Electronic Colloquium on Computational Complexity, Revision 1 of Report No. 131 (2012)

An Information Complexity Approach to Extended Formulations

Mark Braverman* Ankur Moitra †

February 13, 2014

Abstract

We prove an unconditional lower bound that any extended formulation that achieves an $O(n^{1-\epsilon})$ approximation for clique has size $2^{\Omega(n^{\epsilon})}$. There has been considerable recent interest in proving lower bounds for extended formulations. Fiorini et al [14] proved that there is no polynomial sized extended formulation for traveling salesman. Braun et al [7] proved that there is no polynomial sized $O(n^{1/2-\epsilon})$-approximate extended formulation for clique. Here we prove an optimal and unconditional lower bound against extended formulations for clique that matches Håstad's [16] celebrated hardness result. Interestingly, the techniques used to prove such lower bounds have closely followed the progression of techniques used in communication complexity. Here we develop an information theoretic framework to approach these questions, and we use it to prove our main result.

Also we resolve a related question: How many bits of communication are needed to get ϵ-advantage over random guessing for disjointness? Kalyanasundaram and Schnitger [19] proved that a protocol that gets constant advantage requires $\Omega(n)$ bits of communication. This result in conjunction with amplification implies that any protocol that gets ϵ-advantage requires $\Omega(\epsilon^2 n)$ bits of communication. Here we improve this bound to $\Omega(\epsilon n)$, which is optimal for any $\epsilon > 0$.

*Princeton University. Email: mbraverm@cs.princeton.edu. Research supported in part by an Alfred P. Sloan Fellowship, an NSF CAREER award, and a Turing Centenary Fellowship.

†Institute for Advanced Study and Princeton University. Email: moitra@ias.edu. Research supported in part by NSF grant No. DMS-0835373 and by an NSF Computing and Innovation Fellowship.

ISSN 1433-8092

1 Introduction

1.1 Background

A typical goal in combinatorial optimization is to minimize a linear objective function over a discrete set of solutions. A powerful change in perspective is obtained by representing these discrete solutions as vectors and taking their convex hull to get instead a continuous optimization problem. Thus to many discrete sets of combinatorial interest, we associate a polytope and through this paradigm algorithms for linear programming and geometric insights about polytopes and facets can be brought to bear to design new algorithms and approximation algorithms. Unfortunately the natural encoding of a discrete optimization problem as a continuous one often has an exponential number of facets. Nevertheless often it is possible to express a given polytope P as the linear projection of a higher dimensional polytope Q and in so doing to drastically reduce the number of facets needed. In this case, we call Q an *extension* of the polytope P (sometimes we will refer to Q as an *extended formulation* instead). In general P and Q are related by

$$P = \{x | (x, y) \in Q\}$$

and we can draw an analogy with the more familiar case of Boolean formulas: every Boolean formula can be expressed without using quantifiers, but often a Boolean formula can be expressed much more succinctly using quantifiers. Indeed, in many well-known examples in combinatorial optimization, particular polytopes admit a much more compact description as the linear projection of a higher dimensional polytope. We will refer to the minimum number of facets of the higher dimensional polytope as the *extension complexity* of P.

For example, the *permutahedron* is the convex hull of permutations of the vector $\{1, 2, 3, ..., n\}$. Rado gave an exact characterization of this polytope using an exponential number of inequalities but omitting even a single one results in a polytope that strictly contains the permutahedron. However, if we represent each permutation instead as a permutation matrix then the permutahedron can be recovered from a linear projection and the Birkhoff-von Neumann theorem gives a set of $2n^2$ linear constraints (that the matrix be doubly-stochastic) that exactly characterize the convex hull of permutation matrices. Another well-known example is the spanning tree polytope: Edmonds gave an exact characterization of this polytope using an exponential number of facets, and Martin [23] gave a compact extended formulation using $O(n^3)$ constraints. In fact, there are even general principles by which one can obtain a compact extended formulation: If the separation oracle for a linear programming problem itself relies on linear programming, then one can often give a polynomial-sized extended formulation. Also, dynamic programming algorithms typically lead to an extended formulation for the associated polytope whose size is roughly the space used by the algorithm. Lastly, Balas [4] proved that polytopes that admit a compact extended formulation are closed under disjunction.

There has been remarkable success in characterizing the facets of certain polytopes that arise in combinatorial optimization. In some cases, this can lead to faster algorithms (that avoid using the Ellipsoid algorithm). And yet another motivation is that having a convenient representation of a given polytope can make it easier to prove a statement about purely discrete objects. (For example, the T-join polytope and the spanning tree polytope play an important role in recent approximation algorithms for graphic TSP [25] and asymmetric TSP [3] respectively).

Yet there are many polytopes that we do not expect to have a simple description. For example, if we could find a polynomial-sized extended formulation expressing the convex hull of all traveling salesman tours, we could then use any one of a number of efficient algorithms known for linear programming to solve the traveling salesman problem in polynomial time. So certainly, believers in

$NP \subsetneq P\backslash\text{poly}$[1] do not believe such an extended formulation exists (and conversely, if you believe that $P = NP$, it is quite reasonable to believe that an algorithm for 3-SAT would involve linear programming). Remarkably, there are tools for understanding the extension complexity of a given polytope (i.e. for proving that there is no compact extended formulation). To describe these tools, we must introduce some notation:

In a ground breaking work, Yannakakis defined the notion of a *slack matrix*. Given a polytope P with v vertices and f facets, the slack matrix is a $v \times f$ entry-wise nonnegative matrix whose (i, j)th entry is $b_j - \langle a_j, v_i \rangle$ where v_i is the ith vertex of P and $\langle a_j, x \rangle \leq b_j$ is the jth constraint. This matrix provides a tight connection between two combinatorial concepts - the first is the *extension complexity* as defined earlier, and the second is the *nonnegative rank*: The nonnegative rank of a matrix M is the smallest r so that M can be written as $M = \sum_{i=1}^{r} A_i$ where each A_i is a nonnegative, rank one matrix. We will use $rank^+(M)$ to denote the nonnegative rank of M. This concept plays an important role in various methods in machine learning (see [2] and [24]). And Yannakakis proved that the nonnegative rank of the slack matrix is precisely the extension complexity (of a given polytope P). This result is often referred to as Yannakakis's factorization theorem and exactly characterizes one important combinatorial parameter – the extension complexity – in terms of another – the nonnegative rank.

1.2 Extended Formulations and Communication Complexity

So if there are polytopes that we do not expect to have a simple description, how might we go about proving that the nonnegative rank of the slack matrix is indeed superpolynomial? The increasingly sophisticated techniques used to prove strong lower bounds on the nonnegative rank have followed in parallel with the techniques used to prove lower bounds on communication complexity! In order to compare the various works that prove lower bounds on nonnegative rank (and the technique that each one uses) it is helpful to consider each work as providing an answer to the question:

Question 1.1. *If a matrix M has small nonnegative rank, what does that imply?*

Then each work uses a particular consequence of small nonnegative rank to derive a contradiction (with some communication complexity result).

- [14]: If M has small nonnegative rank, then a nonnegative matrix factorization $M = \sum_{i=1}^{r} A_i$ gives a covering of the non-zero entries of M by combinatorial rectangles (the support of each A_i). This in turn contradicts the lower bounds for the nondeterministic communication complexity of unique disjointness (when M is chosen to be the correlation polytope).

Fiorini et al [14] used this connection to prove that there is no polynomial-sized extended formulation for the traveling salesman polytope. An important point is that this result (and others in this line) would follow immediately from $P \neq NP$, but in fact this result holds without any assumption!

- [7]: If M has small nonnegative rank, then there is a large rectangle in M with discrepancy bounded away from zero and this contradicts Razborov's rectangle corruption lemma.

Braun et al [7] use this to prove that any extended formulation for clique that achieves an $O(n^{1/2-\epsilon})$ approximation has size at least 2^{n^ϵ}. What Braun et al [7] need is a technique to prove

[1]Note that even if a polynomial-sized extended formulation did exist, it is not necessarily easy to compute what it is. So in general the existence of a polynomial-sized extended formulation would only imply that if we are given advice about what it is, we could solve NP-hard problems.

lower bounds on the nonnegative rank of matrices that have *no* zero entries, and accomplish this through a generalization of Razborov's rectangle corruption lemma (for disjointness).

Interestingly, the result of [7] falls short of the known NP-hardness results for clique. If $P \neq NP$, then the celebrated result of Håstad [16] and later Khot [21] and Zuckerman [33] directly imply that any extended formulation for clique that achieves an $O(n^{1-\epsilon})$ approximation has super-polynomial size. Can we prove an analogue of this result without the assumption that $P \neq NP$?

In this paper, we give a new method to lower bound nonnegative rank.

- [this work]: If M has small nonnegative rank, then we can use the factorization $M = \sum_i A_i$ to generate a uniform sample from S – the set of pairs of disjoint strings – using too few bits of entropy.

This new method allows us to draw on some of the most powerful tools in communication complexity – *information complexity*. Information complexity has been recently shown [20] to contain as a sub-case many other lower bound techniques for two-player communication complexity. This is the first time that tools from information complexity have been used in the context of nonnegative rank, and we believe that this new method for proving lower bounds will be useful in other applications.

Our main theorem is (see Section 3 for background information on the polytope associated with clique):

Theorem 1.2. *Any extended formulation for clique that achieves an $s+1$-approximation has size at least $2^{\Omega(\frac{n}{s+1})}$.*

Thus any extended formulation that achieves an $O(n^{1-\epsilon})$-approximation must have size at least $2^{\Omega(n^{\epsilon})}$ and our lower bound for extended formulations matches the celebrated hardness results of Håstad [16]. In fact, the best known approximation algorithm for clique is due to Feige [13] and achieves an approximation ratio of $O(n(\log\log n)^2/(\log n)^3)$ and our results even imply that any extended formulation that achieves this bound has size at least $\Omega(n^{\tilde{\Omega}((\log n)^2)})$. This is the *first* example that we know of in which there is a polynomial time approximation algorithm and yet provably there is *no* polynomial-sized extended formulation that achieves the same approximation ratio! Although we remark that it is widely believed that there is no polynomial-sized extended formulation [2] for matching even though there is an efficient algorithm[3].

1.3 Disjointness

In order to explain our main technical contribution (even at a heuristic level) it helps to consider the communication complexity of disjointness. The techniques of Braun et al [7] break down at $n^{1/2}$ precisely because there is a gap in our understanding of the communication complexity of *disjointness*!

In this problem, Alice and Bob are given $a, b \in \{0,1\}^n$ and their goal is to determine whether or not these two strings are disjoint, that is, whether there is an index $i \in [n]$ such that $a_i =$

[2] Yannakakis [31] proved that there is no polynomial-sized *symmetric* extended formulation for matching.

[3] There is a combinatorial algorithm for this problem. And in fact there is also a polynomial time algorithm based on a linear program: even though the characterization of the matching polytope given by Edmonds [12] has exponential size, there is a polynomial time separation oracle for these constraints [26] and hence matching with general weights can be solved by the ellipsoid algorithm. This illustrates a subtlety in not only our lower bounds for extended formulationss but all such works; these lower bounds show that there is no polynomial-sized extended formulation, but it still could be the case that there is an exponential-sized extended formulation that has an efficient separation oracle.

$b_i = 1$. Kalyanasundaram and Schnitger [19] were the first to prove an $\Omega(n)$ lower bound on the communication complexity of any protocol that gets a constant advantage. In fact any linear lower bound in conjunction with amplification implies that any protocol that gets ϵ-advantage requires $\Omega(\epsilon^2 n)$ bits of communication[4]. Since then a number of simpler proofs have been given. Razborov [27] proved an elegant rectangle corruption lemma which yields a linear lower bound. Perhaps the most intuitive lower bound was given in the ground breaking work of Bar-Yossef et al [6] which used information theoretic arguments to prove that an protocol that gets advantage ϵ over random guessing must communicate at least $\Omega(\epsilon^2 n)$ bits. However, this does not quite settle the asymptotic complexity of disjointness.

Question 1.3. *Is there a protocol that gets advantage $1/\sqrt{n}$ for disjointness and uses a subpolynomial amount of communication?*

There is a simple to state technical issue that arises in these proofs, and makes $\Omega(\epsilon^2 n)$ a barrier for these approaches: For example, the work of Bar-Yossef et al [6] works by reducing an n-bit communication problem – disjointness – to a one-bit problem – an AND. Indeed, any lower bound (say, γ) on the information revealed to the participants (about the inputs to Alice and Bob) based on the transcript implies a γn lower bound on the communication complexity of the n-bit problem. Yet, there is a protocol for AND that gets advantage ϵ and reveals only $O(\epsilon^2)$ bits of information (see Lemma 2.1)!

So in a sense, the n-bit problem cannot be reduced to n instances of a one-bit problem! Yet here we are able to circumvent this issue. Instead of considering the advantage of a protocol, we consider a more nuanced quantity – a matrix that describes the probability of outputting a one for each pair of input bits for Alice and Bob. Suppose this matrix is:

$$N = \begin{bmatrix} N_{00}, N_{01} \\ N_{10}, N_{11} \end{bmatrix}$$

Then the advantage of the protocol is at least γ if and only if $N_{00}, N_{10}, N_{01} \geq 1/2 + \gamma$ and $N_{11} \leq 1/2 - \gamma$. Indeed, there is a protocol for AND that gets $N_{00} = 1/2 + 5\epsilon + \Theta(\epsilon^2)$, $N_{10} = N_{01} = 1/2 + \epsilon + \Theta(\epsilon^2)$, $N_{11} = 1/2 - 3\epsilon + \Theta(\epsilon^2)$, and reveals $O(\epsilon^2)$ bits of information. Another way to think about this is that Hellinger distance (see [6] for the details) implies that the information revealed is at least $(N_{00} - N_{11})^2$. Our main insight is:

- The information revealed is at least $\Omega(N_{10} + N_{01} - N_{00} - N_{11})$.
- This does not directly imply a better than $\Omega(\epsilon^2)$ lower bound for one-bit AND (afterall, how could it since there is such a protocol?) but nevertheless any protocol for disjointness can be "smoothed" so that the communication complexity does not increase and yet, in expectation $N_{00} = N_{10} = N_{01}$.

To put it another way, after this "smoothing" operation changing a pair of bits (a_j, b_j) from (say) $(0,0)$ to $(1,0)$ will not change the probability that the protocol outputs one. Hence the one-bit AND problems that we get by considering just a single pair of bits (a_j, b_j) is not an arbitrary problem where the goal is to get advantage ϵ in any way possible but rather to do so in such a way that the probability of outputting a one is the same, independent of whether the input is $(0,0)$, $(0,1)$ or $(1,0)$. Yet for protocols that meet this extra restriction, getting an advantage of ϵ implies

[4] If a protocol gets ϵ-advantage and requires $o(\epsilon^2 n)$ bits of communication, then we can run the protocol $O(1/\epsilon^2)$ times and take the majority vote to get constant advantage, but this would violate the linear lower bound for disjointness.

that at least $\Omega(\epsilon)$ bits are revealed. Hence using the direct sum results for information complexity, we obtain an $\Omega(\epsilon n)$ lower bound for disjointness, thus resolving the asymptotic complexity of disjointness for any $\epsilon > 0$ (not just $\epsilon = \Omega(1)$).

How does this relate to nonnegative rank? As we mentioned, Braun et al [7] gave an interesting generalization of Yannakakis's factorization theorem to the case of approximate extended formulations. And it turns out that if you want to prove a lower bound on the extension complexity of any polytope that approximates P within a multiplicative factor of C, then this quantity is equal to the nonnegative rank of $M + (C-1)J$, where M is the slack matrix of P and J is the all ones matrix. (And indeed nonnegative rank lower bounds are proven using communication complexity lower bounds for disjointness). Then we can think of the factors A_i (in $M + (C-1)J = \sum_i A_i$) intuitively as the "output" matrix N in the preceding discussion, and hence proving a nonnegative rank lower bound for $C = n^{1-\epsilon}$ is akin to proving a lower bound for disjointness when the target advantage $\epsilon \approx 1/C$. This is precisely why the proof of [7] breaks down for $C = \sqrt{n}$ - because it was not known whether or not getting advantage $1/\sqrt{n}$ for disjointness requires a polynomial amount of communication!

However, lower bounds for disjointness *do not* immediately yield lower bounds for nonnegative rank. Intuitively, many techniques in communication complexity aim to show that a player must reveal too much information about his input in order to solve the communication problem. Yet in the setting of nonnegative rank[5], a player can reveal a nonnegative real value (which can contain a great deal of information about his input).

2 A New Lower Bound for Disjointness

Here we prove that any protocol for disjointness that gets advantage ϵ over random guessing must communicate at least $\Omega(\epsilon n)$ bits between the two players. This improves the $\Omega(\epsilon^2 n)$ lower bound due to Razborov [27] and Bar-Yossef et al [6]. The known lower bounds for disjointness were optimal only for $\epsilon = \Omega(1)$ – it was a priori possible that a protocol that gets advantage $1/\sqrt{n}$ requires as much as $\Omega(\sqrt{n})$ communication or as little as $O(1)$ communication. Our lower bound completely resolves the asymptotic complexity of disjointness for any $\epsilon > 0$.

The standard (information complexity) approach for proving lower bounds is to reduce an n-bit problem to a one-bit problem (which in the case of disjoints is a one-bit AND). The difficulty is that there is indeed a protocol that gets advantage ϵ for one-bit AND, but reveals $O(\epsilon^2)$ bits of information to an observer:

Lemma 2.1. *There is a communication protocol for one-bit AND that gets advantage ϵ over random guessing but reveals $O(\epsilon^2)$ bits of information to an observer.*

Proof. We first describe a simpler protocol that almost works: Alice and Bob behave identically, and on input one Alice sends a one with probability $1/2 + 4\epsilon$ and otherwise sends a zero. On input zero Alice sends a zero with probability $1/2 + 4\epsilon$ and otherwise sends a one. (Bob behaves identically). If both players send a one, the protocol outputs one. If both players send a zero, the protocol outputs zero. Otherwise the protocol flips a coin and outputs either one or zero with equal probability. (This is not quite the protocol in the lemma, but a simple modification will complete the proof). The probability that the protocol outputs a zero, for each pair of inputs, is given by the matrix:

$$N = \begin{bmatrix} 1/2 + 4\epsilon, 1/2 \\ 1/2, 1/2 - 4\epsilon \end{bmatrix}$$

[5]Here, one player is given a row and one player is given a column, and their goal is to compute the value of corresponding entry in expectation, using only nonnegative values.

This protocol Γ does not have an advantage over random guessing for all inputs, but we can obtain one that does with a simple modification. The final protocol Π will run Γ with probability $1-2\epsilon$, and with probability 2ϵ will output zero. Then Π has advantage ϵ and a simple calculation shows that it reveals $O(\epsilon^2)$ bits of information to an observer. $\square$

Notice that the probability Π outputs zero is different when the input to Alice and Bob are 0 and 0 or instead 1 and 0 respectively (even though the answer in both of these cases is the same). The standard approach to proving lower bounds (though information complexity) is to use a n-bit protocol to get a one-bit protocol. In fact, we can "symmetrize" any n-bit protocol for disjointness in such a way that the one-bit protocol we obtain has the same probability of outputting zero whether the input to Alice and Bob is 0 and 0, 1 and 0 or 0 and 1 respectively. Our main insight is that this one bit communication problem (with an additional constraint that the top-left, top-right and bottom-left values of N be the same) must reveal $\Omega(\epsilon)$ bits of information.

Theorem 2.2. *Any protocol for disjointness that gets advantage ϵ over random guessing must reveal at least $\Omega(\epsilon n)$ bits of information to the participants.*

Consider a protocol Γ. We will consider the information complexity of Γ measured with respect to a particular distribution on inputs: We will group the n bits into blocks of size exactly three, and for each pair of three bits we will generate $a_j, b_j \in \{0,1\}^3$ uniformly at random from the pairs of length three strings where a_j and b_j have exactly one 1 and two 0's, and a_j and b_j are disjoint. Thus the location of the 1 must be different in a and b, and therefore there are six such pairs.

Given the protocol Γ we will construct a new protocol Π which takes each block of size three and applies a random element π of S_3 to both a_j and b_j. This does not change whether or not a_j or b_j are disjoint, and furthermore has the effect of "smoothing" the protocol Γ: For any pair a_j, b_j (that are disjoint), the probability that Π outputs one is exactly the same. We will abuse notation and let $B_j = i$ be the event that the only one in b_j is the i^{th} bit in the block.

We will be interested in $I(A;\Pi|B) + I(B;\Pi|A)$. The following is well-known:

Fact 2.3. $\frac{1}{2}\Big[\sum_j I(A_j;\Pi|A_{1\ldots j-1}, B_{j\ldots n}) + I(B_j;\Pi|A_{1\ldots j}, B_{j+1\ldots n})\Big] \leq H(\Pi)$

Proof. Using the chain rule for mutual information (see e.g. [8]) we have

$$\frac{1}{2}\Big[\sum_j I(A_j;\Pi|A_{1\ldots j-1}, B_{1\ldots n}) + I(B_j;\Pi|A_{1\ldots n}, B_{j+1\ldots n})\Big] \leq \frac{1}{2}\Big[I(A;\Pi|B) + I(B;\Pi|A)\Big] \leq H(\Pi)$$

Moreover we will use the following inequality for mutual information (again, see [8]): $I(W;X|Y) \leq I(W;X|YZ)$ if $I(X;Z|Y) = 0$. We can apply the inequality above for example by setting $W = \Pi$, $X = A_j$, $Y = A_{1\ldots j-1}B_{j\ldots n}$ and $Z = B_{1\ldots j-1}$. Then it is easy to see that $I(X;Z|Y) = 0$ in this case, and we conclude:

$$I(A_j;\Pi|A_{1\ldots j-1}, B_{j\ldots n}) \leq I(A_j;\Pi|A_{1\ldots j-1}, B_{1\ldots n})$$

Similarly we can set $W = \Pi$, $X = B_j$, $Y = A_{1\ldots j}B_{j+1\ldots n}$ and $Z = A_{j+1\ldots n}$ and again we have $I(X;Z|Y) = 0$ and so

$$I(B_j;\Pi|A_{1\ldots j}, B_{j+1\ldots n}) \leq I(B_j;\Pi|A_{1\ldots n}, B_{j+1\ldots n})$$

and this completes the proof. $\square$

Let $C_j = A_{1\ldots j-1}, B_{j+1\ldots n}$. Then we can write $I(A_j;\Pi|A_{1\ldots j-1}B_{j..n})$ as:

$$\sum_{c,i}\sum_{t} Pr[\Pi = t, C_j = c, B_j = i]D(A_j|C_j = c, B_j = i, \Pi = t\|A_j|C_j = c, B_j = i),$$

where $D(\bullet\|\bullet)$ is the KL-divergence of the two distributions – see [11] for more background on information theory. Note that A_j conditioned on $B_j = i$ is uniform on the set $\{1,2,3\} - \{i\}$.

Let us choose $C_j = A_{1\ldots j-1}, B_{j+1\ldots n}$ according to the distribution on inputs. Then, $I(A_j;\Pi|C_j,B_j)+I(B_j;\Pi|C_j,A_j) = \sum_t \mathbf{E}[adv(t,C_j)]$ where the expectation is over C_j and $adv(t,C_j)$ defined as:

$$\begin{aligned} adv(t,C_j) \quad = \quad & \sum_{i=1,2,3} Pr[\Pi = t, B_j = i|C_j]D(A_j|B_j = i, \Pi = t, C_j\|A_j|B_j = i, C_j) \\ & + \quad Pr[\Pi = t, A_j = i|C_j]D(B_j|A_j = i, \Pi = t, C_j\|B_j|A_j = i, C_j). \end{aligned}$$

Definition 2.4. Let $IC(a,b,c) = (ab+ac)D(\mathcal{B}_{ab/(ab+ac)}\|\mathcal{B}_{1/2})$.

Lemma 2.5. $IC(\alpha_0,\beta_0,\beta_1) + IC(\beta_0,\alpha_0,\alpha_1) \geq \Omega(\alpha_0\beta_1 + \alpha_1\beta_0 - \alpha_0\beta_0 - \alpha_1\beta_1)$

Proof. We have that $D(\mathcal{B}_{\alpha_0\beta_0/(\alpha_0\beta_0+\alpha_0\beta_1)}\|\mathcal{B}_{1/2}) = \Omega\left(\frac{(\beta_0-\beta_1)^2}{(\beta_0+\beta_1)^2}\right)$ and hence it suffices to show that:

$$\frac{2\alpha_0(\beta_0-\beta_1)^2}{\beta_0+\beta_1} + \frac{2\beta_0(\alpha_0-\alpha_1)^2}{\alpha_0+\alpha_1} \geq \alpha_0\beta_1 + \alpha_1\beta_0 - \alpha_0\beta_0 - \alpha_1\beta_1.$$

We multiply both sides by $(\alpha_0+\alpha_1)(\beta_0+\beta_1)$ and the left hand side is:

$$4\alpha_0^2\beta_0^2 + 2\alpha_0^2\beta_1^2 + 2\alpha_1^2\beta_0^2 + 2\alpha_0\alpha_1\beta_1^2 + 2\alpha_1^2\beta_0\beta_1 - 2\alpha_0^2\beta_0\beta_1 - 2\alpha_0\alpha_1\beta_0^2 - 8\alpha_0\alpha_1\beta_0\beta_1.$$

The right hand side is:

$$(\alpha_0+\alpha_1)(\alpha_0-\alpha_1)(\beta_1-\beta_0)(\beta_0+\beta_1) = \alpha_0^2\beta_1^2 + \alpha_1^2\beta_0^2 - \alpha_0^2\beta_0^2 - \alpha_1^2\beta_1^2$$

and subtracting the right hand side from the left hand side we get:

$$(\alpha_0^2\beta_0^2+\alpha_0^2\beta_1^2-2\alpha_0^2\beta_0\beta_1)+(\alpha_0^2\beta_0^2+\alpha_1^2\beta_0^2-2\alpha_0\alpha_1\beta_0^2)+(3\alpha_0^2\beta_0^2+\alpha_1^2\beta_1^2+2\alpha_0\alpha_1\beta_1^2+2\alpha_1^2\beta_0\beta_1-8\alpha_0\alpha_1\beta_0\beta_1)$$

which is at least zero, using the weighted AMGM inequality. □

We will consider a fixed block j, and the matrix $N^t(C_j)$ that gives the probability of $\Pi = t$ (where the output is one) for each pair of inputs for Alice and Bob conditioned on the parts of their input C_j that we have already fixed. (Note that the expectation here is taken over the randomness of the protocol and the remaining bits in the input to Alice and Bob). To simplify notation we will abbreviate $N^t(C_j)$ as N^t. Let us write:

$$N^t = \begin{bmatrix} N^t_{11}, N^t_{12}, N^t_{13} \\ N^t_{21}, N^t_{22}, N^t_{23} \\ N^t_{31}, N^t_{32}, N^t_{33} \end{bmatrix}$$

Note that N^t is a nonnegative. Since Π is a protocol and Alice and Bob can privately sample their remaining bits conditioned on C_j, A_j and B_j we conclude that N^t is a rank one matrix[6]. So we can write $N^t = [a_1,a_2,a_3][b_1,b_2,b_3]^T$. In particular b_i is the probability over $B_{1\ldots j-1}$ that the string $B = B_{1\ldots j-1}, B_j = i, B_{j+1\ldots n}$ is in the rectangle for $\Pi = t$.

[6]We would like to thank Thomas Watson for pointing out an oversight in our earlier version. We had asserted that N^t is rank one, but had neglected to condition on C_j as we had intended to do. Indeed, this is the same conditioning trick that we make use of in our later applications to extended formulations, and is by now standard in information complexity (see e.g. [9]).

Lemma 2.6. $adv(t, C_j) = \Omega(\sum_{i \neq i'} N^t_{ii'}(C_j) - 2\sum_{i=1,2,3} N^t_{ii}(C_j))$

Proof. We will prove this lemma by repeatedly applying Lemma 2.5. For example, set $\alpha_0 = a_3$, $\alpha_1 = a_1$ and $\beta_0 = b_2$, $\beta_1 = b_1$. We can symbolically think of this as "covering" a set of entries in the matrix N^t:

$$\begin{bmatrix} \alpha_1\beta_1, \alpha_1\beta_0, - \\ -, \quad -, \quad - \\ \alpha_0\beta_1, \alpha_0\beta_0, - \end{bmatrix}$$

where we have chosen α_1 and β_1 so that the corresponding entry in N^t is on the diagonal. The term $IC(\alpha_0, \beta_0, \beta_1)$ equals $Pr[\Pi = t|A_j = 3, C_j]D(B_j|A_j = 3, \Pi = t, C_j \| B_j|A_j = 3, C_j)$ and we can think of this as the bottom row in this covering. Similarly, the term $IC(\beta_0, \alpha_0, \alpha_1)$ equals $Pr[\Pi = t|B_j = 2, C_j]D(A_j|B_j = 2, \Pi = t, C_j \| A_j|B_j = 2, C_j)$ and we can think of this as the middle column in the covering. And applying Lemma 2.5 has the effect of adding $N^t_{1,2}$ and $N^t_{3,1}$ and subtracting $N^t_{1,1}$ and $N^t_{3,2}$. We can apply Lemma 2.5 to the following covering scheme:

$$\begin{bmatrix} *, *, - \\ -, -, - \\ *, *, - \end{bmatrix}, \begin{bmatrix} -, *, * \\ -, *, * \\ -, -, - \end{bmatrix}, \begin{bmatrix} -, *, * \\ -, -, - \\ -, *, * \end{bmatrix}, \begin{bmatrix} *, -, * \\ *, -, * \\ -, -, - \end{bmatrix}, \begin{bmatrix} -, -, - \\ *, *, - \\ *, *, - \end{bmatrix}, \begin{bmatrix} -, -, - \\ *, -, * \\ *, -, * \end{bmatrix}$$

always choosing α_1 and β_1 so that the corresponding entry in N^t is on the diagonal. The left hand side double-counts each term in $adv(t, C_j)$ and the right hand side is $\sum_{i \neq i'} N^t_{ii'} - 2\sum_{i=1,2,3} N^t_{ii}$ and this completes the proof. □

We can now complete the proof of the main theorem in this section.

Proof. If we sum $\sum_{i \neq i'} N^t_{ii'}(C_j) - 2\sum_{i=1,2,3} N^t_{ii}(C_j)$ over all transcripts t for which the output of Π is one and take the expectation over C_j, the value is $\Omega(\epsilon)$ by the assumption that (1) the advantage of the protocol is ϵ and (2) for any pair of disjoint a_j, b_j where a_j and b_j each have exactly one one, the probability that the protocol outputs one is the same. Hence we conclude that $I(A_j; \Pi|B_j, C_j) + I(B_j; \Pi|A_j, C_j) = \Omega(\epsilon)$ for each block j, and combining this with Fact 2.3 we conclude that $H(\Pi) = \Omega(\epsilon n)$ since the input restricted to each block is mutually independent of the other blocks. And so the communication complexity of any protocol for disjointness that gets advantage ϵ is $\Omega(\epsilon n)$. □

3 Extended Formulations for Clique

Recall that the nonnegative rank is defined as follows:

Definition 3.1. The nonnegative rank $rank^+(M)$ of a nonnegative matrix M is the smallest r so that M can be written as $M = \sum_{i=1}^r M_i$ where each M_i is a rank-one nonnegative matrix.

The nonnegative rank plays a central role in lower bounds for extended formulations, and here we will explain this connection in more detail. The central polytope in the recent breakthrough lower bounds of Fiorini et al [14] and of Braun et al [7] is the *correlation polytope*.

Definition 3.2. $P = COR(n) = conv\{bb^T | b \in \{0,1\}^n\}$

Recall that the smallest extended formulation (for a given polytope) is exactly the nonnegative rank of the *slack matrix*. However one can prove lower bounds on extended formulations (for a given polytope) by considering only a subset of the constraints. This corresponds to proving a nonnegative rank lower bound on a submatrix of the full slack matrix. Fiorini et al [14] identified a set of constraints on the correlation polytope that are enough to prove strong lower bounds:

Lemma 3.3. *[14] For any $a \in \{0,1\}^n$, the inequality $\langle 2diag(a) - aa^T, x\rangle \leq 1$ is valid for $COR(n)$ and the slack of a vertex $x = bb^T$ is $(1 - a^T b)^2$.*

Hence these constraints define a polytope Q for which $P \subset Q$. In fact, Fiorini et al [14] prove that the extension complexity of $COR(n)$ is $2^{\Omega(n)}$ by considering only pairs of vertices bb^T and constraints (for $a \in \{0,1\}^n$) such that $b^T a \in \{0,1\}$. Whether or not a vertex is on the facet is exactly the unique disjointness problem. An important definition from [7] (and a generalization from the one in [31]) is:

Definition 3.4. [7] Let $P \subset Q$, then the slack matrix $SM(P,Q)$ is a nonnegative matrix where each row corresponds to a vertex v_i in P and a column corresponds to a constraint $\langle a_j, x\rangle \leq b_j$ in Q and the corresponding entry in $SM(P,Q)$ is $b_j - \langle a_j, v_i\rangle$.

Braun et al [7] gave a generalization of Yannakakis's factorization theorem [31]. This is the connection between nonnegative rank and extension complexity (that is relevant when our goal is to show lower bounds for approximate extended formulations) that we will use here:

Theorem 3.5. *[7] The minimum extension complexity of any polytope $P \subset K \subset Q$ is exactly the nonnegative rank of $SM(P,Q)$.*

In fact, the correlation polytope naturally defines a linear encoding for clique: Given a graph G on n nodes, one can choose an objective function $w(G)$ (which is an $n \times n$ matrix that is the direction we are trying to maximize over $COR(n)$) where $w(G)$ is one on each diagonal entry and is zero on i,j and j,i if (i,j) is an edge, and otherwise is -1. The maximum value of $w(G)$ over $COR(n)$ is exactly the maximum clique value of G. (And in fact, these directions are *admissible* in the sense of [7] in that any lower bounds for the nonnegative rank of $SM(P,Q)$ imply that any extended formulation for this linear encoding is large).

What about approximate extended formulations for clique? Clearly, we would like to prove lower bounds on the nonnegative rank of $SM(P,(s+1)Q)$ in order to show that there is no small extended formulation (for clique) that has an approximation factor of s. In this case, the slack matrix $SM(P,(s+1)Q)$ has a particularly simple structure: Restricted to the pairs of strings that have at most one intersection, it is $SM(P,Q) + sJ$ where J is the all ones matrix! Hence, what we need to prove lower bounds on approximate extended formulations of clique is just to prove lower bounds for the nonnegative rank of the slack matrix (generated by the correlation polytope and the constraints defined above) when a large value (say $n^{1-\epsilon}$) is added to each entry that corresponds to a pair of strings with at most one intersection.

4 A New Method to Lower Bound Nonnegative Rank

Let M be a $2^n \times 2^n$ nonnegative matrix. We will associate rows and columns of M with strings $a,b \in \{0,1\}^n$ and furthermore if $a^T b \in \{0,1\}$ then $M_{a,b} = s + 1 - a^T b$. If $s = 0$, Fiorini et al [14] proved that $rank^+(M) \geq 2^{\Omega(n)}$. If $s = O(n^\beta)$ with $\beta < 1/2$, Braun et al [7] proved that $rank^+(M) \geq 2^{\Omega(n^{1-2\beta})}$. Here we prove that even if $s = n^{1-\epsilon}$, then $rank^+(M) \geq 2^{\Omega(n^\epsilon)}$.

In fact, we will prove such a lower bound by constructing a sampling procedure (from a nonnegative matrix factorization of M) and deriving an entropy-theoretic contradiction. Suppose that $f_1, f_2, ...f_r$ and $g_1, g_2, ...g_r$ are nonnegative functions defined on $\{0,1\}^n$ and that furthermore for each $a,b \in \{0,1\}^n$ with $a^T b \in \{0,1\}$ we have

$$\sum_{i=1}^{r} f_i(a) g_i(b) = s + 1 - a^T b$$

Algorithm 1. GENERATE, **Output:** (a, b)

1. Choose an index $i \in [r]$ from γ
2. Choose $(a, b) \in S$ from μ_i
3. Output (a, b)

Definition 4.1. Let $S = \{a, b | a^T b = 0\} \subset \{0,1\}^n \times \{0,1\}^n$.

We will next define a sampling procedure (based on f_i and g_i) that generates a uniformly random pair of strings $(a, b) \in S$:

Definition 4.2. Let $W_i = \sum_{(a,b)\in S} f_i(a)g_i(b)$. Let $\gamma(i) = \frac{W_i}{\sum_{i=1}^r W_i}$ and let $\mu_i(a, b) = \frac{f_i(a)g_i(b)}{W_i}$ for all $(a, b) \in S$ and $\mu_i = 0$ elsewhere.

Claim 4.3. *The procedure* GENERATE *samples a $(a, b) \in S$ uniformly at random.*

Proof. The probability of GENERATE outputting $(a, b) \in S$ is

$$\sum_{i=1}^{r} \gamma(i)\mu_i(a, b) = \sum_{i=1}^{r} \frac{W_i}{\sum_{i'=1}^r W_{i'}} \frac{f_i(a)g_i(b)}{W_i} = \frac{s+1}{\sum_{i'=1}^r W_{i'}}$$

□

Throughout this section, we will let $A_1, A_2, ...A_n$ and $B_1, B_2, ...B_n$ denote the random variables associated with the output (a, b) and additionally we will let I denote the random variable associated with the intermediate random variable i in GENERATE.

Corollary 4.4. $H(A_{1...n}, B_{1...n}) = \log |S| = (\log 3)n$ *and* $nH(A_j) = H(A_{1...n}) = H(B_{1...n}) = nH(B_j) = nH(1/3)$

We proceed to invoke the chain rule (and we will analyze each term using properties of f_i and g_i):

$$\begin{aligned}
H(A_{1...n}, B_{1...n}) &\leq H(A_{1...n}, B_{1...n}, I) = H(I) + H(A_{1...n}, B_{1...n}|I) \\
&= H(I) + \frac{1}{2}\Big[H(B_{1...n}|I) + \sum_{j=1}^{n} H(A_j|A_{1...j-1}, B_{1...n}, I)\Big] \\
&\quad + \frac{1}{2}\Big[H(A_{1...n}|I) + \sum_{j=1}^{n} H(B_j|A_{1...n}, B_{j+1...n}, I)\Big] \\
&\leq \log r + nH(1/3) + \frac{1}{2}\sum_{j=1}^{n} H(A_j|A_{1...j-1}, B_{j...n}, I) + \frac{1}{2}\sum_{j=1}^{n} H(B_j|A_{1...j}, B_{j+1...n}, I)
\end{aligned}$$

where in the last inequality we have used the fact that $H(X|Y) \leq H(X)$.

Definition 4.5. Let $adv_j^A(i) = 1 - H(A_j|A_{1...j-1}, B_j = 0, B_{j+1...n}, I = i)$ and similarly let $adv_j^B(i) = 1 - H(B_j|A_{1...j-1}, A_j = 0, B_{j+1...n}, I = i)$.

Note that if $B_j = 1$, then $A_j = 0$ and the entropy of A_j (conditioned on the given events) is zero. We will use this in conjunction with the following fact about conditional entropy:

Fact 4.6. $H(W|X=x,Y,Z) = \sum_z Pr[Z=z|X=x]H(W|X=x,Y,Z=z)$

Then:

$$\begin{aligned}
H(A_{1\ldots n}, B_{1\ldots n}) &\leq \log r + nH(1/3) + \frac{1}{2}\sum_{j=1}^{n} \mathop{\mathbf{E}}_{I}[H(A_j|A_{1\ldots j-1}, B_{j\ldots n}, I=i)] \\
&\quad + \frac{1}{2}\sum_{j=1}^{n} \mathop{\mathbf{E}}_{I}[H(B_j|A_{1\ldots j}, B_{j+1\ldots n}, I=i)] \\
&= \log r + nH(1/3) + \frac{1}{2}\sum_{j=1}^{n} \mathop{\mathbf{E}}_{I}[Pr[B_j=0|I=i]H(A_j|A_{1\ldots j-1}, B_j=0, B_{j+1\ldots n}, I=i)] \\
&\quad + \frac{1}{2}\sum_{j=1}^{n} \mathop{\mathbf{E}}_{I}[Pr[A_j=0|I=i]H(B_j|A_{1\ldots j-1}, A_j=0, B_{j+1\ldots n}, I=i)] \\
&= \log r + nH(1/3) + \frac{1}{2}\sum_{j=1}^{n} \mathop{\mathbf{E}}_{I}[Pr[B_j=0|I=i](1-adv_j^A(i))] \\
&\quad + \frac{1}{2}\sum_{j=1}^{n} \mathop{\mathbf{E}}_{I}[Pr[A_j=0|I=i](1-adv_j^B(i))] \\
&= \log r + nH(1/3) + 2n/3 - \frac{1}{2}\sum_{j=1}^{n} \mathop{\mathbf{E}}_{I}[Pr[B_j=0|I=i]adv_j^A(i)] \\
&\quad - \frac{1}{2}\sum_{j=1}^{n} \mathop{\mathbf{E}}_{I}[Pr[A_j=0|I=i]adv_j^B(i)]
\end{aligned}$$

Note that $H(1/3) + 2/3 = \log 3$.

Definition 4.7. Let $L_{i,j} = adv_j^A(i)Pr[B_j=0|I=i]$.

It will be easier to analyze $adv_j^A(i)$ directly rather than the quantity $L_{i,j}$ and the following lemma will provide the means for translating lower bounds on $adv_j^A(i)$ into lower bounds on $L_{i,j}$:

Lemma 4.8. *Suppose for all $j \in [n], \mathbf{E}_I[adv_j^A(i)] \geq \frac{1}{s}$, then $\mathbf{E}_I[\sum_{j=1}^n L_{i,j}] = \Omega(\frac{n}{s}) - O(\log r)$.*

Proof. Since $adv_j^A(i) \leq 1$, we get:

$$\begin{aligned}
\mathop{\mathbf{E}}_{I}\left[\sum_{j=1}^{n} L_{i,j}\right] &\geq \frac{1}{4}\mathop{\mathbf{E}}_{I}\left[\sum_{j=1}^{n} adv_j^A(i)\right] - \frac{1}{4}\mathop{\mathbf{E}}_{I}\left[|\{j|Pr[B_j=0|I=i] \leq \frac{1}{4}\}|\right] \\
&\geq \frac{n}{4s} - \frac{1}{4}\mathop{\mathbf{E}}_{I}\left[|\{j|Pr[B_j=0|I=i] \leq \frac{1}{4}\}|\right]
\end{aligned}$$

Our immediate goal is to bound $\mathbf{E}_I[|\{j|Pr[B_j=0|I=i] \leq \frac{1}{4}\}|]$. To accomplish this, we will use the following claim:

Claim 4.9. $H(B_j|I) \leq H(1/3) - \Omega(Pr_I[Pr[B_j=0|I=i] < \frac{1}{4}])$

Proof.

Fact 4.10. $H(c+x) \leq H(c) + H'(c)x - \Omega(x^2)$

Hence:

$$\begin{aligned} H(B_j|I) &= \mathop{\mathbf{E}}_I[H(B_j|I=i)] \\ &\leq H(2/3) + H'(2/3)\mathop{\mathbf{E}}_I[Pr[B_j=0|I=i] - 2/3] - \Omega(\mathop{\mathbf{E}}_I[(Pr[B_j=0|I=i]-2/3)^2]) \\ &= H(1/3) - \Omega(\mathop{\mathbf{E}}_I[(Pr[B_j=0|I=i]-2/3)^2]) \\ &\leq H(1/3) - \Omega(Pr_I[Pr[B_j=0|I=i] < \frac{1}{4}) \end{aligned}$$

□

Next, we can use this claim to prove Lemma 4.8:

$$\begin{aligned} H(B_{1\ldots n}) - H(I) &= H(B_{1\ldots n}|I) \leq \sum_{j=1}^{n} H(B_j|I) \\ &\leq \sum_{j=1}^{n} \left(H(B_j) - \Omega(Pr_I[Pr[B_j=0|I=i] < \frac{1}{4}]) \right) \\ &= nH(1/3) - \Omega(\mathop{\mathbf{E}}_I[|\{j|Pr[B_j=0|I=i] \leq \frac{1}{4}\}|]) \end{aligned}$$

However, $H(B_{1\ldots n}) = nH(1/3)$ and so $\mathbf{E}_I[|\{j|Pr[B_j=0|I=i] \leq \frac{1}{4}\}|] = O(H(I)) = O(\log r)$. □

An analogous lemma holds with A and B exchanged. We will prove the following lemma in the next section:

Lemma 4.11 (Main)**.** *For all* $j \in [n], \mathbf{E}_I[adv_j^A(i) + adv_j^B(i)] \geq \Omega(\frac{1}{s+1})$

In fact in the exact case ($s = 0$), it is much easier to prove an $\Omega(1)$ lower bound (see the discussion after Lemma 5.10). We can now put these pieces together:

$$(\log 3)n = H(A_{1\ldots n}, B_{1\ldots n}) \leq O(\log r) + (\log 3)n - \Omega\left(\frac{n}{s+1}\right)$$

and hence $r \geq 2^{\Omega(\frac{n}{s+1})}$, and this proves our main theorem:

Theorem 4.12. $rank^+(M) \geq 2^{\Omega(\frac{n}{s+1})}$

5 The Typical Advantage

Here we prove the main lemma, and this will complete our main theorem. Throughout this section, we will fix $j \in [n]$.

Definition 5.1. Let $S_j = \{a, b | a_{-j}^T b_{-j} = 0\} \subset \{0,1\}^n \times \{0,1\}^n$

In particular, S_j is the set of pairs of strings a, b so that a and b are disjoint except possibly on the j^{th} coordinate. Note that $|S_j| = 4 \times 3^{n-1}$. Our proof will be based on a single data structure that will allow us to sample from either A_j conditioned on $A_{1\ldots j-1}, B_j = 0, B_{j+1\ldots n}, I = i$ or from B_j conditioned on $A_{1\ldots j-1}, A_j = 0, B_{j+1\ldots n}, I = i$. We now proceed to define this structure.

Definition 5.2. Let $E \subset S_j$ and let $K \subset [r]$, then we will denote $N^j_{E,K}$ as a 2×2 matrix, which for each $a'_j, b'_j \in \{0,1\}$ the (a'_j, b'_j) entry of $N^j_{E,K}$ is equal to

$$\sum_{(a,b)\in E, a_j = a'_j, b_j = b'_j} \sum_{i \in K} f_i(a) g_i(b)$$

Claim 5.3. $N^j_{S_j,[r]} = \frac{|S_j|}{4} \begin{bmatrix} s+1, s+1 \\ s+1, \quad s \end{bmatrix}$

Proof. Consider $(a,b) \in S_j$. Suppose, for example, that $a_j = 0, b_j = 0$ then this pair contributes $s+1$ to the $(0,0)$ entry in $N^j_{S_j,[r]}$. □

Next we define the data structure that will be crucial in our proof. The nodes will correspond to 2×2 matrices given by $N_{E,H}$:

Definition 5.4. The sampling tree T is a depth four tree in which:

- The root node corresponds to $E = S_j$ and $K = [r]$
- The first layer corresponds to nodes $E = S_j$, and $K = \{i\}$
- The second layer corresponds to $E = \{(a,b) \in S_j | a_1 = a'_1, ...a_{j-1} = a'_{j-1}, b_{j+1} = b'_{j+1}, ...b_n = b'_n\}$ and $K = \{i\}$
- The third layer corresponds to $E = \{(a,b) \in S_j | a_1 = a'_1, ...a_{j-1} = a'_{j-1}, a_j = a'_j, b_j = b'_j, b_{j+1} = b'_{j+1}, ...b_n = b'_n\}$ and $K = \{i\}$

Furthermore a node defined by E, K is connected to a node in a lower layer defined by E', K' if and only if $E' \subset E$ and $K' \subset K$.

Claim 5.5. *For any node $u \in T$, we have that $N_u = \sum_{v \in child(u)} N_v$.*

Proof. In fact, by construction if a node u corresponds to E, K then the children v either have $E' = E$ or $K' = K$ and in the first case, the children v define a partition of K (and vice-versa in the other case). □

An important point is that in any leaf node, the corresponding matrix N has only one non-zero entry. Next, we define three different methods for sampling the pair (a_j, b_j) using this tree:

Definition 5.6. Given a 2×2 matrix N, we will define $N(F)$ to be the sum of the first column of N, $N(G)$ to be the sum of the first row of N and $N(T)$ to be the total sum of the entries in N.

Using this definition, we can define F-sampling, G-sampling and T-sampling:

Definition 5.7. F-sampling generates a (a_j, b_j) pair as follows:

- Start at the root node.
- While the current node u is not a leaf node, choose a child v of u with probability $\frac{N_v(F)}{N_u(F)}$.
- Output the (a_j, b_j) pair corresponding to the non-zero entry in the leaf node at termination.

G-sampling and T-sampling are defined analogously with F replaced by G or T respectively. Also if instead we start the F sampling procedure at a first layer node with $K = \{i\}$ we will call it F, i-sampling and similarly for G and T.

The crucial observations are:

Observation 1. *The F-sampling and G-sampling procedures represent a faithful method to sample from the distribution A_j conditioned on $A_{1\ldots j-1}, B_j = 0, B_{j+1\ldots n}, I$ and B_j conditioned on $A_{1\ldots j-1}, A_j = 0, B_{j+1\ldots n}, I$ respectively.*

Observation 2. *The F, i-sampling and G, i-sampling procedures represent a faithful method to sample from the distribution A_j conditioned on $A_{1\ldots j-1}, B_j = 0, B_{j+1\ldots n}, I = i$ and B_j conditioned on $A_{1\ldots j-1}, A_j = 0, B_{j+1\ldots n}, I = i$ respectively.*

Note that the last choice in an F-sampling procedure represents choosing the random variable A_j conditioned on $A_{1\ldots j-1}, B_j = 0, B_{j+1\ldots n}, I$. And similarly the last choice in a G-sampling procedure represents choosing the random variable B_j conditioned on $A_{1\ldots j-1}, A_j = 0, B_{j+1\ldots n}, I$. Hence we can interpret the advantage through these processes:

Corollary 5.8. *$adv_j^A(i)$ is equal to $1 - H$, where H is the expected entropy of the last random choice when performing F, i-sampling. Also $adv_j^B(i)$ is equal to $1 - H$, where H is the expected entropy of the last random choice when performing G, i-sampling.*

Finally, we can use the T-sampling procedure to simulate either F-sampling or G-sampling:

Claim 5.9. *The following procedure is equivalent to F-sampling: Use T-sampling to obtain (a_j, b_j) and if $b_j = 0$, output (a_j, b_j) and otherwise restart the process. Also, the following procedure is equivalent to G-sampling: Use T-sampling to obtain (a_j, b_j) and if $a_j = 0$, output (a_j, b_j) and otherwise restart the process.*

We are now ready to prove the main lemma. We will let N denote the random variable corresponding to the matrix generated by running the T-sampling procedure until reaching the parent of a leaf. Similarly, let N_{00}, N_{10}, N_{10} and N_{11} denote the entries of this matrix. Then:

Lemma 5.10. *N is always a rank one matrix.*

Proof. Since N is the matrix corresponding to the parent of a leaf, we have that E corresponds to a fixed choice $a_{1\ldots j-1}$ for $A_{1\ldots j-1}$, $b_{j+1\ldots n}$ for $B_{j+1\ldots n}$ and $I = i$. Hence the (a'_j, b'_j) entry of N is $F_i(a'_j)G_i(b'_j)$, where $F_i(a'_j)$ is the sum of $f_i(a)$ over all a with $a_j = a'_j$ and $a^T_{j+1\ldots n} b_{j+1\ldots n} = 0$ and similarly $G_i(b'_j)$ is the sum of $g_i(b)$ over all b with $b_j = b'_j$ and $b^T_{1\ldots j-1} a_{1\ldots j-1} = 0$. □

In the exact case ($s = 0$), our proof is particularly simple. Since N is a rank one matrix and must have $N_{11} = 0$, there must be another zero in either the same row or column. Hence choosing an entry in N proportional to its value has entropy at most one bit, and this is already $\log 3 - 1$ bits smaller than the uniform distribution on the top-left, bottom-left and top-right entries, and this already implies an exponential lower bound in the exact case. Since our goal here is to give lower bounds even for large values of s that approach n, we must be more careful in accounting for the entropy lost when the entry N_{11} is allowed to be non-zero.

Definition 5.11. Let $adv(N) = \frac{N_{00}+N_{10}}{N_{00}+N_{10}+N_{10}+N_{11}}[1-H(\frac{N_{00}}{N_{00}+N_{10}})]+\frac{N_{00}+N_{01}}{N_{00}+N_{10}+N_{10}+N_{11}}[1-H(\frac{N_{00}}{N_{00}+N_{01}})]$.

Lemma 5.12. $\mathbf{E}_I[adv_j^A(i) + adv_j^B(i)] \geq \mathbf{E}_N[adv(N)]$

Proof. We can use Corollary 5.8 to get an interpretation of $\mathbf{E}_I[adv_j^A(i)]$ as $1 - H$, where H is the entropy of the last random choice in F-sampling. And using Claim 5.9 we have that $\mathbf{E}_N[\frac{N_{00}+N_{10}}{N_{00}+N_{10}+N_{10}+N_{11}}]$ is the probability that T-sampling results in a sample coupled with one from F-sampling. Hence, the expected entropy lost in the last step of F-sampling is lower bounded by the entropy lost in the last step of T-sampling if the sample satisfies $B_j = 0$. An identical argument for $adv_j^B(i)$ implies the lemma. □

Furthermore N is nonnegative so we can always write

$$N = (N_{01} + N_{10} + N_{00} + N_{11}) \begin{bmatrix} (\frac{1}{2}+\beta)(\frac{1}{2}-\gamma), (\frac{1}{2}+\beta)(\frac{1}{2}+\gamma) \\ (\frac{1}{2}-\beta)(\frac{1}{2}-\gamma), (\frac{1}{2}-\beta)(\frac{1}{2}+\gamma) \end{bmatrix}$$

Note that $\frac{N_{01}+N_{10}-N_{00}-N_{11}}{N_{01}+N_{10}+N_{00}+N_{11}} = 4\beta\gamma$.

Lemma 5.13. *$adv(N) = \Omega(\beta\gamma)$*

Proof. Using the bound $1 - H(\frac{1}{2}+x) \geq x^2$:

$$\begin{aligned} adv(N) &= (\frac{1}{2}-\gamma)H(\frac{1}{2}+\beta) + (\frac{1}{2}+\beta)H(\frac{1}{2}+\gamma) \\ &\geq (\frac{1}{2}-\gamma)\beta^2 + (\frac{1}{2}+\beta)\gamma^2 \end{aligned}$$

Clearly we can assume that $\beta\gamma > 0$ otherwise the lemma is trivial. We can prove the lemma through a simple case analysis:

- **Case:** $\beta > -1/3$ and $\gamma < 1/3$, then $adv(N) = \Omega(\beta^2 + \gamma^2) = \Omega(\beta\gamma)$.
- **Case:** $\beta \leq -1/3$, then $adv(N) \geq (\frac{1}{2}-\gamma)\beta^2$ because the other term is non-negative, and furthermore $\beta^2 = \Omega(1)$ and $(\frac{1}{2}-\gamma) = \Omega(1)$ since $\gamma \leq 0$.
- **Case:** $\gamma \geq 1/3$, then $adv(N) \geq (\frac{1}{2}+\beta)\gamma^2$ and again both terms are $\Omega(1)$ because $\beta \geq 0$.

□

Lemma 5.14. *$\mathbf{E}_N[\frac{N_{01}+N_{10}-N_{00}-N_{11}}{N_{01}+N_{10}+N_{00}+N_{11}}] = \frac{1}{4s+3}$*

Proof. The quantity $\mathbf{E}_N[\frac{N_{01}+N_{10}-N_{00}-N_{11}}{N_{01}+N_{10}+N_{00}+N_{11}}]$ is exactly the probability that under T-sampling the output has $(a_j, b_j) \in \{(0,1),(1,0)\}$ minus the probability that it has $(a_j, b_j) \in \{(0,0),(1,1)\}$. We can analyze this probability directly by considering the matrix corresponding to the root node, which has $(2s+2)\frac{|S_j|}{4}$ total weight in the $(1,0)$ and $(0,1)$ entries and has $(2s+1)\frac{|S_j|}{4}$ total weight in the $(0,0)$ and $(1,1)$ entries. Hence the difference of the probabilities of these two events is exactly $\frac{1}{4s+3}$. □

And using Lemma 5.12, Lemma 5.13 and Lemma 5.14 we conclude that $\mathbf{E}_I[adv_j^A(i) + adv_j^B(i)] = \Omega(\frac{1}{s+1})$. This concludes the proof of the main lemma.

Acknowledgements

We would like to thank Avi Wigderson for many helpful discussions at an early stage of this work.

References

[1] S. Arora, B. Bollobás and L. Lovász. Proving integrality gaps without knowing the linear program. *FOCS*, pp. 313-322, 2002.

[2] S. Arora, R. Ge, R. Kannan and A. Moitra. Computing a nonnegative matrix factorization - provably. *STOC*, pp. 145–162, 2012.

[3] A. Asadpour, M. Goemans, A. Madry, S. Oveis Gharan and A. Saberi. An $O(\log n/\log\log n)$-approximation algorithm for the asymmetric traveling salesman problem. *SODA*, pp. 379–389, 2010.

[4] E. Balas. Disjunctive programming and a hierarchy of relaxations for discrete optimization problems. *SIAM J. Algebraic Discrete Methods*, pp. 466-486, 1985.

[5] B. Barak, M. Braverman, X. Chen and A. Rao. How to compress interactive communication. *STOC*, pp. 67–76, 2010.

[6] Z. Bar-Yossef, T.S. Jayram, R. Kumar and D. Sivakumar. An information statistics approach to data stream and communication complexity. *JCSS*, pp. 702–732, 2004.

[7] G. Braun, S. Fiorini, S. Pokutta and D. Steurer. Approximation limits of linear programs (beyond hierarchies). *FOCS*, pp. 480–489, 2012.

[8] M. Braverman. Interactive information complexity. *STOC* pp. 505–524, 2012.

[9] M. Braverman and A. Rao. Information Equals Amortized Communication. *FOCS*, pp. 748–757, 2011.

[10] A. Chattopadhyay and T. Pitassi. The story of set disjointness. *SIGACT News*, pp. 59–85, 2010.

[11] T. Cover and J. Thomas. *Elements of Information Theory*. J. Wiley and Sons, 1991.

[12] J. Edmonds. Maximum matching and a polyhedron and $0,1$-vertices. *Journal of Research of the National Bureau of Standards*, pp. 125–130, 1965.

[13] U. Feige. Approximating maximum clique by removing subgraphs. *SIAM J. Discrete Math*, pp. 219–225, 2004.

[14] S. Fiorini, S. Massar, S. Pokutta, H. Tiwary and R. de Wolf. Linear vs semidefinite extended formulations: exponential separations and strong lower bounds. *STOC*, pp. 95–106, 2012.

[15] S. Fiorini, T. Rothvoß and H. Tiwary. Extended formulations for polygons. Arxiv, 2011.

[16] J. Håstad. Clique is hard to approximate within $n^{1-\epsilon}$ *Acta Mathematica*, pp. 105–142, 1999. Preliminary version in *FOCS* 1996.

[17] V. Kaibel. Extended formulations in combinatorial optimization. *Optima*, pp. 2-7, 2011.

[18] V. Kaibel, K. Pashkovich and D. Theis. Symmetry matters for the sizes of extended formulations. *IPCO*, pp. 135-148, 2010.

[19] B. Kalyanasundaram and G. Schnitger. The probabilistic communication complexity of set intersection. *SIAM J. Discrete Math*, pp. 545-557, 1992.

[20] I. Kerenidis, S. Laplante, V. Lerays, J. Roland and D. Xiao. Lower bounds on information complexity via zero-communication protocols and applications. *FOCS* 2012, to appear.

[21] S. Khot. Improved inapproximability results for MaxClique, chromatic number and approximate graph coloring. *FOCS*, pp. 600–609, 2001.

[22] E. Kushilevitz and N. Nisan. *Communication Complexity*. Cambridge University Press, 1997.

[23] K. Martin. Using separation algorithms to generate mixed integer model reformulations. *Operations Research Letters*, pp. 119-128, 1991.

[24] A. Moitra. An almost optimal algorithm for computing nonnegative rank. *SODA* 2013, to appear.

[25] S. Oveis Gharan, A. Saberi and M. Singh. A randomized rounding approach to the traveling salesman problem. *FOCS*, pp. 550–559, 2011.

[26] M. Padberg and M. Rao. Odd minimum cut-sets and b-matchings. *Mathematics of Operations Research*, pp. 67–80, 1982.

[27] A. Razborov. On the distributional complexity of disjointness. *Theoretical Computer Science*, pp. 385-390, 1992.

[28] T. Rothvoss. Some 0/1 polytopes need exponential size extended formulations. Arxiv, 2011.

[29] S. Vavasis. On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization*, pp. 1364-1377, 2009.

[30] L. Wolsey. Using extended formulations in practice. *Optima*, pp. 7-9, 2011.

[31] M. Yannakakis. Expressing combinatorial optimization problems by linear programs. *JCSS*, pp. 441–466, 1991. Preliminary version in *STOC* 1988.

[32] G. Ziegler. *Lectures on Polytopes*. Springer-Verlag, 1995.

[33] D. Zuckerman. Linear degree extractors and the inapproximability of max clique and chromatic number. *STOC*, pp. 681–690, 2006.

ECCC ISSN 1433-8092

http://eccc.hpi-web.de