \tau < 1/2$ and $\rho \geq 1$ be parameters. Let $M \in \mathbb{R}_+^{2^n \times 2^n}$ be the following random matrix*

$$M(a, b) := \begin{cases} \rho - u_{ab} & \text{if } a \cap b = \emptyset, \\ \rho - 1 + u_{ab} & \text{if } |a \cap b| = 1 \\ \geq 0 & \text{otherwise} \end{cases}$$

with $u_{ab} \in \{0, 1\}$ mutually independent random variables with $\mathbb{P}\left[u_{ab} = 1\right] = \tau$ for all a, b. Then

$$\operatorname{rank}_+(M) \geq \left(1 - o\left(\frac{\sqrt{\tau}}{\rho - \tau}\right)\right) \frac{\rho - \tau}{\rho} 2^{n(1-2\tau)/8(\rho-\tau)}$$

with high probability.

Proof. The proof is similar to Lemma 5.8, but now we have to account for the noise in the matrix, too. We fix M, implicitly conditioning on it, and start with (15), which can be proved similarly:

$$\mathbb{I}\left[A, B; \Pi \mid D = 0, C\right] \geq \mathbb{H}\left[A, B \mid D = 0, C\right] - \sum_{i \in [n]} \mathbb{H}\left[A_i, B_i \mid D = 0, C, \Pi\right].$$

We again estimate both terms separately. We start with $\mathbb{H}\left[A, B \mid D = 0, C\right]$.

Let $D(a, b, c)$ be the value of D when $A = a$, $B = b$ and $C = c$. For a given C let k be the number of pairs (a, b) with $D(a, b, C) = 0$ and $u_{ab} = 1$ (i.e., $M(a, b) = \rho - 1$). This appears in probabilities involving $D = 0$:

$$\begin{aligned}\mathbb{P}\left[D = 0 \mid C\right] &= \frac{2^n \rho - k}{\sum_{x,y} M(x, y)}, \\ \mathbb{P}\left[D = 0\right] &= \mathbb{E}\left[\frac{2^n \rho - k}{\sum_{x,y} M(x, y)}\right] = \frac{2^n \rho - \mathbb{E}\left[k\right]}{\sum_{x,y} M(x, y)}, \\ \mathbb{P}\left[A = a, B = b, C = c \mid D = 0\right] &= \frac{\rho - u_{ab}}{2^n (2^n \rho - \mathbb{E}\left[k\right])}, \qquad \text{provided } D(a, b, c) = 0.\end{aligned}$$

Estimating the entropy via the largest probability in the distribution:

$$\mathbb{H}\left[A, B, C \mid D = 0\right] \geq \log \frac{2^n (2^n \rho - \mathbb{E}\left[k\right])}{\rho}.$$

Therefore

$$\mathbb{H}\left[A, B \mid D = 0, C\right] = \mathbb{H}\left[A, B, C \mid D = 0\right] - \underbrace{\mathbb{H}\left[C \mid D = 0\right]}_{\leq n} \geq \log \frac{2^n \rho - \mathbb{E}\left[k\right]}{\rho}. \tag{17}$$

Now we establish the concentration of $\mathbb{E}[k]$ with high probability depending on M. Observe that k is the sum of the $u_{a,b}$ for all disjoint pairs of subsets a, b of $[n]$, and for a pair (a, b) there are $2^{|[n]\setminus(a\cup b)|}$ ways of choosing C to have $D = 0$, hence

$$\mathbb{E}[k] = \sum_{\substack{a,b\subseteq[n]\\ a\cap b=\emptyset}} \frac{2^{|[n]\setminus(a\cup b)|}}{2^n} u_{ab} = \sum_{\ell=0}^{n} 2^{-\ell} X_\ell, \qquad \text{where} \quad X_\ell := \sum_{\substack{|a|+|b|=\ell\\ a\cap b=\emptyset}} u_{ab}. \tag{18}$$

By Chernoff's bound,

$$\mathbb{P}\left[|X_\ell - \tau N_\ell| > \sqrt{\tau} N_\ell^{3/4}\right] \leq 2\exp\left(-N_\ell^{1/2}\right),$$

where $N_\ell := \binom{n}{\ell}2^\ell$ is the number of disjoint pairs a, b with $|a| + |b| = \ell$.

Restricting to $|n/2 - \ell| \leq n^{3/4}$, clearly $N_\ell \geq 2^{n/2 - n^{3/4}}$, and

$$\mathbb{P}\left[\exists \ell : |n/2-\ell| \leq n^{3/4}, |X_\ell - \tau N_\ell| \leq \sqrt{\tau} N_\ell^{3/4}\right] \leq 4n^{3/4}\exp(-2^{n/4-n^{3/4}/2}).$$

Outside the range $|n/2 - \ell| \leq n^{3/4}$ we have, again by Chernoff's bound

$$\sum_{\ell:|n/2-\ell|>n^{3/4}} 2^{-\ell} X_\ell \leq \sum_{\ell:|n/2-\ell|>n^{3/4}} 2^{-\ell} N_\ell = \sum_{\ell:|n/2-\ell|>n^{3/4}} \binom{n}{\ell} \leq 2^{n+1}\exp(-n^{1/2}).$$

Hence with probability $1 - \exp(-\Omega(n))$

$$\begin{aligned}
\sum_{\ell=n/2-n^{3/4}}^{n/2+n^{3/4}} 2^{-\ell} X_\ell &= \sum_{\ell=n/2-n^{3/4}}^{n/2+n^{3/4}} 2^{-\ell}\left(\tau N_\ell + \sqrt{\tau} O(N_\ell^{3/4})\right) \\
&= \sum_{\ell=n/2-n^{3/4}}^{n/2+n^{3/4}} \binom{n}{\ell}\left(\tau + \sqrt{\tau} O(2^{-1/4\ell})\right) \\
&= 2^n(1 - O(\exp(-n^{1/2})))\tau + \sqrt{\tau} O((1+2^{-1/4})^n).
\end{aligned}$$

Therefore

$$\mathbb{E}[k] = \sum_{\ell=n/2-n^{3/4}}^{n/2+n^{3/4}} 2^{-\ell} X_\ell + O(2^n \exp(-n^{1/2})) = 2^n(\tau - o(\sqrt{\tau})). \tag{19}$$

We can now estimate the entropy of A, B:

$$\mathbb{H}[A, B \mid D = 0, C] \geq \log \frac{2^n\rho - 2^n(\tau - o(\sqrt{\tau}))}{\rho} = n + \log\frac{\rho - \tau}{\rho} - o\left(\frac{\sqrt{\tau}}{\rho}\right). \tag{20}$$

Finally we estimate $\mathbb{H}[A_i, B_i \mid D = 0, C, \Pi]$. Given $D_j = 0$ for all $j \neq i$, the distribution of A_i, B_i can be written, with a normalizing K, as:

$$\mathbb{P}\left[A_i = x, B_i = y \,\middle|\, \forall j \neq i : D_j = 0\right] = \begin{cases} (\rho - X_{xy}^{(i)}/2^{n-1})/K, & \text{if } (x,y) \neq (1,1) \\ (\rho - 1 + X_{xy}^{(i)}/2^{n-1})/K, & \text{if } (x,y) = (1,1), \end{cases} \tag{21}$$

where

$$X_{xy}^{(i)} := \sum_{\substack{a,b\subseteq[n]:a_i=x,b_i=y,\\ a\cap b\setminus\{i\}=\emptyset}} 2^{-|a\setminus\{i\}|-|b\setminus\{i\}|} u_{ab}.$$

This expression is $\mathbb{E}[k]$ for the submatrix defined by $a_i = x$ and $b_i = y$, which is a version of M for an $n - 1$-element set. Hence similar to (19), with probability $1 - \exp(-\Omega(n))$

$$X_{xy}^{(i)} = 2^{n-1}(\tau - o(\sqrt{\tau})). \tag{22}$$

Under these circumstances we now apply Lemma 5.1 to the conditional distribution $D_j = 0$ for all $j \neq i$ with Π replaced by $(\Pi, C_j : j \neq i)$, and adding a subscript i to the other variables A, B, C, D in the lemma:

$$
\begin{aligned}
\mathbb{H}\left[A_i, B_i \,|\, D = 0, C, \Pi\right] &\leq \frac{3\sqrt{\alpha \max(\beta, \gamma)} + \alpha\sqrt{\delta / \min(\beta, \gamma)}}{2\alpha + \beta + \gamma} \\
&= \frac{3(\rho - \tau - o(\sqrt{\tau})) + (\rho - \tau - o(\sqrt{\tau}))\sqrt{\frac{\rho - 1 + \tau + o(\sqrt{\tau})}{\rho - \tau - o(\sqrt{\tau})}}}{4(\rho - \tau - o(\sqrt{\tau}))} \\
&\leq \frac{3(\rho - \tau - o(\sqrt{\tau})) + (\rho - \tau - o(\sqrt{\tau}))\left(1 - \frac{1 - 2\tau + o(\sqrt{\tau})}{2(\rho - \tau - o(\sqrt{\tau}))}\right)}{4(\rho - \tau - o(\sqrt{\tau}))} \\
&= 1 - \frac{1 - 2\tau}{8(\rho - \tau)} + o\left(\frac{\sqrt{\tau}}{\rho - \tau}\right).
\end{aligned}
\tag{23}
$$

All in all, with probability $1 - \exp(-\Omega(n))$ both (20) and (23) hold, and the latter for all i. Actually, the arguments above show the error term $o(\sqrt{\tau}/(\rho - \tau))$ in (23) is exponentially small in n independent of i, which we will use in the final statement. Putting everything together, we obtain

$$
\begin{aligned}
\mathbb{I}\left[A, B; \Pi \,|\, D = 0, C\right] &\geq \mathbb{H}\left[A, B \,|\, D = 0, C\right] - \sum_{i \in [n]} \mathbb{H}\left[A_i, B_i \,|\, D = 0, C, \Pi\right] \\
&\geq n + \log \frac{\rho - \tau}{\rho} - o\left(\frac{\sqrt{\tau}}{\rho - \tau}\right) - n\left(1 - \frac{1 - 2\tau}{8(\rho - \tau)} + o\left(\frac{\sqrt{\tau}}{\rho - \tau}\right)\right) \\
&= \frac{n(1 - 2\tau)}{8(\rho - \tau)} + \log \frac{\rho - \tau}{\rho} - o\left(\frac{\sqrt{\tau}}{\rho - \tau}\right)
\end{aligned}
$$

□

Remark 5.12 (Limits of Theorem 5.11). Observe that the matrix in Theorem 5.11 contains a $2^{n/2} \times 2^{n/2}$-submatrix of disjoint strings (supporting the rows on $[n/2]$ and the columns on $[n] \setminus [n/2]$). If the noise is large enough then these entries are very close to be chosen fully at random, so that in this case the rank is roughly $2^{n/2}$ with high probability; hence so is the nonnegative rank then. Therefore the bound in Theorem 5.11 is only meaningful for smaller levels of noise.

In view of Remark 5.12 we now turn to the case of adversarial flips of entries. Using a similar technique as in Theorem 5.11 we establish:

Theorem 5.13 (Adversarial flipping of bits). *Let* $0 \leq \tau \leq 1/10$ *and* $\rho \geq 1$*. Furthermore, let* $M \in \mathbb{R}_+^{2^n \times 2^n}$ *be a matrix*

$$
M(a, b) := \begin{cases} \rho - u_{ab} & \text{if } a \cap b = \emptyset, \\ \rho - 1 + u_{ab} & \text{if } |a \cap b| = 1 \\ \geq 0 & \text{otherwise} \end{cases}
$$

with $u_{ab} \in \{0, 1\}$ *for all* a, b *so that at most a* τ*-fraction of the* $u_{a,b}$ *are* 1 *for the families*

$$
\begin{gathered}
\left\{a, b \,|\, |a| + |b| = \ell, a \cap b = \emptyset, a_i = x, b_i = y\right\}, \\
\left\{a, b \,|\, |a| + |b| = \ell, a \cap b = \{i\}\right\},
\end{gathered}
$$

for all $0 \leq \ell \leq n$, $1 \leq i \leq n$ *and* $x, y \in \{0, 1\}$*. Then* $\operatorname{rank}_+(M) \geq \frac{\rho - \tau}{\rho} 2^{n\rho(1 - 10\tau)/8(\rho - \tau)^2}$.

Proof. The proof is similar to Theorem 5.11, hence we include only the differences here. First, in (18) we have $X_\ell \leq \tau N_\ell$, hence

$$
\mathbb{E}\left[k\right] \leq \sum_{\ell = 0}^{n} \tau N_\ell = \tau 2^n. \tag{24}
$$

Therefore via (17),

$$\mathbb{H}\left[A,B\,|\,D=0,C\right] \geq \log\frac{2^n\rho-\mathbb{E}\left[k\right]}{\rho} \geq n+\log\frac{\rho-\tau}{\rho}.$$

Formula (21) still provides the distribution of A_i, B_i given $D_j = 0$ for all $j \neq i$. Similarly to (24), $X^{(i)}_{xy} \leq \tau 2^{n-1}$ follows from the assumptions. Instead of (23), we obtain

$$\begin{aligned}\mathbb{H}\left[A_i,B_i\,|\,D=0,C,\Pi\right] &\leq \frac{3\sqrt{\alpha\max(\beta,\gamma)}+\alpha\sqrt{\delta/\min(\beta,\gamma)}}{2\alpha+\beta+\gamma}\\ &\leq \frac{3\rho+\rho\sqrt{(\rho-1+\tau)/(\rho-\tau)}}{4(\rho-\tau)}\\ &\leq \frac{3\rho+\rho\left(1-\frac{1-2\tau}{\rho-\tau}\right)}{4(\rho-\tau)}\\ &= \frac{\rho}{\rho-\tau}\left(1-\frac{1-2\tau}{8(\rho-\tau)}\right).\end{aligned} \tag{25}$$

All in all, we obtain

$$\begin{aligned}\mathbb{I}\left[A,B;\Pi\,|\,D=0,C\right] &\geq \mathbb{H}\left[A,B\,|\,D=0,C\right]-\sum_{i\in[n]}\mathbb{H}\left[A_i,B_i\,|\,D=0,C,\Pi\right]\\ &\geq n+\log\frac{\rho-\tau}{\rho}-n\frac{\rho}{\rho-\tau}\left(1-\frac{1-2\tau}{8(\rho-\tau)}\right)\\ &= n\frac{\rho(1-10\tau)+8\tau^2}{8(\rho-\tau)^2}+\log\frac{\rho-\tau}{\rho}.\end{aligned}$$

□

5.3 Application to (approximate) extended formulations

We will now prove that the approximate extension complexity of the pair P, Q from Subsection 4.1 remains high even if vertices of P and facets of Q are removed. Recall that the vertices of P are considered as possible cliques, and the facets of Q as stable sets.

Let P_k be the convex hull of vertices of P corresponding to subsets of size k. Similarly, let Q_k be the polyhedra defined by facets of Q corresponding to subsets of size k. Restricting to fixed size cliques and subgraphs, Lemma 5.8 readily provides

Corollary 5.14. *Let* $0 < \varepsilon \leq 1$. *Then*

$$\log \operatorname{xc}(P_k, \rho Q_k) \geq \frac{n}{8\rho} - O(n^{1-\varepsilon}) \qquad \textit{for } k = n/4 + O(n^{1-\varepsilon}).$$

Let P^α be the polytope obtained from P by removing at most an α-fraction of the vertices corresponding to cliques of size k for every $0 \leq k \leq n$. Similarly, let Q^β be the polyhedron obtained from Q by removing at most a β-fraction of the facets corresponding to subgraphs of size k for every k. For adversarial removal, we apply Corollary 5.7.

Corollary 5.15. *Let* $0 < \alpha, \beta < 1$. *Then*

$$\log \operatorname{xc}(P^\alpha, \rho Q^\beta) \geq \left(\frac{1}{8\rho} - (\alpha+\beta)\right) n - \log 3.$$

Let P^α_k be the polytope obtained from P_k by removing at most an α-fraction of the vertices. Similarly, let Q^β_k be the polyhedron obtained from Q_k by removing at most a β-fraction of the facets. Once more, we combine adversarial removal with fixed size objects as in Corollary 5.10.

Corollary 5.16. *Let* $0 < \alpha, \beta < 1$ *and* $0 < \varepsilon \leq 1$. *Then*

$$\log \operatorname{xc}(P^\alpha_k, \rho Q^\beta_k) \geq \left(\frac{1}{8\rho} - (\alpha+\beta)\mathbb{H}\left[1/4\right]\right) n - O(n^{1-\varepsilon}) \qquad \textit{for } k = n/4 + O(n^{1-\varepsilon}).$$

6 Approximate fooling sets

If M is a nonnegative matrix, a *fooling set* $\mathcal{F}$ for M is a set of row-column indices so that $M(a,b) \neq 0$ for all $(a,b) \in \mathcal{F}$ and for distinct $(a_1,b_1),(a_2,b_2) \in \mathcal{F}$ either $M(a_1,b_2) = 0$ or $M(a_2,b_1) = 0$. Using Lemmas 3.4 and 3.5 we obtain a strengthened version of the fooling set method by relaxing the above condition.

Corollary 6.1 (Information theoretic fooling set method)**.** *Let M be a nonnegative matrix and let $\mathcal{F} = \{(a_i,b_i) \mid i \in [\ell]\}$ be a set such that $M(a,b) \neq 0$ for all $(a,b) \in \mathcal{F}$. Then*

$$\log \operatorname{rank}_+(M) \geq \gamma(\mathcal{F},M) := \log|\mathcal{F}| - \frac{1}{|\mathcal{F}|} \sum_{\substack{(a_1,b_1),(a_2,b_2)\in\mathcal{F} \\ (a_1,b_1)\neq(a_2,b_2)}} \sqrt{\frac{M(a_1,b_2)M(a_2,b_1)}{M(a_1,b_1)M(a_2,b_2)}},$$

where $\gamma(\mathcal{F},M)$ is the information bound *of M from $\mathcal{F}$. The function $\gamma(\mathcal{F},\cdot)$ is continuous in M.*

Proof. The continuity of $\gamma(\mathcal{F},\cdot)$ is clear, so we only prove the lower bound.

Let A,B be the random row-column pair in the induced distribution of M. Let Z be a uniform random variable taking values in $\mathcal{F}$, and Π_0 a seed with range size of the nonnegative rank of M, which exists by Lemma 3.4. We define a random variable Π by setting its conditional distribution given Z, namely, $(\Pi|Z = (a,b)) := (\Pi_0|A = a, B = b)$. Thus Π may differ from Π_0, nevertheless Π inherits (7) of Lemma 3.5 from Π_0. Together with Lemma 2.5 applied to Z and Π:

$$\begin{aligned}\log \operatorname{rank}_+(M) \geq \mathbb{I}\left[Z;\Pi\right] &\geq \log|\mathcal{F}| - \frac{1}{|\mathcal{F}|} \sum_{\substack{z_1,z_2\in\operatorname{range}(Z) \\ z_1\neq z_2}} \left(1 - h^2(\Pi|Z = z_1;\Pi|Z = z_2)\right) \\ &\geq \log|\mathcal{F}| - \frac{1}{|\mathcal{F}|} \sum_{\substack{(a_1,b_1),(a_2,b_2)\in\mathcal{F} \\ (a_1,b_1)\neq(a_2,b_2)}} \sqrt{\frac{M(a_1,b_2)M(a_2,b_1)}{M(a_1,b_1)M(a_2,b_2)}}.\end{aligned}$$ □

Remark 6.2. Observe that $\sqrt{\frac{M(a_1,b_2)M(a_2,b_1)}{M(a_1,b_1)M(a_2,b_2)}}$ measures the deviation of being rank-1 for the 2×2 submatrix formed by a_1,a_2,b_1,b_2. In particular, it is 1 if and only if the submatrix is rank-1.

A lower bound similar to Corollary 6.1 can also be obtained with a trace-based method Gillis et al. [2013].

6.1 Application to (approximate) extended formulations

We can immediately strengthen known fooling set results in terms of inapproximability. We can typically do better than the following corollary by taking the actual values in a slack matrix.

Corollary 6.3 (Weak inapproximability from fooling sets)**.** *Let M be a nonnegative matrix and let $\mathcal{F} = \{(a_i,b_i) \mid i \in [\ell]\}$ be a fooling set for M. We define $\delta_- := \min_{(a,b)\in\mathcal{F}} M(a,b)$ and $\delta_+ := \max_{(a_1,b_1),(a_2,b_2)\in\mathcal{F}} M(a_1,b_2)$. Let us assume $\delta_- \leq 2\delta_+$. Then for any $(\rho-1)$-shift $\tilde{M}$ of M with*

$$\rho - 1 \leq \frac{\delta_-^2}{\delta_+} \cdot \left(\frac{\log\left(|\mathcal{F}|/\alpha\right)}{|\mathcal{F}| - 1}\right)^2$$

and $0 < \alpha < |\mathcal{F}|$ *we have* $\operatorname{rank}_+(\tilde{M}) \geq \alpha$.

Proof. As $\tilde{M}$ is a $(\rho - 1)$-shift, we have $\tilde{M}(a, b) = M(a, b) + (\rho - 1)$ for all a, b. We obtain with Corollary 6.1,

$$\begin{aligned}\log \operatorname{rank}_+(\tilde{M}) &\geq \log |\mathcal{F}| - \frac{1}{|\mathcal{F}|} \sum_{\substack{(a_1,b_1),(a_2,b_2)\in\mathcal{F} \\ (a_1,b_1)\neq(a_2,b_2)}} \sqrt{\frac{\tilde{M}(a_1, b_2)\tilde{M}(a_2, b_1)}{\tilde{M}(a_1, b_1)\tilde{M}(a_2, b_2)}} \\ &= \log |\mathcal{F}| - \frac{1}{|\mathcal{F}|} \sum_{\substack{(a_1,b_1),(a_2,b_2)\in\mathcal{F} \\ (a_1,b_1)\neq(a_2,b_2)}} \sqrt{\frac{(M(a_1, b_2) + \rho - 1)(M(a_2, b_1) + \rho - 1)}{(M(a_1, b_1) + \rho - 1)(M(a_2, b_2) + \rho - 1)}}.\end{aligned}$$

As $\mathcal{F}$ is a fooling set for M, we obtain that the latter is bounded from below by

$$\begin{aligned}&\log |\mathcal{F}| - \frac{1}{|\mathcal{F}|} \sum_{\substack{(a_1,b_1),(a_2,b_2)\in\mathcal{F} \\ (a_1,b_1)\neq(a_2,b_2)}} \sqrt{\frac{(\rho - 1)(\rho - 1 + \delta_+)}{(\rho - 1 + \delta_-)(\rho - 1 + \delta_-)}} \\ =& \log |\mathcal{F}| - (|\mathcal{F}| - 1) \cdot \frac{\sqrt{(\rho - 1)(\rho - 1 + \delta_+)}}{\rho - 1 + \delta_-}.\end{aligned}$$

We now require

$$\log |\mathcal{F}| - (|\mathcal{F}| - 1) \cdot \frac{\sqrt{(\rho - 1)(\rho - 1 + \delta_+)}}{\rho - 1 + \delta_-} \geq \log \alpha \tag{26}$$

and approximating the solution for $\rho - 1$ we obtain the claim. For the convenience of the reader, we present a short verification: let

$$K := \left(\frac{\log\left(|\mathcal{F}| / \alpha\right)}{|\mathcal{F}| - 1}\right)^2.$$

It follows from the assumptions of the corollary:

$$\frac{(\rho - 1 + \delta_-)^2}{(\rho - 1)(\rho - 1 + \delta_+)} = 1 + \underbrace{\frac{\delta_-^2}{(\rho - 1)\delta_+}}_{\geq 1/K} - \underbrace{\frac{(\delta_+ - \delta_-)^2}{\delta_+ (\rho - 1 + \delta_+)}}_{\leq \frac{(\delta_+ - \delta_-)^2}{\delta_+^2} \leq 1} \geq \frac{1}{K},$$

which is just a rearranging of (26). □

Corollary 6.4 (Inapproximability of $[0, 1]^n$). *Let P be a combinatorial n-cube and let Q be a ρ-approximate EF of P with $\rho - 1 = (4n)^{-2}$. Then* $\operatorname{size}(Q) \geq \sqrt{2} \cdot n$.

Proof. The fooling set $\mathcal{F}$ for $[0, 1]^n$ provided in Fiorini et al. [2013] has size $2n$ and $\delta_- = \delta_+ = 1$. With $\alpha := \sqrt{2}n$ we obtain with Corollary 6.3 that for $\rho - 1 \leq (2(2n - 1))^{-2}$, the $(\rho - 1)$-shift of the slack matrix has nonnegative rank at least $\sqrt{2} \cdot n$, proving the claim together with Braun et al. [2012]. □

Remark 6.5. Compare the result in Corollary 6.4 with the approximation Q of $[0, 1]^n$ given by the simplex defined by the nonnegativity constraints $x \geq 0$ and the inequality $ex \leq n$ (altogether $n + 1$ inequalities) where $e = (1, \ldots, 1)$. Now $\max_P x_1 = 1$, however $\max_Q x_1 = n$.

In a similar way we can generalize [Fiorini et al., 2013, Proposition 5.10] and [Fiorini et al., 2013, Proposition 5.11]. In fact, with Corollary 6.3 every fooling set for a matrix can be turned into a lower bound on the nonnegative rank of a shift of that matrix, leading to lower bounds for the approximate extension complexity.

Corollary 6.6 (Inapproximability of the bipartite matching polytope)**.** *Let $n \geq 4$ and P be the bipartite matching polytope and let Q be a ρ-approximate EF of P with $\rho = 1 + ((2-\varepsilon)/(n+\varepsilon)(n^2 + 2n - 1))^2$. Then* $\operatorname{size}(Q) \geq n^2 + \varepsilon n$.

It is not too hard to see that the approximate fooling set method is stronger than the fooling set method. However, it is also subject to limitations stemming from the continuity of $\gamma(\mathcal{F}, \cdot)$.

Example 6.7. Let $P \subseteq \mathbb{R}^n$ be a regular n-gon. Then $\operatorname{xc}(P) = \Theta(\log n)$. Now suppose that we perturb P to $\tilde{P}$ so that $\tilde{P}$ is a generic n-gon. Then $\operatorname{xc}(\tilde{P}) = \Omega(\sqrt{n})$ (see Ben-Tal and Nemirovski [2001], Fiorini et al. [2012b])

Let $\tilde{M}$ be a slack matrix for $\tilde{P}$. Suppose we start from an approximate fooling set $\mathcal{F}$ for $\tilde{M}$. By slightly perturbing $\tilde{M}$ to be a slack matrix for P we obtain by continuity of γ that

$$|\gamma(\mathcal{F}, M) - \gamma(\mathcal{F}, \tilde{M})| < \varepsilon,$$

and we have that $2^{\gamma(\mathcal{F},M)} \leq \operatorname{rank}_+(P) = O(\log n)$ by Lemma 6.1. Thus $2^{\gamma(\mathcal{F},\tilde{M})} = O(\log n)$ and so we cannot obtain a strong lower bound for generic n-gons via the information theoretic fooling set method.

Another example is given by a matrix that is close to the slack matrix of the matching polytope.

Example 6.8. Let $M \in \mathbb{R}_+^{2^n \times 2^n}$ be the partial matrix that is defined as

$$M(a,b) := \begin{cases} |a \cap b| - \varepsilon & \text{if } a \cap b \neq \emptyset, \\ \geq 0 & \text{otherwise.} \end{cases}$$

Then for $\varepsilon > 0$ we have $2^n - (n+1) \leq \operatorname{rank}(M) \leq \operatorname{rank}_+(M)$, (which follows from a reduction to disjointness by Razborov [2012], see below). On the other hand, for $\varepsilon = 0$ we have $\operatorname{rank}(M) = n$. With a similar argument as in Example 6.7, we cannot find an information theoretic fooling set for M with $\varepsilon > 0$ small, of size larger than n.

The following proof of $\operatorname{rank}(M) \geq 2^n - (n+1)$, has been suggested by one of the reviewers, improving our previous lower bound of $\binom{n}{n/2}$. Let $N(a,b) := |a \cap b| - \varepsilon$, then N has rank $n+1$. We claim that $M - N$ has full rank 2^n, which immediately implies $\operatorname{rank}(M) \geq 2^n - (n+1)$.

Let $T(a,b) := (M - N)(a, [n] \setminus b)$, i.e., T is obtained from $M - N$ by permuting columns, in particular T has the same rank as $M - N$. By definition, $T(a,b) = 0$ for $a \not\subseteq b$, and $T(a,b) \geq \varepsilon > 0$ if $a \subseteq b$. Hence using a total ordering on the subsets of $[n]$ extending the inclusion relation $\subseteq$, the matrix T becomes upper diagonal with positive diagonal entries, and therefore T is full-dimensional.

7 Concluding remarks

We introduced a new framework to lower bound the nonnegative rank of a matrix in terms of common information, which is in turn estimated via the Hellinger distance. We believe that this framework is more widely applicable to lower bound the nonnegative rank of many other matrices and hence can be used to lower bound the extension complexity of a variety of polytopes. Our estimations on the common information are (almost) optimal for the UDISJ matrix and its variants. Also, our approach immediately generalizes to higher dimensional tensors and the estimations remain virtually the same.

We would like to conclude with several open questions.

Question 7.1. *For which other explicit nonnegative matrices can we compute strong lower bounds on the common information?*

Question 7.2. *Does the rectangle covering bound/rectangle corruption bound have an information theoretic analog?*

These bounds subsume the fooling set bound and the bound of the logarithm of the number of faces. The latter two are incomparable in general, e.g., the fooling set bound is better for $[0,1]^n$ ($2n$ vs. $n \log 3$) and worse for the regular n-gon ($\Theta(\log n)$ vs. 5) (see Fiorini et al. [2013]).

Question 7.3. *Is the approximate fooling set method limited in a way similar to the classical fooling set method?*

Question 7.4. *Is $\frac{\varepsilon}{8 \ln 2} n$ the exact bound on common information in Theorem 4.1? Is there a better condition providing larger common information?*

Acknowledgements

We are indebted to Samuel Fiorini for providing extensive feedback through the various stages of this work and who helped to significantly improve the presentation. We would also like to thank Daniel Dadush, Kostya Pashkovich, Santosh Vempala, and Ronald de Wolf for extremely valuable feedback as well as Jérémie Roland for pointing us to Jain et al. [2013].

Part of this work was conducted at the Dagstuhl Seminar 13082 on *Communication Complexity, Linear Optimization, and lower bounds for the nonnegative rank of matrices*[1]. The authors would like to thank the organizers for providing such a stimulating environment. Research reported in this paper was partially supported by NSF grant CMMI-1300144.

References

S. Arora, R. Ge, R. Kannan, and A. Moitra. Computing a nonnegative matrix factorization–provably. In *Proceedings of the 44th symposium on Theory of Computing*, pages 145–162. ACM, 2012.

D. Avis and H. R. Tiwary. On the extension complexity of combinatorial polytopes. *ArXiv e-prints*, Feb. 2013.

Z. Bar-Yossef, T. Jayram, R. Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. *Journal of Computer and System Sciences*, 68(4): 702–732, 2004. doi: 10.1016/j.jcss.2003.11.006. URL `http://www.sciencedirect.com/science/article/pii/S0022000003001855`.

A. Ben-Tal and A. Nemirovski. On polyhedral approximations of the second-order cone. *Math. Oper. Res.*, 26:193–205, 2001. doi: 10.1287/moor.26.2.193.10561.

G. Braun and S. Pokutta. The matching polytope does not admit fully-polynomial size relaxation schemes. *arXiv e-prints*, page 1403.6710 [cs.CC], 2014.

G. Braun, S. Fiorini, S. Pokutta, and D. Steurer. Approximation Limits of Linear Programs (Beyond Hierarchies). In *53rd IEEE Symp. on Foundations of Computer Science (FOCS 2012)*, pages 480–489, 2012. ISBN 978-1-4673-4383-1. doi: 10.1109/FOCS.2012.10.

G. Braun, S. Fiorini, and S. Pokutta. Average case polyhedral complexity of the maximum stable set problem. *submitted*, 2013a.

G. Braun, R. Jain, T. Lee, and S. Pokutta. Information-theoretic approximations of the nonnegative rank. *submitted*, 2013b.

M. Braverman and A. Moitra. An information complexity approach to extended formulations. *Electronic Colloquium on Computational Complexity (ECCC)*, 19(131), 2012.

M. Braverman, A. Garg, D. Pankratov, and O. Weinstein. Information lower bounds via self-reducibility. *Electronic Colloquium on Computational Complexity*, 12(177), 2012a.

[1] `http://www.dagstuhl.de/13082`

M. Braverman, A. Garg, D. Pankratov, and O. Weinstein. From information to exact communication. In *Electronic Colloquium on Computational Complexity (ECCC)*, volume 19, page 171, 2012b.

S. Chan, J. Lee, P. Raghavendra, and D. Steurer. Approximate constraint satisfaction requires large LP relaxations. accepted for FOCS, 2013.

M. Conforti, G. Cornuéjols, and G. Zambelli. Extended formulations in combinatorial optimization. *4OR*, 8:1–48, 2010. doi: 10.1007/s10288-010-0122-z.

T. Cover and J. Thomas. *Elements of information theory*. Wiley-interscience, 2006.

Y. Faenza, S. Fiorini, R. Grappe, and H. R. Tiwary. Extended formulations, non-negative factorizations and randomized communication protocols. In *Proceedings of the Second international conference on Combinatorial Optimization (ISCO 2012)*, pages 129–140, 2012.

S. Fiorini, S. Massar, S. Pokutta, H. R. Tiwary, and R. de Wolf. Linear vs. Semidefinite Extended Formulations: Exponential Separation and Strong Lower Bounds. *Proceedings of STOC 2012*, 2012a.

S. Fiorini, T. Rothvoß, and H. Tiwary. Extended formulations for polygons, Oct. 2012b.

S. Fiorini, V. Kaibel, K. Pashkovich, and D. O. Theis. Combinatorial bounds on nonnegative rank and extended formulations. *Discrete Mathematics*, 313:67–83, 2013.

N. Gillis. Sparse and unique nonnegative matrix factorization through data preprocessing. *Journal of Machine Learning Research*, 13:3349–3386, Nov. 2012.

N. Gillis, F. Glineur, and D. O. Theis. personal communication. 2 2013.

J. Håstad. Clique is hard to approximate within $1-\varepsilon$. *Acta Mathematica*, 182(1):105–142, 1999.

R. Jain, Y. Shi, Z. Wei, and S. Zhang. Efficient protocols for generating bipartite classical distributions and quantum states. *Proceedings of SODA 2013*, 2013.

V. Kaibel. Extended formulations in combinatorial optimization. *Optima*, 85:2–7, 2011.

J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on information theory*, 37(I):145–151, Jan. 1991. doi: 10.1109/18.61115.

A. Moitra. An almost optimal algorithm for computing nonnegative rank. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2013)*, pages 1454–1464, 2013.

K. Pashkovich. *Extended Formulations for Combinatorial Polytopes*. PhD thesis, Magdeburg Universität, 2012.

S. Pokutta and M. Van Vyve. A note on the extension complexity of the knapsack polytope. *to appear in Operations Research Letters*, 2013.

A. A. Razborov. On the distributional complexity of disjointness. *Theoret. Comput. Sci.*, 106(2): 385–390, 1992.

A. A. Razborov. personal communication. 2012.

T. Rothvoß. Some 0/1 polytopes need exponential size extended formulations, 2011. arXiv:1105.0036.

T. Rothvoß. The matching polytope has exponential extension complexity. *Proceedings of STOC*, pages 263–272, 2014. doi: 10.1145/2591796.2591834.

Y. Shitov. An upper bound for nonnegative rank. *Journal of Combinatorial Theory*, pages 126–132, Mar. 2014.

H. S. Witsenhausen. Values and bounds for the common information of two discrete random variables. *SIAM Journal on Applied Mathematics*, 31(2):313–333, 1976.

R. d. Wolf. Nondeterministic quantum query and communication complexities. *SIAM Journal on Computing*, 32(3):681–699, 2003.

A. Wyner. The common information of two dependent random variables. *Information Theory, IEEE Transactions on*, 21(2):163–179, 1975.

M. Yannakakis. Expressing combinatorial optimization problems by linear programs (extended abstract). In *Proc. STOC 1988*, pages 223–228, 1988.

M. Yannakakis. Expressing combinatorial optimization problems by linear programs. *J. Comput. System Sci.*, 43(3):441–466, 1991. doi: 10.1016/0022-0000(91)90024-Y.

S. Zhang. Quantum strategic game theory. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 39–59. ACM, 2012.