## *Quantum entanglement, sum of squares, and the log rank conjecture.*

**Note:** These are very rough notes (essentially copied from the introduction of the corresponding paper). A better version should be up shortly, but in the meantime, reading the paper and looking at the lecture video is probably preferable.

*Entanglement* is one of the more mysterious and subtle phenomena of quantum mechanics. The formal definition is below Definition 2, but roughly speaking, a quantum state $\rho$ on two systems $A$ and $B$ is *entangled* if a quantum measurement of one system can effect the other system. A non-entangled state is called *separable*. This type of "spooky interaction at a distance" is responsible for many of the more counter-intuitive features of quantum mechanics. Entanglement is also used by all algorithms for quantum computers that obtain speedups over the best known classical algorithms, and it may be necessary for such speedups **?**.91.147902.

One of the ways in which the complexity of entanglement is manifested is that even given the full description of a quantum state $\rho$ as a density matrix, there is no known efficient algorithm for determining whether $\rho$ is entangled or not. Indeed, the best known algorithms take time which is *exponential* in the dimension of the state (which itself is exponential in the number of underlying qubits). This is in contrast to the classical case, where there is an efficient algorithm that given a probability distribution $\mu$ over a universe $A \times B$, can check whether or not $\mu$ is a *product distribution* by simply computing the rank of $\mu$ when viewed as a matrix.

Given the inherently probabilistic and noisy setting of quantum computing, arguably the right question is not to determine entanglement exactly, but rather to distinguish between the case that a state $\rho$ is separable, and the case that it is $\epsilon$-far from being separable, in the sense that there exists some *measurement* $M$ that accepts $\rho$ with probability $p$ but accepts every separable state with probability at most $p - \epsilon$. This problem is known as the *Quantum Separability Problem* with parameter $\epsilon$. Gharibian **?**, improving on Gurvits **?**, showed that this problem is NP hard when $\epsilon$ is inversely polynomial in the dimension of the state. Harrow and Montanaro Harrow and Montanaro [2013] showed that, assuming the Exponential Time Hypothesis, there is no $n^{o(\log n)}$ time algorithm for this problem for $\epsilon$ which is a small constant.

A tightly related problem, which is the one we focus on in this

paper, is the *Best Separable State (BSS)* problem.[1] In the BSS problem the input is a measurement $\mathcal{M}$ on a two part system and two numbers $1 \geq c > s \geq 0$ and the goal is to distinguish between the YES case that there is a separable state that $\mathcal{M}$ accepts with probability at least $c$ and the NO case that $\mathcal{M}$ accepts every separable state with probability at most $s$. In particular, certifying that a particular measurement $\mathcal{M}$ satisfies the NO case is extremely useful since it implies that $\mathcal{M}$ can serve as *entanglement witness* **?** (**?**), in the sense that achieving acceptance probability with $\mathcal{M}$ larger than $s$ certifies the presence of entanglement in a state. Such entanglement witnesses are used to certify entanglement in experiments and systems such as candidate computing devices **?**, and so having an efficient way to certify that they are sound (do not accept separable states) can be extremely useful.

Similarly to the quantum separability problem, the BSS problem is NP hard when $c - s = 1/poly(n)$ **?** and Harrow and Montanaro Harrow and Montanaro [2013] (Corollary 13(i)) show that (assuming the ETH) there is no $n^{o(\log n)}$ time algorithm for $BSS_{1,1/2}$. An outstanding open question is whether the Harrow and Montanaro [2013] result is *tight*: whether there is a quasi-polynomial time algorithm for $BSS_{c,s}$ for some constants $1 \geq c > s \geq 0$. This question also has a complexity interpretation. A measurement on a two part system can be thought of as a *verifier* (with hardwired input) that interacts with two provers. Requiring the state to be *separable* corresponds to stipulating that the two provers are not entangled. Thus it is not hard to see that an algorithm for $BSS_{c,s}$ corresponds to an algorithm for deciding all languages in the complexity class $QMA(2)$ of *two prover quantum Merlin Arthur* systems with corresponding completeness and soundness parameters $c$ and $s$ respectively. In particular a quasi-polynomial time algorithm for $BSS_{0.99,0.5}$ would imply that $QMA(2) \subseteq EXP$, resolving a longstanding problem in quantum complexity.[2]

In 2004, Doherty, Parrilo and Spedialieri Doherty et al. [2004] proposed an algorithm for the BSS problem based on the *Sum of Squares* semidefinite programming hierarchy **?** (Lasserre [2001]). It is not known whether this algorithm can solve the $BSS_{c,s}$ problem (for constants $c > s$) in quasi-polynomial time. However , Christandl and Yard **?**.1993683 showed that it runs in quasi-polynomial time when the measurement $\mathcal{M}$ is restricted to a special class of measurements known as *one-way local operations and classical communications* (1-LOCC). and Harrow Brandão and Harrow [2015] showed that similar performance for these types of measurements can be achieved by an algorithm based on searching on an appropriately defined $\epsilon$-net.

[1] Using the connection between optimization and separation oracles in convex programming, one can convert a sufficiently good algorithm for the search variant of one of these problems to the other. See Harrow and Montanaro [2013] (Sec. 4.2) for a thorough discussion of the relations between these and many other problems.

[2] For more on information on this problem and its importance, see the presentations in the recent workshop http://qma2016.quics.umd.edu/ that was dedicated to it.

## *Non quantum motivations*

The BSS problem is actually quite natural and well motivated from classical considerations. As we'll see in Section [sec:techniques] below, it turns out that at its core lies the following problem:

**1. Definition (Rank one vector in subspace problem).** Let $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$ and $\epsilon > 0$. The *$\epsilon$ rank one vector problem over* $\mathbb{F}$ is the task of distinguishing, given a linear subspace $\mathcal{W} \subseteq \mathbb{F}^{n^2}$, between the case that there is a nonzero rank one matrix $L \in \mathcal{W}$ and the case that $\|L - M\|_F \geq \epsilon \|L\|_F$ for every rank one $L$ and $M \in \mathcal{W}$.[3]

[3] For a $k \times m$ matrix $A$, we denote by $\|A\|_F$ its *Frobenius* norm, defined as $\sqrt{\sum_{i,j} |A_{i,j}|^2} = \mathrm{Tr}(AA^*)^{1/2}$, which is the same as taking the $\ell_2$ norm of the matrix when considered as an $km$-dimensional vector.

This is arguably a natural problem in its own right. While solving this problem exactly (i.e., determining if there is a rank one solution to a set of linear equations) is the same as the NP hard task of solving *quadratic equations*, it turns out that we can obtain non-trivial algorithmic results by considering the above notion of approximation. Indeed, our main result implies an $\exp(\tilde{O}(\sqrt{n}))$ time algorithm for this problem for any constant $\epsilon > 0$ in both the real and complex cases.

## *Our results*

In this work we give a $2^{\tilde{O}(\sqrt{n})}$ time algorithm for the $BSS_{1,s}$ problem for every constant $s < 1$. We now make the necessary definitions and state our main result.[4]

[4] For the sake of accessibility, as well as to emphasize the connections with non-quantum questions, we use standard linear algebra notation in this paper as opposed to Dirac's ket notation that is more common in quantum mechanics. A vector $u$ is a column vector unless stated otherwise, and $u^*$ denotes the complex conjugate transpose of the vector $u$. If $u$ is real, then we denote its transpose by $u^\top$. See the lecture notes **?** for a more complete coverage of separability and entanglement.

**2. Definition (Separable states).** A *quantum state* on a system of $m$ elementary states (e.g., a $\log m$-qubit register) is an $m \times m$ complex Hermitian matrix $\rho$ (known as a *density matrix*) such that $\mathrm{Tr}\,\rho = 1$. A quantum state $\rho$ is *pure* if it is of the form $\rho = ww^*$ for some unit vector $w \in \mathbb{C}^m$. Otherwise we say that $\rho$ is *mixed*. Note that every mixed state $\rho$ is a convex combination of pure states.

If $m = n^2$, and we identify $[m]$ with $[n] \times [n]$ then an $m$-dimension pure quantum state $\rho = ww^* \in \mathbb{C}^{m^2}$ is *separable* if the vector $w \in \mathbb{C}^m$ is equal to $uv^*$ for some $u, v \in \mathbb{C}^n$. A general state $\rho$ is *separable* if it is a convex combination of separable pure states. That is, $\rho = \mathbb{E}(uv^*)(uv^*)^*$ where the expectation is taken over a distribution supported over pairs of unit vectors $u, v \in \mathbb{C}^n$. A state that is not separable is called *entangled*.

A quantum *measurement operator* is an $m \times m$ complex Hermitian matrix $\mathcal{M}$ such that $0 \preceq \mathcal{M} \preceq I$. The probability that a measurement $\mathcal{M}$ accepts a state $\rho$ is $\mathrm{Tr}(\rho\mathcal{M})$.

**3. Theorem (Main result).** *For every $s < 1$, there is a $2^{\tilde{O}(\sqrt{n})}$ time algorithm, based on $\tilde{O}(\sqrt{n})$ rounds of the sos hierarchy, that on input an $n^2 \times n^2$ measurement operator $\mathcal{M}$, distinguishes between the following two cases:*

- *YES: There exists a separable state $\rho \in \mathbb{C}^{n^2 \times n^2}$ such that $\mathrm{Tr}(\rho\mathcal{M}) = 1$.*
- *NO: For every separable $\rho \in \mathbb{C}^{n^2 \times n^2}$, $\mathrm{Tr}(\rho\mathcal{M}) \leq s$*

To our knowledge, this algorithm is the first for this problem that beats the brute force bound of $2^{O(n)}$ time for general measurements.

Like the algorithms of Doherty et al. [2004] (**?**.1993683), our algorithm is based on the *sum of squares* SDP hierarchy, but we introduce new techniques for analyzing it that we believe are of independent interest. As we discuss in Section [sec:conclusions], it is a fascinating open question to explore whether our techniques can be quantitatively strengthened to yield faster algorithms and/or extended for other problems such as the 2 to 4 norm and small set expansion problems, that have been shown to be related to the BSS problem by Barak et al. [2012] (albeit in a different regime of parameters than the one we deal with in this work). As we remark below, this question seems related to other longstanding open questions in computer science and in particular to the *log rank conjecture* in communication complexity Lovász and Saks [1988].

[Imperfect completeness] [rem:perfect-completeness] We state our results for the case of perfect completeness for simplicity, but all of the proofs extend to the case of "near perfect completeness" where in the YES case we replace the condition $\mathrm{Tr}(\rho\mathcal{M}) = 1$ with the condition $\mathrm{Tr}(\rho\mathcal{M}) = 1 - \frac{1}{n}$ (see the proof of Theorem [thm:alg-analysis]). It is an interesting open problem to find out whether our results can extend to the setting where in the YES case $\mathrm{Tr}(\rho\mathcal{M}) = 1 - \epsilon$ for some absolute constant $\epsilon$. We conjecture that this is indeed the case.

While the natural setting for quantum information theory is the *complex numbers*, much of the power and interest already arises in the case of the real numbers, which is more natural for the sos algorithm (though it does have complex-valued generalization). Hence in this version of this paper we focus solely on the case that all operators, subspaces, matrices are *real*. We note that there is a natural mapping of an $n \times n$ complex matrix $A + iB$ (with real $n \times n$ matrices $A, B$)into the $2n \times 2n$ real matrix $\left(\begin{smallmatrix} A & B \\ -B & A \end{smallmatrix}\right)$. Note that a complex rank one decomposition $A + iB = (x + iy)(z + iw)^*$ for $x, y, z, w \in \mathbb{R}^n$ will translate into a rank two decomposition $\left(\begin{smallmatrix} x & y \\ -y & x \end{smallmatrix}\right)\left(\begin{smallmatrix} z & w \\ -w & z \end{smallmatrix}\right)^*$. We can use a higher dimensional version of our result (see **??** to find such

decompositions, though we defer the complete derivation to the full version of this paper.

## *Our techniques*

Our algorithm follows a recent paradigm of constructing rounding algorithms for the sum of squares sdp by considering its solutions as "pseudo distributions" **?**. These can be thought of as capturing the uncertainty that a computationally bounded solver has about the optimal solution of the given problem, analogously to the way that probability distributions model uncertainty in the classical information-theoretic Bayesian setting.

Somewhat surprisingly, our main tool in analyzing the algorithm are techniques that arose in proof of the currently best known upper bound for the *log rank conjecture* Lovász and Saks [1988]. This conjecture has several equivalent formulations, one of which is that every $N \times N$ matrix $A$ with Boolean (i.e., 0/1) entries and rank at most $n$, contains a submatrix of size at least $2^{-\operatorname{poly}\log(n)}N \times 2^{-\operatorname{poly}\log(n)}N$ that is of rank one.[5] The best known bound on the log rank conjecture is by Lovett Lovett [2014] who proved that every such matrix contains a submatrix of size at least $2^{-\tilde{O}(\sqrt{n})}N \times 2^{-\tilde{O}(\sqrt{n})}N$.

[5] The original formulation of the log rank conjecture is that every such matrix has communication complexity at most $\operatorname{poly}\log(n)$, and Nisan and Wigderson Nisan and Wigderson [1994] showed that this is equivalent to the condition that such matrices contains a monochromatic submatrix of the above size. Every monochromatic submatrix is rank one, and every rank one submatrix of size $s \times s$ of a Boolean valued matrix contains a monochromatic submatrix of size at least $\frac{s}{2} \times \frac{s}{2}$.

Our algorithm works by combining the following observations:

1. Lovett's proof can be generalized to show that *every* $N \times N$ rank $n$ real (or complex) matrix $A$ (not necessarily with Boolean entries) contains a $2^{-\tilde{O}(\sqrt{n})}N \times 2^{-\tilde{O}(\sqrt{n})}N$ submatrix that is *close* to rank one in Frobenius norm.

2. If $\mu$ is an *actual* distribution over solutions to the sos program for the BSS problem on dimension $n$, then we can transform $\mu$ into an $N \times N$ rank $n$ matrix $A = A(\mu)$ such that extracting an approximate solution from $A$ in time $2^{\tilde{O}(k)}$ can be done if $A$ contains an approximately rank one submatrix of size at least $2^{-k}N \times 2^{-k}N$.

3. Moreover all the arguments used to establish steps 1 and 2 above can be encapsulated in the sum of squares framework, and hence yield an algorithm that extracts an approximately optimal solution to the BSS problem from a degree $\tilde{O}(\sqrt{n})$ pseudo-distribution $\mu$ that "pretends" to be supported over exact solutions.

Thus, even though in the sos setting there is no actual distribution $\mu$, and hence no actual matrix $A$, we can still use structural results on this "fake" (or "pseudo") matrix $A$ to obtain an *actual* rounding

algorithm. We view this as a demonstration of the power of the "pseudo distribution" paradigm to help in the discovery of new algorithms, that might not seem as natural without placing them in this framework.

### *Rounding from rank one reweighings*

We now give a more detailed (yet still quite informal) overview of the proof. As mentioned above, we focus on the case that the $n^2 \times n^2$ measurement matrix $\mathcal{M}$ is *real* (as opposed to *complex*) valued.

Let $\mathcal{W} \subseteq \mathbb{R}^{n^2}$ be the subspace of vectors $X$ such that $X^\top \mathcal{M} X = \|X\|^2$ (this is a subspace since $\mathcal{M} \preceq I$ and hence $\mathcal{W}$ is the eigenspace of $\mathcal{M}$ corresponding to the eigenvalue 1). We pretend that the sos algorithm yields a distribution $\mu$ over rank one matrices of the form $X = uv^\top$ such that $X \in \mathcal{W}$. When designing a rounding algorithm, we only have access to *marginals* of $\mu$, of the form $\mathbb{E}_\mu f(X)$ for some "simple" function $f$ (e.g., a low degree polynomial). We need to show that we can use such "simple marginals" of $\mu$ to extract a single rank one matrix $u_0 v_0^\top$ that has large projection into $\mathcal{W}$.

We start with the following simple observation:

**4. Lemma.** *If $\mu$ is a distribution over matrices $X$ in a subspace $\mathcal{W} \subseteq \mathbb{R}^{n^2}$ such that the expectation $\mathbb{E}_\mu X$ is approximately rank one, in the sense that $\|L - \mathbb{E}_\mu X\|_F \leq \epsilon \|L\|_F$ for some rank one matrix $L$, then $\mathrm{Tr}(\mathcal{M}\rho) \geq 1 - 2\epsilon^2$ where $\rho$ is the pure separable state $\rho = LL^\top / \|L\|_F^2$.*

Since $\mu$ is supported over matrices in $\mathcal{W}$, $\mathbb{E}_\mu X$ is in $\mathcal{W}$. But this means that the $\ell_2$ (i.e., Frobenius) norm distance of $L$ to the subspace $\mathcal{W}$ is at most $\epsilon \|L\|_F$. Since $\mathrm{Tr}(XX^\top \mathcal{M}) = \mathrm{Tr}(X^\top \mathcal{M} X) = \|X\|_F^2$ for every $X \in \mathcal{W}$, the value $\mathrm{Tr}(LL^\top M)$ will be at least as large as the norm squared of the projection of $L$ to $\mathcal{W}$.

In particular this means that if we were lucky and the condition of Lemma 4's statement occurs, then it would be trivial for us to extract from the expectation $\mathbb{E}_\mu X$ (which is a very simple marginal) a rank one matrix that is close to $\mathcal{W}$, and hence achieves probability $1 - \epsilon$ in the measurement $\mathcal{M}$. Note that even if every matrix in the support of $\mu$ has unit norm, the matrix $L$ could be of significantly smaller norm. We just need that there is some dimension-one subspace on which the cancellations among these matrices are significantly smaller than the cancellations in the rest of the dimensions.

Of course there is no reason we should be so lucky, but one power that the marginals give us is the ability to *reweigh* the original dis-

tribution $\mu$. In particular, for every "simple" non-negative function $\zeta : \mathbb{R}^{n^2} \to \mathbb{R}_+$, we can compute the marginal $\mathbb{E}_{\mu_\zeta} X$ where $\mu_\zeta$ is the distribution over matrices where $\mathbb{P}_{\mu_\zeta}[X]$ (or $\mu_\zeta(X)$ for short) is proportional to $\zeta(X)\mu(X)$. A priori in the degree $k$ sos algorithm we are only able to reweigh using functions $\zeta$ that are polynomials of degree at most $k$, but for the purposes of this overview, let us pretend that we can reweigh using any function that is not too "spiky" and make the following definition:

Let $\mu$ be a probability distribution. We say that a probability distribution $\mu'$ is a *k-spike reweighing* of $\mu$ if $\Delta_{KL}(\mu'\|\mu) \leq k$ where $\Delta_{KL}(\mu'\|\mu)$ denotes the Kullback-Leibler divergence of $\mu'$ and $\mu$, defined as $\mathbb{E}_{X\sim\mu'} \log(\mu'(X)/\mu(X))$.

Thus at least on a "moral level", the following theorem should be helpful for proving our main result:

**5. Theorem (Rank one reweighing).** *Let $\mu$ be any distribution over rank one $n \times n$ matrices and $\epsilon > 0$. Then there exists an $\sqrt{n}\,\mathrm{poly}(1/\epsilon)$-spike reweighing $\mu'$ of $\mu$ and a rank one matrix $L$ such that*

$$\|L - \tilde{\mathbb{E}}_{\mu'} X\|_F \leq \epsilon \|L\|_F \tag{1}$$

One of the results of this paper is a proof of Theorem 5 (see **??**). It turns out that this can be done using ideas from the works on the log rank conjecture.

## *From monochromatic rectangles to rank one reweighings*

What does Theorem Theorem 5 has to do with the log rank conjecture? To see the connection let us imagine that the distribution $\mu$ is *flat* in the sense that it is a uniform distribution over rank one matrices $\{u_1 v_1^\top, \ldots, u_N v_N^\top\}$ (this turns out to be essentially without loss of generality) and consider the $n \times N$ matrices $U$ and $V$ whose columns are $u_1, \ldots, u_N$ and $v_1, \ldots, v_N$ respectively. The $n \times n$ matrix $\tilde{\mathbb{E}}_\mu u_i v_i^\top$ is proportional to $UV^\top$. This matrix has the same spectrum (i.e., singular values) as the $N \times N$ matrix $U^\top V$. Hence, $UV^\top$ is close to a rank one matrix if and only if $U^\top V$ is, since in both cases this happens when the square of the top singular value dominates the sum of the squares of the rest of the singular values. Now a flat distribution $\mu'$ with $\Delta_{KL}(\mu'\|\mu) \leq k$ corresponds to the uniform distribution over $\{u_i v_i^\top\}_{i\in I}$ where $I \subseteq [N]$ satisfies $|I| \geq 2^{-k}N$. We can see that $\mathbb{E}_{\mu'} u_i v_i^\top$ will be approximately rank one if and only if the submatrix of $U^\top V$ corresponding to $I$ is approximately rank one. Using these

ideas it can be shown that Theorem [thm:rank-one-reweighing] is equivalent to the following theorem:[6]

**6. Theorem (Rank one reweighing—dual formulation).** *Let $A$ be any $N \times N$ matrix of rank at most $n$. Then there exists a subset $I \subseteq [N]$ with with $|I| \geq \exp(-\sqrt{n}\,\mathrm{poly}(1/\epsilon))N$ and a rank one matrix $L$ such that*

$$\|L - A_{I,I}\|_F \leq \epsilon \|L\|_F \tag{2}$$

*where $A_{I,I}$ is the submatrix corresponding to restricting the rows and columns of $A$ to the set $I$.*

One can think of Theorem 6 as an approximate and robust version of Lovett's result Lovett [2014] mentioned above. Lovett showed that every $N \times N$ matrix of rank $n$ with *Boolean* entries has a $2^{-\tilde{O}(\sqrt{n})}N \times 2^{-\tilde{O}(\sqrt{n})}N$ submatrix that is of exactly rank 1. We show that the condition of Booleanity is not needed if one is willing to relax the conclusion to having a submatrix that is only *approximately* rank 1. It is of course extremely interesting in both cases whether the bound of $\tilde{O}(\sqrt{n})$ can be improved further, ideally all the way to $polylog(n)$. In the Boolean setting, such a bound might prove the log rank conjecture,[7] while in our setting such a bound (assuming it extends to "pseudo matrices") would yield a quasipolynomial time algorithm for BSS, hence showing that $QMA(2) \subseteq EXP$. It can be shown that as stated, Theorem [thm:rank-one-reweighing] is tight. However there are different notions of being "close to rank one" that could be useful in both the log-rank and the quantum separability setting, for which there is hope to obtain substantially improved quantitative bounds. We discuss some of these conjectural directions in **??**.

[6] To show this formally we use the fact that by Markov, every distribution $\mu'$ with $\Delta_{KL}(\mu' \| U_{[N]}) = \log N - H(\mu') = k$ is $\epsilon$-close to a distribution with min entropy $\log N - O(k/\epsilon)$ and every distribution of the latter type is a convex combination of flat distributions of support at least $N2^{-O(k/\epsilon)}$.

[7] We note a caveat that this depends on the notion of "approximate" used. Gavinsky and Lovett Gavinsky and Lovett [2014] showed that to prove the log rank conjecture it suffices to find a in a rank $n$ Boolean matrix a rectangle of measure $\exp(-\mathrm{polylog}(n))$ that is *nearly monochromatic* in the sense of having a $1 - 1/O(n)$ fraction of its entries equal. In this paper we are more concerned with rectangles whose distance to being rank one (or monochromatic) is some $\varepsilon > 0$ that is only a small constant or $1/\mathrm{polylog}(n)$. [fn:approx-monochromatic]

## *Overview of proof*

In the rest of this technical overview, we give a proof sketch of Theorem 6 and then discuss how the proof can be "lifted" to hold in the setting of sum of square pseudo-distributions. The condition that a matrix $A$ is of rank $n$ is the same as that $A = UV^\top$ where $U, V$ are two $n \times N$ matrices with columns $u_1, \ldots, u_N$ and $v_1, \ldots, v_N$ respectively (i.e., $A_{i,j} = \langle u_i, v_j \rangle$ for all $i, j \in [N]$). We will restrict our attention to the case that all the columns of $U$ and $V$ are of unit norm. (This restriction is easy to lift and anyway holds automatically in our intended application.) In this informal overview, we also restrict attention to the *symmetric* case, in which $A = A^\top$ and can be written as $A = UU^\top$ and also assume that $U$ is *isotropic*, in the sense that $\mathbb{E}_{i \in [N]} u_i u_i^\top = \frac{1}{n}\,\mathrm{Id}$.

Our inspiration is Lovett's result Lovett [2014] which establishes a stronger conclusion for Boolean matrices. In particular, our proof follows Rothvoß's proof Rothvoß [2014] of Lovett's theorem, though the non-Boolean setting does generate some non-trivial complications. The $N \times N$ matrix $A$ satisfies that $A_{i,j} = \langle u_i, u_j \rangle$. An equivalent way to phrase our goal is that we want to find a subset $I \subseteq [N]$ over the indices such that:

<dl> <dt>**(i)**</dt> <dd>

$|I| \geq \exp(-\tilde{O}(\sqrt{n}))N$.

</dd> <dt>**(ii)**</dt> <dd>

If $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_n$ are the eigenvalues of $\mathbb{E}_{i \in I} u_i u_i^\top$ then $\epsilon^2 \lambda_1^2 \geq \sum_{j=2}^n \lambda_j^2$

</dd> </dl>

We will chose the set $I$ *probabilistically* and show that **(i)** and **(ii)** above hold in *expectation*. It is not hard to use standard concentration of measure bounds to then deduce the desired result but we omit these calculations from this informal overview.

Our initial attempt for the choice of $I$ is simple, and is directly inspired by Rothvoß [2014]. We choose a random standard Gaussian vector $g \in N(0, \frac{1}{n} \mathrm{Id})$ (i.e., for every $i$, $g_i$ is an independent standard Gaussian of mean zero and variance $1/n$). We then define $I_g = \{i : \langle g, u_i \rangle \geq \sqrt{k/n}\}$ where $k = \tilde{O}(\sqrt{n})$ is a parameter to be chosen later. Since $u_i$ is a unit vector, $\langle g, u_i \rangle$ is a Gaussian of variance $1/n$, and so for every $i$, the probability that $i \in I_g$ is $\exp(-O(k))$ hence satisfying **(i)** in expectation.

The value $\lambda_1$ of $\mathbb{E}_{i \in I} u_i u_i^\top$ will be at least $\Omega(k/n)$ in expectation. Indeed, we can see that the Gaussian vector $g$ that we choose (which satisfies $\|g\|^2 = 1 \pm o(1)$ with very high probability) will satisfy that $g^\top \left( \mathbb{E}_i u_i u_i^\top \right) g = \mathbb{E}_{i \in I_g} \langle u_i, g \rangle^2 \geq k/n$ and hence in expectation the top eigenvalue of $\mathbb{E}_i u_i u_i^\top$ will be at least $(1 - o(1))k/n$.

So, if we could only argue that in expectation it will hold that $\sum_{j=1}^n \lambda_j^2 \ll k^2/n^2 = \mathrm{polylog}(n)/n$ then we'd be done. Alas, this is not necessarily the case. However, if this does fail, we can see that we have made progress, in the sense that by restricting to the indices in $I$ we raised the Frobenius norm of $\mathbb{E} u_i u_i^\top$ from the previous value of $1/n$ (under the assumption that $U$ was isotropic) to $polylog(n)/n$. Our idea is to show that this holds in general: we can select a Gaussian vector $g$ and define the set $I_g$ as above such that by restricting to the indices in $I_g$ we either get an approx rank one matrix or we

increase the Frobenius norm of our expectation matrix by at least an $(1+\epsilon)$ factor for an appropriately chosen $\epsilon > 0$. Since the latter cannot happen more than $\log n/\epsilon$ times, the final set of indices still has measure $\exp(-\tilde{O}(\sqrt{n}))$.

In further rounds, if our current set of indices is $I$ and the matrix (after subtracting from each vector $u_i$ its expectation) $U_I = \mathbb{E}_{i\in I} u_i u_i^\top = \sum_{j=1}^n \lambda_j v_j v_j^\top$ is not approximately rank one, then rather than choosing $g$ as a standard Gaussian, we choose it from the distribution $N(0, U_I)$ where we use $U_I$ as the covariance matrix. The expected norm of $g$ is simply $\mathrm{Tr}(U_I)$ which equals 1. For every $i$, the random variable $\langle u_i, g\rangle$ is a Gaussian with mean zero and variance $\sum_{j=1}^n \langle u_i, v_j\rangle \lambda_j$. But for every $j$ in expectation over $i$, $\mathbb{E}\langle u_i, v_j\rangle^2 = \lambda_j$ and so it turns out that we can assume that this random variable has variance $\sum \lambda_j^2 = \|U_I\|_F^2$.

This means that if we choose $I' = \{i \in I : \langle u_i, g\rangle \geq \sqrt{k}\|U_I\|_F\}$ we get a subset of $I$ with measure $\exp(-O(k))$. But now the new matrix $U_{I'} = \mathbb{E}_{i\in I'} u_i u_i^\top$ will have an eigenvalue of at least $k\|U_I\|_F^2$ magnitude which is much larger than $\|U_I\|_F$ since we chose $k \gg \sqrt{n}$. Hence $U_{I'}$ has significantly larger Frobenius norm than $U_I$.

The above arguments can be made precise, and we do so in Section [sec:structure-thm].

## *Rectangle lemma for pseudo-distributions*

The above is sufficient to show that given $N \times n$ matrices $U = (u_1|\cdots|u_N)$ and $V = (v_1|\cdots|v_n)$ (which we view as inducing a distribution over rank one matrices by taking $u_i v_i^\top$ for a random $i$), we can condition on a not too unlikely event (of probability $\exp(-\tilde{O}(\sqrt{n}))$ to obtain that $\mathbb{E}\, u_i v_i^\top$ is roughly rank one. But in the sos setting we are *not* given such matrices. Rather we have access to an object called a "pseudo-distribution" $\mu$ which we behaves to a certain extent as if it is such a distribution, but for which it is not actually the case. In particular, we are not able to sample from $\mu$, or condition it on arbitrary events, but rather only compute $\mathbb{E}_\mu f(X)$ for polynomials $f$ of degree at most $\tilde{O}(\sqrt{n})$, and even these expectations are only "pseudo expectations" in the sense that they do not need to correspond to any actual probability distribution.

To lift the arguments above to the sos setting, we need to first show that if $\mu$ was an actual distribution, then we could perform all of the above operations using only access to $\tilde{O}(\sqrt{n})$ degree moments of $\mu$. Then we need to show that our *analysis* can be captured by the

degree $\tilde{O}(\sqrt{n})$ sos proof systems. Both these steps, which are carried out in Section [??] of this paper, are rather technical and non-trivial, and we do not describe them in this overview.

For starters, we need to move from *conditioning* a probability distribution to *reweighing* it. All of our conditioning procedures above had the form of restricting to $i$'s such that $\xi(i) \geq \sqrt{k}$ where $\xi(i)$ was probabilistically chosen so that for every $i$ $\xi(i)$ is a a mean zero and standard deviation one random variable satisfying $\mathbb{P}[\xi(i) = \ell] = \exp(-\Theta(\ell^2))$. We replace this conditioning by *reweighing* the distribution $i$ with the function $\zeta(i) = \exp(\sqrt{k}\xi(i))$. Note that iterative conditioning based on functions $\xi_1, \ldots, \xi_t$ can be replaced with reweighing by the product function $\zeta_1, \ldots, \zeta_t$. We then show that these $\zeta_j$ functions can be approximated by polynomials of $\tilde{O}(k)$ degree.

The arguments above allow us to construct a rounding algorithm that at least makes sense syntactically, in the sense that it takes the $\tilde{O}(\sqrt{n})$ degrees moments of $\mu$ and produces a rank one matrix that is a candidate solution to the original matrix. To analyze this algorithm, we need to go carefully over our analysis before, and see that all the arguments used can be embedded in the sos proof system with relatively low degree. Luckily we can rely on the recent body of works that establishes a growing toolkit of techniques to show such embeddings **?**.

## *References*

Boaz Barak, Fernando G. S. L. Brandão, Aram Wettroth Harrow, Jonathan A. Kelner, David Steurer, and Yuan Zhou. Hypercontractivity, sum-of-squares proofs, and their applications. In *STOC*, pages 307–326. ACM, 2012.

Fernando G. S. L. Brandão and Aram Wettroth Harrow. Estimating operator norms using covering nets. *CoRR*, abs/1509.05065, 2015.

Andrew C Doherty, Pablo A Parrilo, and Federico M Spedalieri. Complete family of separability criteria. *Physical Review A*, 69(2): 022308, 2004.

Dmitry Gavinsky and Shachar Lovett. En route to the log-rank conjecture: New reductions and equivalent formulations. In *ICALP (1)*, volume 8572 of *Lecture Notes in Computer Science*, pages 514–524. Springer, 2014.

Aram Wettroth Harrow and Ashley Montanaro. Testing product states, quantum merlin-arthur games and tensor optimization. *J. ACM*, 60 (1):3, 2013.

Jean B. Lasserre. An explicit exact SDP relaxation for nonlinear 0-1 programs. In *IPCO*, volume 2081 of *Lecture Notes in Computer Science*, pages 293–303. Springer, 2001.

László Lovász and Michael E. Saks. Lattices, möbius functions and communication complexity. In *FOCS*, pages 81–90. IEEE Computer Society, 1988.

Shachar Lovett. Communication is bounded by root of rank. In *STOC*, pages 842–846. ACM, 2014.

Noam Nisan and Avi Wigderson. On rank vs. communication complexity. In *FOCS*, pages 831–836. IEEE Computer Society, 1994.

Thomas Rothvoß. A direct proof for lovett's bound on the communication complexity of low rank matrices. *CoRR*, abs/1409.6366, 2014.