Tensor Prediction, Rademacher Complexity and Random 3-XOR

Boaz Barak * Ankur Moitra †

April 3, 2015

Abstract

We study the tensor prediction problem, where the goal is to accurately predict the entries of a low rank, third-order tensor (with noise) given as few observations as possible. We give algorithms based on the sixth level of the sum-of-squares hierarchy that work with $m = \tilde{O}(n^{3/2})$ observations, and we complement our result by showing that any attempt to solve tensor prediction with $m = O(n^{3/2-\epsilon})$ observations through the sum-of-squares hierarchy needs $\Omega(n^\epsilon)$ rounds and consequently would run in moderately exponential time. In contrast, information theoretically $m = \tilde{O}(n)$ observations suffice.

This work is part of a broader agenda of studying computational vs. statistical tradeoffs through the sum-of-squares hierarchy. In particular, for linear inverse problems (such as tensor prediction) the natural sum-of-squares relaxation gives rise to a sequence of norms. Our approach is to characterize their Rademacher complexity. Moreover, both our upper and lower bounds are based on connections between this, the task of strongly refuting random 3-XOR formulas, and the resolution proof system.

---

* Microsoft Research New England, Cambridge, MA.
† Massachusetts Institute of Technology. Department of Mathematics and the Computer Science and Artificial Intelligence Lab. Email: moitra@mit.edu. This work is supported in part by a grant from the MIT NEC Corporation and a Google Research Award.
1 Introduction

One of the major recent advances in machine learning has been the development of a broad range of algorithms for solving linear inverse problems based on convex optimization. These are algorithms that recover an unknown object \( M \in \mathbb{R}^N \) from \( m \) linear measurements (where \( m \ll N \)) based on the assumption that \( M \) possesses some special structure. Perhaps the most famous examples are compressed sensing and matrix completion. In the former, it was shown that an \( n \)-dimensional, \( k \)-sparse vector can be recovered from \( O(k \log n/k) \) random linear measurements \([26, 25, 34]\). In the latter, it was shown that an \( n \times n \) matrix that is incoherent and has rank \( r \) can be recovered from observing \( O(nr \log^2 n) \) of its entries \([24, 27, 67]\).

The general form of a linear inverse problem can be thought of as attempting to solve the following optimization problem:

\[
(P_0) \quad \min_{X} \text{obj}_0(X) \quad \text{s.t.} \quad L_i(X) = L_i(M) \quad \text{for} \quad i = 1, 2, \ldots, m
\]

where each \( L_i(\cdot) \) is a known linear function, corresponding to an observation of \( M \). In compressed sensing, we choose \( \text{obj}_0(X) = \|X\|_0 \) (i.e., the number of nonzero entries of \( X \)) and our goal is to find the sparsest solution to a given linear system. In matrix completion, we choose \( \text{obj}_0(X) = \text{rank}(X) \) and our goal is to find the lowest rank matrix that agrees with the entries we have observed. The challenge is that for these and many other linear inverse problems, the natural formulation is computationally hard in the worst-case. However, it turns out that in many cases one can obtain a relaxation

\[
(P_1) \quad \min_{X} \text{obj}_1(X) \quad \text{s.t.} \quad L_i(X) = L_i(M) \quad \text{for} \quad i = 1, 2, \ldots, m
\]

which can be solved efficiently, and for many settings of interest succeeds in recovering \( M \) exactly. The results for compressed sensing are achieved by setting \( \text{obj}_1(X) = \|X\|_1 \). And the results for matrix completion are achieved by setting \( \text{obj}_1(X) = \|X\|_* \), where \( \| \cdot \|_* \) is the nuclear norm which is given by the sum of the singular values. The resulting algorithms for both of these problems are computationally efficient, and succeed in recovering \( M \) from a number of measurements that is close to the information theoretic minimum. In contrast, there are some problems where the number of measurements for the best known efficient recovery is significantly more than the information theoretic minimum. In this paper, we study perhaps the most natural such problem — the tensor prediction problem, which we define in the next subsection.

We remark that there have been many further successes in using simple convex programs to solve linear inverse problems, such as for phase retrieval \([20]\), sparse principal component analysis \([22]\), super-resolution \([21]\) and compressed sensing off-the-grid \([73]\). In fact, all of these algorithms can be thought of as using the following general recipe \([30]\): First, define \( \mathcal{A} \) to be the convex hull of all \( X \) satisfying \( \text{obj}_0(X) = 1 \) as well as some normalization condition. Then, solve the optimization problem

\[
(A) \quad \min \|X\|_{\mathcal{A}} \quad \text{s.t.} \quad L_i(X) = L_i(M) \quad \text{for} \quad i = 1, 2, \ldots, m
\]

where \( \|X\|_{\mathcal{A}} \) is the minimum \( \alpha \) such that \( X/\alpha \in \mathcal{A} \) (the norm \( \| \cdot \|_{\mathcal{A}} \) is often called the atomic norm). All of the above algorithms for solving linear inverse problems can thought of as being derived from this common framework\(^2\) and operating in a setting where \( (a) \) the atomic norm is easy to compute and \( (b) \) it is possible to recover the underlying object \( M \) efficiently, using close to the information theoretic minimum number of measurements (or observations). In contrast, in our setting the atomic norm is hard to compute and it turns out that even powerful hierarchies of approximations to it will not be able to approach the information theoretic limits in polynomial time.

We will show that using the optimization problem \((A)\), it is possible to solve the tensor prediction problem for an \( n \times n \times n \) tensor that is incoherent and has rank \( r \) with \( m = O(nr) \) observations, but the resulting algorithm runs in exponential time. Our main algorithmic result is that the sixth level of the sum-of-squares hierarchy \([65, 59]\) can solve the tensor prediction problem with \( m = O(n^{3/2}r) \) observations. Our main negative result is that for \( r = O(1) \) and for any \( c > 0 \), if \( m = O(n^{3/2-c}) \) then the \( n^{2c} \) level fails to give any non-trivial prediction. Hence any attempt to solve tensor prediction with fewer observations through the

\(^1\)For every \( p \geq 1 \), we use \( \|X\|_p \) to denote the \( L_p \)-norm of \( X \), defined as \( (\sum_i |X_i|^p)^{1/p} \).

\(^2\)This is because the \( L_1 \) norm is the atomic norm corresponding to the convex hull of unit \( L_2 \) norm, 1-sparse vectors, while the nuclear norm is the atomic norm corresponding to the convex hull of all unit Frobenius norm, rank one matrices.
sum-of-squares hierarchy would run in moderately exponential time. Both our upper and lower bounds are obtained by connecting the tensor prediction problem to the task of refuting random constraint satisfaction problems, and the resolution proof system. This connection is made via bounding the Rademacher complexity of the sequence of norms that arise from the sum-of-squares hierarchy. We now introduce the tensor prediction problem in more detail.

1.1 Tensor Prediction

Here we introduce the tensor prediction problem, where there is an unknown third-order tensor that is close to low-rank and we would like to accurately predict its entries. Our goal is to do so while making use of as few observations as possible. In particular, let

\[ M = \sum_{\ell=1}^{r} a_\ell \otimes a_\ell \otimes a_\ell + H \]  

be a third order tensor\(^3\) that has a low-rank component of rank \(r\) and where \(H \in \mathbb{R}^{n^3}\) is tensor whose entries represent noise in our observations. In particular, its \((i,j,k)\)th entry will be denoted by \(h_{i,j,k}\). All of our results will generalize to asymmetric tensors, and to higher-order tensors. See Appendix A and Appendix B.

In our model, we will assume that each vector \(a_\ell\) is a unit vector in \(\mathbb{R}^n\) and that each vector \(a_\ell\) has all of its entries bounded by \(\frac{C}{\sqrt{n}}\) for some parameter \(C\).\(^4\) This should be thought of as being in the spirit of the usual incoherence assumptions [24]. Let \(\Omega \subset [n] \times [n] \times [n]\) represent the locations of the entries that we observe and \(\eta = \frac{1}{n^3} \sum_{(i,j,k) \in \Omega} |h_{i,j,k}|\) be the average noise in a typical observation. In our model, we will assume that our observations are chosen uniformly at random and without repetition, and throughout this paper we set \(|\Omega| = m\). We let \(T = \sum_{\ell=1}^{r} a_\ell \otimes a_\ell \otimes a_\ell\) be the low rank component of \(M\). The goal in the tensor prediction problem is to approximately recover \(T\) from as few observations as possible.

Further Motivation. Tensor completion is a natural generalization of matrix completion, as well as a canonical example of a hard linear inverse problem, and thus understanding its complexity is interesting as a mathematical problem. But it is also a basic learning problem in its own right. There are many types of data (e.g., clinical tests) where each observation is described by three or more attributes, it which case it makes sense to organize these observations into a tensor and to try to predict the missing entries. Moreover the natural assumption is that these tensors are close to low rank, and this is what makes it possible to predict the missing entries accurately, from few observations (just as in the matrix case). See Section 1.4 for further discussion and its relation to tensor decomposition.

A few remarks are in order. First, note that the usual tensor completion problem is simultaneously weaker in that it assumes \(\eta = 0\) and stronger in that it asks to exactly recover the original low rank component. The prediction task is arguably better suited to many realistic applications, where the tensors are only approximately low rank. We also remark that unlike some prior works on tensor completion [56], we do not require the vectors \(a_\ell\)’s to be orthogonal to one another, or even to have small pairwise inner-products. The factors can have arbitrary overlap, and this is a crucial aspect of the generality of our results.

Now that we have described the model, let us formally describe our prediction goal. We will work within the framework of statistical learning theory, where our goal is to output a hypothesis \(X\) that has low error:

\[ \text{err}(X) = \frac{1}{n^3} \sum_{i,j,k} \left| X_{i,j,k} - M_{i,j,k} \right| \]

We remark that if \(r = O(1)\) then the maximum entry in \(M\) is \(O(1/n^{3/2})\) and so the estimate \(X = 0\) already achieves error \(O(1/n^{3/2})\). And our goal will be to understand when we can achieve asymptotically smaller error. This type of guarantee has a number of appealing properties, such as allowing us to obtain a \((1 \pm o(1))\) multiplicative factor approximation for almost all entries of a typical tensor (see Corollary 1.2).

---

\(^3\)Recall that for two vectors \(a \in \mathbb{R}^n, b \in \mathbb{R}^m\), \(a \otimes b\) denotes their tensor product which is a vector in \(\mathbb{R}^{nm}\) such that \((a \times b)_{i,j} = a_i b_j\) for all \(i, j\).

\(^4\)We think of \(C\) as a large constant. However this parameter can be chosen to be polylogarithmic at the expense of increasing the sample complexity in our bounds by comparable polylogarithmic factors.
Our algorithms will be based on solving the following optimization problem:

$$(K) \quad \min_{X} \|X\|_K \text{ s.t. } \frac{1}{m} \sum_{(i,j,k) \in \Omega} |X_{i,j,k} - M_{i,j,k}| \leq 2\eta$$

where $\| \cdot \|_K$ is some norm. We will often use $K_k$ to emphasize that we are interested in a sequence of norms and that $K_k$ comes from the $k$th level of the sum-of-squares relaxation to the atomic norm. In the above optimization problem, our goal is to find a low-complexity hypothesis $X$ that has error at most $2\eta$ on the training set (empirical error), where $\| \cdot \|_K$ is used to measure the complexity.

The key question we should ask, when analyzing a norm is: “How well does it generalize?” Thus we are interested in the following quantity:

$$E \left[ \sup_{\|X\|_K \leq 1} \left| \text{err}(X) - \frac{1}{m} \sum_{(i,j,k) \in \Omega} |X_{i,j,k} - M_{i,j,k}| \right| \right]$$

which governs, over all hypotheses, what is the largest gap between its true error and its empirical error. It is standard to work with the Rademacher complexity — $R_m(K)$ — to bound this quantity. Roughly, the Rademacher complexity of $K$ is a measure of its complexity based on how well it correlates with a random labeling of the data points in our sample, which for us reside in the domain $[n] \times [n] \times [n]$. That is, $R_m(K)$ is the expectation of $\sup_{X \in K} \frac{1}{m} \sum_{\omega \in \Omega} \sigma_\omega X_\omega$ where each $\sigma_\omega$ is a $\pm 1$ random variable (see Definition 4.1). Bounds on it in turn can be used to get bounds on the generalization error that hold with high probability (see Theorem 4.2). Srebro and Shraibman [72] used this framework for matrix completion, but in our case bounding the Rademacher complexity will turn out to be much more involved.

1.2 Our Results

1.2.1 Upper bounds

We obtain a number of algorithmic results for a variety of interrelated problems. These results all utilize a common analysis. Once we bound the Rademacher complexity of a particular norm that arises from the sixth level of the sum-of-squares hierarchy, we can immediately obtain results for achieving low error rates for tensor prediction (Theorem 1.1), algorithms for strongly refuting random 3-XOR (Theorem 1.4), and achieving $2\eta + o(n^{-3/2})$ error with $m = \omega(n^{3/2} \log^4 n)$ observations. This algorithm comes from solving (K) with an appropriately chosen norm. We prove:

**Theorem 1.1** (Main theorem). Suppose that $\text{max}_{i,j,k} |\eta_{i,j,k}| \leq \sqrt{\frac{m}{\log 2/\delta}} \eta$. There is a polynomial time algorithm that, when given $m$ observations distributed uniformly at random (and without repetition) from the entries of the tensor $M$ in $(*)$, outputs a hypothesis $X$ that with probability $1 - \delta$ satisfies

$$\text{err}(X) \leq C^3 r \sqrt{\frac{\log^4 n}{mn^{3/2}}} + 2C^3 r \sqrt{\frac{\log 2/\delta}{mn^3}} + 2\eta$$

The crucial fact is that the error is $o(r^{1/2}/n^{3/2})$, when $m = \omega(n^{3/2} r \log^4 n)$, and so this type of error is asymptotically smaller than the magnitude of a typical entry. In fact, we can explicitly apply our algorithm to a random tensor, and obtain the following interesting corollary. Let $T = \sum_{i=1}^r \sigma_i a_i \otimes a_i \otimes a_i$, where $\sigma_i$ are $\pm 1$ random variables. We will further assume just for the following corollary that each factor $a_i$ has each of its entries at least $\frac{1}{D\sqrt{n}}$ for some parameter $D$ that is not necessarily a constant. Finally suppose that $\text{max}_{i,j,k} |\eta_{i,j,k}|$ is asymptotically smaller than the magnitude of a typical entry in $T$. Then:

$$\text{err}(X) \leq C^3 r \sqrt{\frac{\log^4 n}{mn^{3/2}}} + 2C^3 r \sqrt{\frac{\log 2/\delta}{mn^3}} + 2\eta$$

This is standard to work with the Rademacher complexity — $R_m(K)$ — to bound this quantity. Roughly, the Rademacher complexity of $K$ is a measure of its complexity based on how well it correlates with a random labeling of the data points in our sample, which for us reside in the domain $[n] \times [n] \times [n]$. That is, $R_m(K)$ is the expectation of $\sup_{X \in K} \frac{1}{m} \sum_{\omega \in \Omega} \sigma_\omega X_\omega$ where each $\sigma_\omega$ is a $\pm 1$ random variable (see Definition 4.1). Bounds on it in turn can be used to get bounds on the generalization error that hold with high probability (see Theorem 4.2). Srebro and Shraibman [72] used this framework for matrix completion, but in our case bounding the Rademacher complexity will turn out to be much more involved.

1.2 Our Results

1.2.1 Upper bounds

We obtain a number of algorithmic results for a variety of interrelated problems. These results all utilize a common analysis. Once we bound the Rademacher complexity of a particular norm that arises from the sixth level of the sum-of-squares hierarchy, we can immediately obtain results for achieving low error rates for tensor prediction (Theorem 1.1), algorithms for strongly refuting random 3-XOR (Theorem 1.4), and achieving $2\eta + o(n^{-3/2})$ error with $m = \omega(n^{3/2} \log^4 n)$ observations. This algorithm comes from solving (K) with an appropriately chosen norm. We prove:

**Theorem 1.1** (Main theorem). Suppose that $\text{max}_{i,j,k} |\eta_{i,j,k}| \leq \sqrt{\frac{m}{\log 2/\delta}} \eta$. There is a polynomial time algorithm that, when given $m$ observations distributed uniformly at random (and without repetition) from the entries of the tensor $M$ in $(*)$, outputs a hypothesis $X$ that with probability $1 - \delta$ satisfies

$$\text{err}(X) \leq C^3 r \sqrt{\frac{\log^4 n}{mn^{3/2}}} + 2C^3 r \sqrt{\frac{\log 2/\delta}{mn^3}} + 2\eta$$

The crucial fact is that the error is $o(r^{1/2}/n^{3/2})$, when $m = \omega(n^{3/2} r \log^4 n)$, and so this type of error is asymptotically smaller than the magnitude of a typical entry. In fact, we can explicitly apply our algorithm to a random tensor, and obtain the following interesting corollary. Let $T = \sum_{i=1}^r \sigma_i a_i \otimes a_i \otimes a_i$, where $\sigma_i$ are $\pm 1$ random variables. We will further assume just for the following corollary that each factor $a_i$ has each of its entries at least $\frac{1}{D\sqrt{n}}$ for some parameter $D$ that is not necessarily a constant. Finally suppose that $\text{max}_{i,j,k} |\eta_{i,j,k}|$ is asymptotically smaller than the magnitude of a typical entry in $T$. Then:

$$\text{err}(X) \leq C^3 r \sqrt{\frac{\log^4 n}{mn^{3/2}}} + 2C^3 r \sqrt{\frac{\log 2/\delta}{mn^3}} + 2\eta$$
**Corollary 1.2.** There is a polynomial time algorithm that, when given \( m = \tilde{\Omega}(n^{3/2}D^6C^6r) \) observations from \( M \) outputs a hypothesis \( X \) that with probability \( 1 - o_r(1) \) satisfies
\[
x_{i,j,k} = (1 \pm o(1))T_{i,j,k}
\]
for a \( 1 - o_r(1) \) fraction of the entries \((i,j,k)\).

Here we use the notation \( o_r(1) \) to denote a term that goes to zero when \( r \) is super-constant. We emphasize that the factors \( a_i \) are not themselves random, and can have arbitrary inner product. We only require that their signs \( \sigma_i \) are random. Our algorithm allows us to predict almost all of the entries of a weakly random tensor, within a multiplicative \( 1 \pm o(1) \) factor, and works even in the presence of some modest amount of noise. Moreover we only need to observe an \( o(1) \) fraction of the entries even for high rank tensors where \( r = n^{3/2 - \epsilon} \). This is well beyond linear! In fact, these results do not actually require a lower bound on each of the entries in \( a_i \), but rather that the entries of \( T \) are anti-concentrated. In the setting above, we use the results of [35] to establish this; and our results also hold for a random tensor via [28].

**Rademacher Complexity and Random 3-XOR.** As we discussed, our results are obtained by bounding the Rademacher complexity of a particular norm — \( K_6 \) — that arises from the sixth level sum-of-squares hierarchy. The full set of inequalities that define this norm is somewhat involved (Definition 3.4), in contrast to the nuclear norm that has proven to be so effective in the context of matrix completion. Nevertheless, one can compute this norm efficiently by solving an \( O(n^6) \)-sized semidefinite program. Our main technical theorem is the following:

**Theorem 1.3.** \( R^m(K_6) \leq O(\sqrt{\frac{\log^3 n}{mn^{3/2}}}) \)

It turns out that such a bound is also intimately related to the task of strong refutation for random 3-XOR. We define this task precisely in Section 2, but informally the goal is to give a polynomial time algorithm to certify that a random 3-XOR formula has no satisfying assignment that gets a non-trivial advantage over random guessing. Hence we obtain the following corollary:

**Theorem 1.4 (informal).** There is a polynomial time algorithm to strongly refute random 3-XOR, with \( m = \omega(n^{3/2}\log^4 n) \) clauses.

This result complements the well-known lower bound of Grigoriev [47] that \( n^{2.8} \) level sum-of-squares relaxation cannot (even weakly) refute a random 3-XOR with \( m = n^{3/2 - \delta} \) clauses. Our results show that with just a few more clauses, even the sixth level sum-of-squares relaxation is no longer fooled and in fact can strongly refute random 3-XOR. See Corollary 6.5. We discuss previous work on refuting random CSPs in Section 2.

### 1.2.2 Lower Bounds

While our tensor prediction algorithm of Theorem 1.1 runs in polynomial time, it is possible to do much better and approach the information theoretic limit, but with an algorithm that runs in exponential time. Let \( \| \cdot \|_A \) denote the atomic norm (Definition 3.2). Its Rademacher complexity satisfies the following bound:

**Lemma 1.5.** \( R^m(A) = O(\frac{1}{\sqrt{mn^2}}) \)

We prove this result in Section 4. Note that this quantity is \( o(1/n^{3/2}) \) when \( m = \omega(n) \), and hence if we plug it into \( \| \cdot \|_A \) into (K) then we could get an analogue of Theorem 1.1 whose generalization error is much smaller. In particular, we could improve Corollary 1.2 to use \( m = \tilde{O}(nr) \) samples instead. The difficulty is that there is no known polynomial time algorithm to compute it (even approximately), and moreover it is closely related to the injective norm of a tensor which is known to be \( NP \)-hard to approximate [48, 50]. Nevertheless there was optimism that there would be a sequence of semidefinite programs to approximate the atomic norm whose performance would approach the information theoretic limit for tensor prediction/completion [30]. No quantitative bounds were known for how many levels we need to get a good enough approximation, and in fact here we show a lower bound that matches our upper bound. Again, connecting our problem to the task of strongly refuting random 3-XOR formulas, appealing to known results [47, 69] we show:
Theorem 1.6. For any $\epsilon > 0$ and any $c < 2$ and $k = \Omega(n^c)$, the $k$-round sum-of-squares relaxation — $K_k$ — for the atomic norm has $R^m(K_k) \geq \frac{1 - o(1)}{m^{3/2}}$ for $m = n^{3/2 - \epsilon}$.

We prove this result in Section 7.

In particular even if our observations are chosen randomly to be either $+1/\sqrt{n}$ or $-1/\sqrt{n}$ and do not come from a low-order tensor at all, there is still a hypothesis $X$ with $\|X\|_{K_k} \leq 1$ that agrees exactly with $1 - o(1)$ of our observations. The sum-of-squares hierarchy has been proposed as a natural, universal algorithm that for a broad range of problems encapsulates the limits of what we know how to accomplish in polynomial time [14]. To this end, it is natural to believe that our upper and lower bounds characterize a sharp phase transition in the tradeoff between sample complexity and computational complexity. Moreover this sort of tradeoff is not present in any of the previous work on solving linear inverse problems.

1.3 Tradeoffs via Hierarchies?

There has been considerable recent interest in understanding tradeoffs between sample complexity and computational complexity [17, 29]. This is an exciting area, but the question is: How do we know whether some problem possesses an inherent tradeoff or we just have not found the right algorithm that runs efficiently and solves the problem with the information-theoretic minimum number of samples? We could hope to prove computational lower bounds under widely believed assumptions such as $P \neq NP$. However there are fundamental obstructions to using such assumptions to prove lower bounds for average case problems such as those that arise in learning [6, 19]. We could also hope to derive many lower bounds from a single assumption. But despite some early successes [17], as a rule reductions tend not to preserve the “naturalness” of the input distribution.

In this sense, we can think of the present work as the natural next step: Can we take a powerful family of algorithms, and prove sample complexity vs. computational complexity tradeoffs? Indeed, a wide range of algorithm design can be thought of as happening within the sum-of-squares hierarchy, so methodologically it is a powerful way to explore the intersection between machine learning, statistics and algorithm design. Moreover, any new algorithm that accomplishes something in this domain that is provably well beyond the reach of the sum-of-squares hierarchy would be a breakthrough in its own right. Here we have shown that the sixth level of the sum-of-squares hierarchy achieves essentially the best sample complexity, and that there is a sharp phase transition when given fewer than $n^{3/2}$ observations. We believe that there will be many further examples where imposing (rough) computational constraints will create new phase transitions in the amount of data we need.

1.4 Further Related Work

We discuss related work on refuting random CSPs in Section 2, but here we discuss tensor decompositions. Tensor decompositions have numerous applications in learning, specifically learning the parameters of various latent variable models. See Chapter 3 in [62], which includes an exposition of the applications to phylogenetic reconstruction [63], HMMs [63], mixture models [54], topic modeling [3] and community detection [4]. The setting for these problems is that one is given access to the entire tensor (perhaps up sampling noise from estimating the moments of a distribution from random samples) and the goal is to find its low rank decomposition. In contrast, in our setting of tensor prediction one observes a tiny fraction of the tensor, and our goal is to (approximately) fill in the missing entries. Thus, algorithms for these two tasks of tensor prediction and decomposition are highly complimentary ways to work with tensor data, first to fill it in from few observations, then to use it to estimate the parameters of a model.

We remark that in most if not all of the applications of tensor decompositions, it is crucial that the factors are not constrained to be orthogonal and this generality is what enables tensor decompositions to have so many further applications in learning, in comparison to matrix decompositions. Finally, most algorithms for tensor decomposition work up to $r = n$ for third order tensors, but here we are able to work with tensors whose rank is $r = n^{3/2 - \epsilon}$. (There is recent work on third order tensor decomposition, for $r = n^{3/2 - \epsilon}$ however these works require the tensor to be random [44, 52].)
2 Connections Between Tensor Prediction and 3-XOR

Before we begin our analysis, let us elaborate on the connections between tensor prediction and the task of strongly refuting random 3-XOR formulas. We do so because it is useful to keep both views of this problem in mind, and each offers different sorts of insights. Our key observation is the following connection:

Observation 1. Any norm $K_k$ that satisfies $R^n(K_k) = o(1/n^{3/2})$ can be used to strongly refute a random 3-XOR formula.

Let us make this more precise: Throughout this paper, it will be more convenient to think of a 3-XOR formula as a collection of constraints of the form $v_i v_j v_k = Z_{i,j,k}$ where $v_i, v_j, v_k, Z_{i,j,k} \in \{\pm 1\}$. This is a standard transformation where $-1$ is associated with the value true and $1$ is associated with the value false.

Now we can transform a 3-XOR formula with $m$ clauses into a collection of $m$ observations according to the following rules:

(a) A clause $v_i \cdot v_j \cdot v_k = Z_{i,j,k}$ is mapped to an observation at location $(i,j,k)$ whose value is $Z_{i,j,k}$.

(b) We solve the following optimization problem

\[
\max \eta \text{ s.t. } \|X\|_{K_k} \leq 1 \text{ and } \frac{1}{m} \sum_{(i,j,k) \in \Omega} X_{i,j,k} Z_{i,j,k} \geq \frac{2\eta}{n^{3/2}}
\]

which (informally) asks to find $X$ with $\|X\|_{K_k} \leq 1$ that has maximal agreement with the observations.

(c) Let $opt$ be the optimum value, and output $val = 1/2 + opt$.

The key is that the fact that $A \subseteq K_k$ together with any bound $R^n(K_k) = o(1/n^{3/2})$ readily implies that the quantity $val$ satisfies:

(a) $val$ is an upper bound on the fraction of clauses that can be satisfied

(b) $val$ is $1/2 + o(1)$ with high probability

We refer to this as strong refutation. See Corollary 6.5 for further details.

In this language, our approach is to find an algorithm for strongly refuting a random 3-XOR formula with $m \geq cn^{3/2} \log^4 n$ clauses. We then embed it into the sixth level of the sum-of-squares hierarchy. Our results are inspired by the algorithm of Coja-Oghlan, Goerdt and Lanka [31] which provides an upper bound on the fraction of clauses that can be satisfied in a random 3-SAT formula, where $val$ is $7/8 + o(1)$ with high probability provided that $m \geq cn^{3/2} \log^4(1/n)$. We remark that their result improves on a long line of work on refuting a random 3-SAT formula [46, 40, 39, 41] with roughly the same number of clauses, and one of the consequences of our work is that algorithms that go further and provide strong bounds on the fraction of clauses that can be satisfied can actually have algorithmic implications in machine learning.

Our lower bounds also follow from this connection. We make use of the results in [47, 69] which brought powerful tools [16] from proof complexity to bear, to show that the Lasserre hierarchy needs $\Omega(n^{2c})$ rounds to refute a random 3-XOR formula where $m \leq n^{3/2 - \epsilon}$. We then give straightforward reductions between the hierarchy that these results were originally proved for, and the natural hierarchy in our setting. We remark that both the upper and lower bounds extend to all order $d$ tensors (see Section 6, Section 7). In the case of even $d$, the upper bounds were known and follow from unfolding the tensor into an $n^{d/2} \times n^{d/2}$ matrix. This seems to have been observed by several authors. Here our contribution is in completing the picture for odd $d$.

3 Norms and Pseudo-expectation

Here we will introduce the various norms that we work with throughout this paper. We will be interested in bounding their Rademacher complexity. In many cases, these norms will arise from the sum-of-squares hierarchy which will allow us to take a parallel view where we think of an element $X$ with $\|X\|_{K} \leq 1$ as defining a pseudo-expectation operator (see below).
Definition 3.1. A vector $a$ is $C$-balanced if $a$ is a unit vector in $\mathbb{R}^n$ and $\|a\|_\infty \leq \frac{C}{\sqrt{n}}$.

We will be interested in the following family of convex optimization problems:

$$(K) \quad \min \|X\|_K \text{ s.t. } \frac{1}{m} \sum_{(i,j,k) \in \Omega} |X_{i,j,k} - T_{i,j,k}| \leq 3\eta$$

where $\Omega$ is the set of observed entries and $\|\cdot\|_K$ is a norm. This type of convex optimization problem plays a crucial role in designing algorithms for a wide range of linear inverse problems [68, 24, 27, 22, 20, 30], and the general recipe is to define an atomic norm (see [30]) which in our case is the following:

Definition 3.2 (atomic norm). Let $A \subseteq \mathbb{R}^3$ be defined as

$$A = \{X_{i,j,k} \mid X_{i,j,k} = \mathbb{E}_{a \sim \mu}[a_i a_j a_k]\}$$

where $\mu$ is a distribution on $C$-balanced vectors. The atomic norm of $X$ is the infimum over $\alpha$ such that $X/\alpha \in A$.

The difficulty is that there are no known algorithms to compute this norm (even approximately), and the resulting optimization problem $(K)$ for this choice of norm may very well be computationally hard. In fact, if we remove the restriction that $\mu$ be a distribution on $C$-balanced vectors and only require it to be a distribution on unit vectors then this norm is dual to the injective norm, and both it and its dual are known to be $NP$-hard to approximate for third order tensors [48, 61, 45, 50].

Instead, we will work with the norm — $K_\alpha$ — induced by the sixth level sum-of-squares relaxation to the atomic norm. Next we introduce the notion of a pseudo-expectation operator from [10, 11, 14]:

Definition 3.3 (Pseudo-expectation [10]). Let $k$ be even and let $P_k^n$ denote the linear subspace of all polynomials of degree at most $k$ on $n$ variables. A linear operator $\mathcal{L} : P_k^n \rightarrow \mathbb{R}$ is called a degree $k$ pseudo-expectation operator if it satisfies the following conditions:

“normalization”: $\mathcal{L}(1) = 1$.

“nonnegativity”: $\mathcal{L}(P^2) \geq 0$, for any degree at most $k/2$ polynomial $P$.

Moreover suppose that $p \in P_k^n$ with $\deg(p) = k'$. We say that $\mathcal{L}$ satisfies the constraint $\{p = 0\}$ if $\mathcal{L}(pq) = 0$ for every $q \in P_{k-k'}^n$. And we say that $\mathcal{L}$ satisfies the constraint $\{p \geq 0\}$ if $\mathcal{L}(pq^2) \geq 0$ for every $q \in P_{[k-k']/2}^n$.

The rationale behind this definition is that if $\mu$ is a distribution on vectors in $\mathbb{R}^n$ then the operator $\mathcal{L}(p) = \mathbb{E}_{Y \sim \mu}[p(Y')]$ is a degree $d$ pseudo-expectation operator for every $d$ — i.e. it meets the conditions of Definition 3.3. However the converse is in general not true. Nonetheless, we will often use the suggestive notation $\mathbb{E}[p]$ to denote the value $\mathcal{L}(p)$. We are now ready to define the norm that will be used in our upper bounds:

Definition 3.4 ($SOS_k$ norm). We let $K_k$ be the set of all $X \in \mathbb{R}^{n^3}$ such that there exists a degree $k$ pseudo-expectation operator on $P_k^n$ satisfying the constraints

(a) $\{\sum_{i=1}^n Y_i^2 = 1\}$

(b) $\{Y_i^2 \leq C^2/n\}$ for all $i \in \{1, 2, \ldots, n\}$ and

(c) $X_{i,j,k} = \mathbb{E}[Y_i Y_j Y_k]$ for all $i, j$ and $k$.

The $SOS_k$ norm of $X \in \mathbb{R}^{n^3}$ is the infimum over $\alpha$ such that $X/\alpha \in K_k$.

---

5Here we have made a shift in convention, where the constraints are on how close the entries of $X$ are to those in $T$. Recall that we do not have access to $T$ but rather to $M$ where $M_{i,j,k} = T_{i,j,k} + \eta_{i,j,k}$. However any solution to the optimization problem $(K)$ as defined in the introduction, will with probability $1 - \delta$ be a feasible solution to the optimization problem above (by the triangle inequality and a Chernoff bound), and moreover it will have the same norm. Our main interest is in showing that a feasible solution with low norm has nearly as good true error as its empirical error.
Throughout most of this paper, we will be interested in the case $k = 6$.

The constraints in Definition 3.3 can be expressed as an $O(n^6)$-sized semidefinite program. This implies that given any set of polynomial constraints of the form \{ $p = 0$, $p \geq 0$ \}, one can efficiently find a degree $k$ pseudo-distribution satisfying those constraints if one exists. This is often called the degree $k$ Som-of-Squares algorithm [71, 64, 59, 65]. Hence we can compute the SOS6 norm of any tensor $X$ within arbitrary accuracy in polynomial time. We emphasize that the SOS6 norm is a relaxation for the atomic norm and $A \subseteq K_6$. In particular, this implies that $\|X\|_A \leq \|X\|_{k_6}$ for every tensor $X$.

## 4 Rademacher Complexity

Our first goal is to bound the Rademacher complexity of the resolution norm. The following are standard arguments from empirical process theory, and we repeat them here for completeness. See [58, 15] for further details. Let

$$\Omega = \{(i_1, j_1, k_1), (i_2, j_2, k_2), \ldots, (i_m, j_m, k_m)\}$$

be a random set of $m$ samples (chosen uniformly at random and without repetition) which correspond to observed entries of the tensor $T$. Similarly let $\Omega'$ be an independent set of $m$ samples from the same distribution, again without repetition. Then we are interested in bounding the worst generalization error over the concept class, as given below:

$$\mathbb{E}_{\Omega} \left[ \sup_{\|X\|_K \leq 1} \left| \frac{1}{m} \sum_{\ell=1}^{m} |X_{i_\ell,j_\ell,k_\ell} - T_{i_\ell,j_\ell,k_\ell}| - \mathbb{E}_{\Omega'} \left[ |X_{i,j,k} - T_{i,j,k}| \right] \right| \right]$$

Then we can write

$$(*) \leq \mathbb{E}_{\Omega,\Omega'} \left[ \sup_{\|X\|_K \leq 1} \left| \frac{1}{m} \sum_{\ell=1}^{m} |X_{i_\ell,j_\ell,k_\ell} - T_{i_\ell,j_\ell,k_\ell}| - \mathbb{E}_{\Omega'} \left[ X_{i,j,k} - T_{i,j,k} \right] \right| \right]$$

where the last line follows by the concavity of $\sup(\cdot)$. Now we can introduce Rademacher (random $\pm 1$) variables $\{\sigma_\ell\}_\ell$ in which case we can rewrite the right hand side of the above expression as follows:

$$(*) \leq \mathbb{E}_{\Omega,\Omega',\sigma} \left[ \sup_{\|X\|_K \leq 1} \left| \frac{1}{m} \sum_{\ell=1}^{m} \sigma_\ell \left( |X_{i_\ell,j_\ell,k_\ell} - T_{i_\ell,j_\ell,k_\ell}| - |X_{i'_\ell,j'_\ell,k'_\ell} - T_{i'_\ell,j'_\ell,k'_\ell}| \right) \right| \right]$$

$$\leq \mathbb{E}_{\Omega,\Omega',\sigma} \left[ \sup_{\|X\|_K \leq 1} \left| \frac{1}{m} \sum_{\ell=1}^{m} \sigma_\ell |X_{i_\ell,j_\ell,k_\ell} - T_{i_\ell,j_\ell,k_\ell}| \right| + \frac{1}{m} \sum_{\ell=1}^{m} \sigma_\ell |X_{i'_\ell,j'_\ell,k'_\ell} - T_{i'_\ell,j'_\ell,k'_\ell}| \right| \right]$$

$$\leq 2 \mathbb{E}_{\Omega,\sigma} \left[ \frac{1}{m} \sum_{\ell=1}^{m} \sigma_\ell |T_{i_\ell,j_\ell,k_\ell}| \right] + 2 \mathbb{E}_{\Omega,\sigma} \left[ \sup_{\|X\|_K \leq 1} \frac{1}{m} \sum_{\ell=1}^{m} \sigma_\ell |X_{i_\ell,j_\ell,k_\ell}| \right]$$

where the second and fourth inequalities use the triangle inequality. The second term above is called the Rademacher complexity (of $K$). Our goal is to bound it.

**Definition 4.1.** The Rademacher complexity $R^m(\mathcal{H})$ of a class of hypotheses $\mathcal{H}$ is

$$\mathbb{E}_{\Omega,\sigma} \left[ \sup_{f \in \mathcal{H}} \frac{1}{m} \sum_{\ell=1}^{m} \sigma_\ell |f(i_\ell,j_\ell,k_\ell)| \right]$$
The following result is well-known, and allows us to bound the generalization error (with high probability) using the Rademacher complexity.

**Theorem 4.2.** Let \( \delta \in (0, 1) \) and suppose each \( f \in \mathcal{H} \) has bounded loss — i.e. \( |f(i,j,k) - T_{i,j,k}| \leq a \) and that samples are drawn i.i.d. Then with probability at least \( 1 - \delta \), for every function \( f \in \mathcal{H} \), we have

\[
\mathbb{E}[|f(i,j,k) - T_{i,j,k}|] \leq \frac{1}{m} \sum_{\ell=1}^{m} |f(i_{\ell},j_{\ell},k_{\ell}) - T_{i_{\ell},j_{\ell},k_{\ell}}| + 2 R^m(\mathcal{H}) + 2a \sqrt{\frac{\ln(1/\delta)}{m}}
\]

This follows by applying McDiarmid’s inequality to the above symmetrization argument. See for example [8].

We also remark that in our setting, samples are not drawn i.i.d. from a distribution because our observations are chosen without replacement. However for our setting of \( m \), each observed entry is expected to be seen a total of \( 1 + o(1) \) times and we can move back and forth between both sampling models at negligible loss in our bounds, and hence we can still make use of the above theorem.

Note that since the values of \( T_{i,j,k} \) and \( X_{i,j,k} \) are invariant under re-ordering their indices, it will actually be more convenient to think of \( \Omega \) as a collection of ordered triples of indices. We can do this by choosing an ordering of the indices, for each observation, uniformly at random. Then we will be interested in the following random tensor, whose values are determined by \( \Omega \) and \( \{\sigma_{\ell}\}_{\ell} \):

**Definition 4.3.** Let \( Z \) be an \( n \times n \times n \) random tensor such that

\[
Z_{i,j,k} = \begin{cases} 
0, & \text{if } (i,j,k) \notin \Omega \\
\sigma_{\ell} & \text{where } (i,j,k) = (i_{\ell},j_{\ell},k_{\ell})
\end{cases}
\]

Then according to our convention, \( Z_{i,j,k} \) is not a symmetric tensor because the triples in \( \Omega \) are ordered. This change will actually simplify our calculations, because in the next sections, when we sum over indices \( i,j,k \) we can consider these as sums over all ordered triples. In particular we have

\[
\sum_{\ell=1}^{m} \sigma_{\ell} X_{i_{\ell},j_{\ell},k_{\ell}} = \sum_{i,j,k} Z_{i,j,k} X_{i,j,k} = \langle Z, X \rangle
\]

where the sum is over all ordered triples, and we have introduced the notation \( \langle \cdot, \cdot \rangle \) to denote the natural inner-product between tensors.

We remark that it is straightforward to bound the Rademacher complexity of the atomic norm.

**Lemma 4.4.** \( R^m(\mathcal{A}) = O\left( \frac{c^3}{\sqrt{mn^2}} \right) \)

**Proof.** Using the definition of \( Z \) above, it is easy to see that

\[
\mathbb{E}_{\Omega, a} \left[ \sup_{\|X\|_A \leq 1} \left| \sum_{\ell=1}^{m} \sigma_{\ell} X_{i_{\ell},j_{\ell},k_{\ell}} \right| \right] = \sup_{C\text{-balanced } a} |\langle Z, a \otimes a \otimes a \rangle| \]

We can now adapt the discretization approach in [42], although our task is considerably simpler because we are constrained to \( C\)-balanced \( a \)'s. In particular, let

\[
T = \left\{ a \big| a \text{ is } C\text{-balanced and } a \in \left( \epsilon \mathbb{Z} \sqrt{m} \right)^n \right\}
\]

Then \( |T| \leq O(1/\epsilon)^n \). It will be more convenient to work with the asymmetric case. In particular, suppose that \( |\langle Z, a \otimes b \otimes c \rangle| \leq M \) for all \( a, b, c \in T \). Then for an arbitrary, but balanced \( a \) we can expand it as \( a = \sum_{i} \epsilon^i a^i \) where each \( a^i \in T \) and similarly for \( b \) and \( c \). And now

\[
|\langle Z, a \otimes b \otimes c \rangle| \leq \sum_{i} \sum_{j} \sum_{k} \epsilon^i \epsilon^j \epsilon^k |\langle Z, a^i \otimes b^j \otimes c^k \rangle| \leq (1 - \epsilon)^{-3} M
\]
Moreover since each entry in $a \otimes b \otimes c$ has magnitude at most $\frac{C^3}{m^{3/2}}$ and we can apply a Chernoff bound to obtain that for any particular $a, b, c \in T$ we have

$$|\langle Z, a \otimes b \otimes c \rangle| \leq O\left(\frac{C^3 \sqrt{m \log 1/\delta}}{n^{3/2}}\right)$$

with probability at least $1 - \delta$. Finally, if we set $\delta = (\Omega(\epsilon))^{-n}$ and we set $\epsilon = 1/2$ we get that

$$R_m^a(A) \leq (1 - \epsilon)^{-3} m \max_{a, b, c \in T} |\langle Z, a \otimes b \otimes c \rangle| = O\left(\frac{C^3 \sqrt{mn^2}}{n^{3/2}}\right)$$

and this completes the proof.

The important point is that the Rademacher complexity is $o\left(\frac{1}{n^{3/2}}\right)$ as soon as $m = \omega(n)$, which is nearly the best we could hope for. However the only known algorithm for solving $(K)$ with this norm runs in exponential time.

## 5 Resolution in $K_6$

Here we will reduce the question of bounding the Rademacher complexity of $K_6$ to a question about the spectral norm of a random matrix, albeit one whose entries are dependent. We can think of the arguments in this section as embedding a particular algorithm (based on resolution) for strongly refuting a random 3-XOR formula into the sum-of-squares hierarchy.

Let $X \in K_6$. Then there is a degree six pseudo-expectation meeting the conditions of Definition 3.4. Using Cauchy-Schwartz we have:

$$\left(\langle Z, X \rangle\right)^2 = \left(\sum_i \sum_{j,k} Z_{i,j,k} \mathbb{E}[Y_i Y_j Y_k]\right)^2 \leq n \left(\sum_i \left(\sum_{j,k} Z_{i,j,k} \mathbb{E}[Y_i Y_j Y_k]\right)^2\right)$$

To simplify our notation, we will define the following polynomial

$$Q_{i,Z}(Y) = \sum_{j,k} Z_{i,j,k} Y_j Y_k$$

which we will use repeatedly. If $d$ is even then any degree $d$ pseudo-expectation operator satisfies the constraint $(\mathbb{E}[p])^2 \leq \mathbb{E}[p^2]$ for every polynomial $p$ of degree at most $d/2$ (e.g., see Lemma A.4 in [9]). Hence the right hand side of (1) can be bounded as:

$$n \left(\sum_i \left(\mathbb{E}[Y_i Q_{i,Z}(Y)]\right)^2\right) \leq n \sum_i \mathbb{E}\left[\left(Y_i Q_{i,Z}(Y)\right)^2\right]$$

It turns out that bounding the right-hand side of (2) boils down to bounding the spectral norm of the following matrix.

**Definition 5.1.** Let $A$ be the $n^2 \times n^2$ matrix whose rows and columns are indexed over ordered pairs $(j, j')$ and $(k, k')$ respectively, defined as

$$A_{j,j',k,k'} = \sum_i Z_{i,j,k} Z_{i,j',k'}$$

We can now make the connection to resolution more explicit: If $Z_{i,j,k}, Z_{i,j',k'} \in \{\pm 1\}$ then we can think of them as 3-XOR constraints. We map these to the constraints $x_i \cdot x_j \cdot x_k = Z_{i,j,k}$ and $x_i \cdot x_j' \cdot x_{k'} = Z_{i,j',k'}$.

In this case, resolving them (i.e. multiplying them) we obtain a 4-XOR constraint

$$x_j \cdot x_k \cdot x_j' \cdot x_{k'} = x_j \cdot x_j' \cdot x_k \cdot x_{k'} = Z_{i,j,k} Z_{i,j',k'}$$

Thus $A$ captures the effect of resolving 3-XOR constraints. However we note that the entries in $A$ are not independent, so bounding its maximum singular value will require some care. We remark that the rows of
\[ A \text{ are indexed by } (j, j') \text{ and the columns are indexed by } (k, k'), \text{ and it is crucial that } j \text{ and } j' \text{ come from different 3-XOR clauses, as do } k \text{ and } k'. \]

It will be more convenient to decompose this matrix and reason about its two types of contributions separately. To that end, we let \( R \) be the \( n^2 \times n^2 \) matrix whose non-zero entries are of the form
\[
R_{j,j,k,k} = \sum_i Z_{i,j,k}Z_{i,j,k}
\]
and all of its other entries are set to zero. Then let \( B \) be the \( n^2 \times n^2 \) matrix whose entries are of the form
\[
B_{j,j',k,k'} = \begin{cases} 
0, & \text{if } j = j' \text{ and } k = k' \\
\sum_i Z_{i,j,k}Z_{i,j',k'} & \text{else}
\end{cases}
\]

And clearly we have \( A = B + R \). Finally:

**Lemma 5.2.**
\[
\sum_i \mathbb{E}\left[\left(Y_iQ_{1,Z}(Y)\right)^2\right] \leq \frac{C^2}{n}\|B\| + \frac{C^6m}{n^3}
\]

**Proof.** Recall that the pseudo-expectation operator satisfies \( \{Y_i^2 \leq C^2/n\} \) for all \( i \in \{1, 2, \ldots, n\} \), and hence we have
\[
\sum_i \mathbb{E}\left[\left(Y_iQ_{1,Z}(Y)\right)^2\right] \leq \frac{C^2}{n}\sum_i \mathbb{E}\left[\left(Q_{1,Z}(Y)\right)^2\right] = \frac{C^2}{n}\sum_{j,k,j',k'} \mathbb{E}\left[Z_{i,j,k}Z_{i,j',k'}Y_jY_{j'}Y_kY_{k'}\right]
\]

Now let \( Y \in \mathbb{R}^n \) be a vector of variables where the \( i \)th entry is \( Y_i \). Then we can re-write the right hand side as a matrix inner-product:
\[
\frac{C^2}{n}\sum_{j,k,j',k'} \mathbb{E}\left[Z_{i,j,k}Z_{i,j',k'}Y_jY_{j'}Y_kY_{k'}\right] = \frac{C^2}{n} \langle A, \mathbb{E}[(Y \otimes Y)(Y \otimes Y)^T]\rangle
\]

Recall that \( A = B + R \). We will bound the contribution of each separately.

**Claim 5.3.** \( \mathbb{E}[(Y \otimes Y)(Y \otimes Y)^T] \) is positive semi-definite and has trace at most one

**Proof.** It is easy to see that a quadratic form on \( \mathbb{E}[(Y \otimes Y)(Y \otimes Y)^T] \) corresponds to \( \mathbb{E}[p^2] \) for some \( p \in P^2_n \) and this implies the first part of the claim. And also
\[
\text{Tr}(\mathbb{E}[(Y \otimes Y)(Y \otimes Y)^T]) = \sum_{j,k} \mathbb{E}[Y_j^2Y_k^2] = 1
\]

where the last equality follows because the pseudo-expectation operator satisfies \( \{\sum_{i=1}^n Y_i^2 = 1\} \).

Hence we can bound the contribution of the first term as \( \frac{C^2}{n}\langle B, \mathbb{E}[(Y \otimes Y)(Y \otimes Y)^T]\rangle \leq \frac{C^2}{n}\|B\| \). Now we proceed to bound the contribution of the second term:

**Claim 5.4.** \( \mathbb{E}[Y_j^2Y_k^2] \leq \frac{C^4}{n^2} \)

**Proof.** It is easy to verify that the following equality holds:
\[
\frac{C^4}{n^2} - Y_j^2Y_k^2 = \left(\frac{C^2}{n} - Y_j^2\right)\left(\frac{C^2}{n} - Y_k^2\right) + \left(\frac{C^2}{n} - Y_j^2\right)Y_j^2 + \left(\frac{C^2}{n} - Y_k^2\right)Y_k^2
\]

Moreover the pseudo-expectation of each of the three terms above is nonnegative, by construction. This implies the claim.

Moreover each entry in \( Z \) is in the set \( \{-1, 0, +1\} \) and there are precisely \( m \) non-zero terms. Thus it is easy to see that the sum of the absolute values of all entries in \( R \) is at most \( m \). Hence we have:
\[
\frac{C^2}{n}\langle R, \mathbb{E}[(Y \otimes Y)(Y \otimes Y)^T]\rangle \leq \sum_{j,k} R_{j,j,k,k} \mathbb{E}[Y_j^2Y_k^2] \leq \frac{C^6m}{n^3}
\]

And this completes the proof of the lemma.
6 Spectral Bounds, and Putting it All Together

Recall the definition of $B$ given in the previous section. However in this section, it will be more convenient to index $B$ as follows:

$$B_{j,k,j',k'} = \begin{cases} 0, & \text{if } j = k \text{ and } j' = k' \\ \sum_i Z_{i,j,j'} Z_{i,k,k'} & \text{else} \end{cases}$$

Note that the rows of $B$ are still indexed by pairs that come from different 3-XOR clauses, and the same is true for the columns.

We will also work with an asymmetric version of this matrix instead. Let us consider the following random process: For $r = 1, 2, ..., O(\log n)$ partition the set of all ordered triples $(i, j, k)$ into two sets $S_r$ and $T_r$. We will use this ensemble of partitions to define an ensemble of matrices $\{B_r\}_{r=1}^{O(\log n)}$: Set $U_{i,j,j'}^{r}$ as equal to $Z_{i,j,j'}$ if $(i, j, j') \in S_r$ and zero otherwise. Similarly set $V_{i,k,k'}^{r}$ equal to $Z_{i,k,k'}$ if $(i, k, k') \in T_r$ and zero otherwise. Also let $E_{i,j,j',k,k',r}$ be the event that there is no $r' < r$ where $(i, j, j') \in S_{r'}$ and $(i, k, k') \in T_{r'}$ or vice-versa. Now let

$$B_{j,k,j',k'}^{r} = \sum_i U_{i,j,j'}^{r} V_{i,k,k'}^{r} \mathbb{1}_E$$

where $\mathbb{1}_E$ is short-hand for the indicator function of the event $E_{i,j,j',k,k',r}$. The idea behind this construction is that each pair of triples $(i, j, j')$ and $(i, k, k')$ that contributes to $B$ will be contribute to some $B^r$ with high probability. Moreover it will not contribute to any later matrix in the ensemble. Hence with high probability

$$B = \sum_{r=1}^{O(\log n)} B^r$$

Throughout the rest of this section, we will suppress the superscript $r$ and work with a particular matrix in the ensemble, $B$. Now let $\ell$ be even and consider

$$\text{Tr}(BB^TBB^T...BB^T)$$

As is standard, we are interested in bounding $E[\text{Tr}(BB^TBB^T...BB^T)]$ in order to bound $\|B\|$. But note that $B$ is not symmetric. Also note that the random variables $U$ and $V$ are not independent, but whether or not they are non-zero is non-positively correlated and their signs are mutually independent. Expanding the trace above we have

$$\text{Tr}(BB^TBB^T...BB^T) = \sum_{j_1,k_1,j_2,k_2} ... \sum_{j_{\ell-1},k_{\ell-1}} B_{j_1,k_1,j_2,k_2} B_{j_3,k_3,j_2,k_2} ... B_{j_1,k_1,j_\ell,k_\ell}$$

$$= \sum_{j_1,k_1} \sum_{j_2,k_2} \sum_{i_1} ... \sum_{j_{\ell},k_{\ell}} U_{i_1,j_1,j_2} V_{i_1,k_1,k_2} \mathbb{1}_{E_1} U_{i_2,j_3,j_2} V_{i_2,k_3,k_2} \mathbb{1}_{E_2} ... U_{i_{\ell},j_1,j_\ell} V_{i_{\ell},k_1,k_\ell} \mathbb{1}_{E_{\ell}}$$

where $\mathbb{1}_{E_1}$ is the indicator for the event that the entry $B_{j_1,k_1,j_2,k_2}$ is not covered by an earlier matrix in the ensemble, and similarly for $\mathbb{1}_{E_2}, ..., \mathbb{1}_{E_{\ell}}$.

Notice that there are $2\ell$ random variables in the above sum (ignoring the indicator variables). Moreover if any $U$ or $V$ random variable appears an odd number of times, then the contribution of the term to $E[\text{Tr}(BB^TBB^T...BB^T)]$ is zero. We will give an encoding for each term that has a non-zero contribution, and all we need from this encoding is that it be injective.

Fix a particular term in the above sum where each random variable appears an even number of times. Let $s$ be the number of distinct values for $i$. Moreover let $i_1, i_2, ..., i_s$ be the order that these indices first appear. Now let $r_1$ and $q_1$ denote the number of distinct values for $j$ and $k$ respectively that appear with $i_1$ — i.e. $r_1$ is the number of distinct $j$‘s that appear as $U_{i_1,j,*}$ or $U_{i_1,*,*}$. We give our encoding below. It is more convenient to think of the encoding as any way to answer the following questions about the term.

(a) What is the order $i_1, i_2, ..., i_s$ of the first appearance of each distinct value of $i$?

(b) For each $i$ that appears, what is the order of each distinct value of $j$ that appears with it? Similarly, what is the order for each distinct value of $k$ that appears with it?
(c) For each step (i.e. a new variable in the term when reading from left to right), has the value of \( i \) been visited already? Also, has the value for \( j \) been visited? Has the value for \( k \) been visited? Note that whether or not \( j \) has been visited depends on what the value of \( i \) is, and if \( i \) is a new value then \( j \) must be new too, by definition. Finally, if any value has already been visited, which earlier value is it?

Let \( r = r_1 + r_2 + \ldots + r_s \) and let \( q = q_1 + q_2 + \ldots + q_s \). Then the number of possible answers to (a) and (b) is at most \( n^s \) and \( n^s n^q \) respectively. It is also easy to see that the number of answers to (c) that arise over the sequence of \( \ell \) steps is at most \( 8^\ell (srq)^\ell \). We remark that much of the work on bounding the maximum eigenvalue of a random matrix is in removing any \( \ell \) type terms, and so needs to encode visiting random variables that have already been visited more compactly. See [74] for further discussion. However such terms will only cost us polylogarithmic factors in our bound on \( \|B\| \) in any case.

It is easy to see that this encoding is injective, since given the answers to the above questions one can simulate each step and recover the sequence of random variables. Next we establish some easy facts that allow us to bound \( E[\text{Tr}(B^TB^TB^T \ldots B^T)] \).

**Claim 6.1.** For any term that has a non-zero contribution to \( E[\text{Tr}(B^TB^TB^T \ldots B^T)] \), we must have \( s, r, q \leq \ell/2 \)

**Proof.** Recall that there are \( 2\ell \) random variables in the product and precisely \( \ell \) of them correspond to \( U \) variables and \( \ell \) of them to \( V \) variables. Suppose that \( s > \ell/2 \). Then there must be at least one \( U \) variable and at least one \( V \) variable that occur exactly once, which implies that its expectation is zero. Similarly suppose \( r > \ell/2 \). Then there must be at least one \( U \) variable that occurs exactly once, which also implies that its expectation is zero. An identical argument holds for \( q \) and with \( V \) variables. \( \square \)

**Claim 6.2.** For any valid encoding, \( s \leq r \) and \( s \leq q \).

**Proof.** This holds because in each step where the \( i \) variable is new and has not been visited before, by definition the \( j \) variable is new too (for the current \( i \)) and similarly for the \( k \) variable. \( \square \)

Finally, if \( s, r \) and \( q \) are defined as above for the term

\[
U_{i_1,j_1,j_2} V_{i_1,k_1,k_2} U_{i_2,j_3,j_2} V_{i_2,k_3,k_2} \ldots U_{i_t,j_t,j_t} V_{i_t,k_1,k_t}
\]

then its expectation is at most \( p^r p^q \) where \( p = m/n^3 \) because there are exactly \( r \) distinct \( U \) variables and \( q \) distinct \( V \) variables whose value is in the set \( \{-1,0,+1\} \) and whether or not a variable is non-zero is non-positively correlated and the signs are mutually independent.

This now implies the main lemma:

**Lemma 6.3.** \( E[\text{Tr}(B^TB^TB^T \ldots B^T)] \leq n^{3\ell/2} p^r (\ell)^{3\ell+3} \)

**Proof.** Note that the indicator variables only have the effect of zeroing out some terms that could otherwise contribute to \( E[\text{Tr}(B^TB^TB^T \ldots B^T)] \). Returning to the task at hand, we have

\[
E[\text{Tr}(B^TB^TB^T \ldots B^T)] \leq \sum_{s,r,q} n^s n^r n^q p^r p^q (srq)^\ell \leq \sum_{s,r,q} n^s n^r n^q p^r p^q (\ell)^{3\ell}
\]

where the sum is over all valid triples \( s, r, q \) and hence \( s, r, q \leq \ell/2 \) and \( s \leq r \) and \( s \leq q \) using Claim 6.1 and Claim 6.2. Now \( pn \) is much less than one, so we can upper bound the above as

\[
E[\text{Tr}(B^TB^TB^T \ldots B^T)] \leq \sum_{s,r,q} n^s (pn)^r (pn)^q (\ell)^{3\ell} \leq n^s (pn)^s (pn)^s (\ell)^{3\ell} = n^{3s} p^{2s} (\ell)^{3\ell+3}
\]

Hence \( E[\text{Tr}(B^TB^TB^T \ldots B^T)] \leq n^{3\ell/2} p^r (\ell)^{3\ell+3} \) and this completes the proof. \( \square \)

**Theorem 6.4.** With high probability, \( \|B\| \leq O\left(\frac{m \log^3 n}{n^{3/2}}\right) \)
Proof. We proceed by using Markov’s inequality:
\[
\Pr[\|B\| \geq n^{3/2}p(2\ell)^3] = \Pr\left[\|B\| \geq \left(n^{3/2}p(2\ell)^3\right)^\ell \right] \leq \frac{\mathbb{E}[\text{Tr}(BB^TBB^T...BB^T)]}{n^{3\ell/2}p(2\ell)^{3\ell}} \leq \frac{\ell^3}{2^{3\ell}}
\]
and hence setting \(\ell = \Theta(\log n)\) we conclude that \(\|B\| \leq 8n^{3/2}p\log^3 n\) holds with high probability. Moreover, \(B = \sum_{r=1}^{O(\log n)} B^r\) also holds with high probability. If this equality holds and each \(B^r\) satisfies \(\|B^r\| \leq 8n^{3/2}p\log^3 n\), we have
\[
\|B\| \leq \max_r O(\|B^r\| \log n) = O\left(\frac{m \log^4 n}{n^{3/2}}\right)
\]
where we have used the fact that \(p = m/n^3\). This completes the proof of the theorem.

\[\square\]

Proofs of Theorem 1.3, Theorem 1.1, Theorem 1.4 and Corollary 1.2

At last with these spectral bounds in hand, we can now prove Theorem 1.3:

Proof. Recall the formula for the Rademacher complexity that we gave in Section 5. Let \(Y\) be an arbitrary element of \(K_6\). Then using Lemma 5.2 and Theorem 6.4 we have
\[
\left(\langle Z, Y \rangle\right)^2 \leq n \left(\sum_i \left(\sum_{j,k} Z_{i,j,k} Y_{i,j,k}\right)^2\right) \leq C^2\|B\| + C^6 m n^2 = O\left(\frac{m \log^4 n}{n^{3/2}}\right)
\]
Hence the Rademacher complexity can be bounded as
\[
\frac{1}{m} \left(\langle Z, Y \rangle\right) \leq O\left(\sqrt{\frac{\log^4 n}{m n^{3/2}}}\right)
\]
which completes the proof of the theorem.

Recall that bounds on the Rademacher complexity readily imply bounds on the generalization error (see Theorem 4.2). Hence we can now prove Theorem 1.1:

Proof. We can make use of the optimization problem \((K)\) plugging in the norm \(\|\cdot\|_{K_6}\). Since this norm comes from the sixth level of the sum-of-squares hierarchy, it follows that the resulting convex program is an \(n^6\)-sized semidefinite program and there is an efficient algorithm to solve it to arbitrary accuracy. Moreover we can always plug in \(X = T\), and the bounds on the noise and standard concentration bounds readily imply that with high probability \(X = T\) is a feasible solution. Hence with high probability, the convex program has a feasible solution \(X\) with \(\|X\|_{K_6} \leq r\). Now if we take any \(X\) returned by the convex program, its empirical error is at most \(2\eta\), and since \(\|X\|_{K_6} \leq r\) with high probability, the bounds on the Rademacher complexity (Theorem 1.3) and standard results about how to translate such bounds into bounds on the generalization error (Theorem 4.2) now imply our main theorem.

\[\square\]

Next we give an application to strongly refuting random 3-XOR formulas. Recall, the algorithm referred to in the corollary below was given in Section 2, and we analyze it here.

Corollary 6.5. There is a polynomial time algorithm that outputs a quantity \(\text{val}\) that satisfies

(a) \(\text{val}\) is an upper bound on the fraction of clauses that can be satisfied

(b) \(\text{val}\) is \(1/2 + o(1)\) with high probability on a random 3-XOR formula with \(m\) constraints whenever \(m = \omega(n^{3/2} \log^4 n)\).
Proof. Recall that we have chosen the convention that a 3-XOR formula $\phi$ is a collection of constraints of the form $v_i v_j v_k = Z_{i,j,k}$ where $v_i, v_j, v_k, Z_{i,j,k} \in \{\pm 1\}$. Then any assignment

$$\sigma : \{v_i\}_{i=1}^n \to \{\pm 1\}$$

can be transformed into a unit vector $a \in \mathbb{R}^n$ where $a_i \equiv v_i / \sqrt{n}$. It is immediate that $a$ is 1-balanced and hence $a \otimes^3 \in A$. Recall that we output $val = 1/2 + opt$ where $opt$ is the optimum value for solving $(R)$ with $K_a$ using the transformation from $\phi$ to $\Omega$ outlined in Section 2. Since $\mathcal{A} \subseteq K_a$ it follows that any assignment that satisfies a $c$-fraction of the $m$ clauses will yield a feasible solution to $(R)$ whose objective function is at least $2c - 1$ and

$$c \leq 1/2 + (2c - 1) \leq 1/2 + opt$$

as desired. Thus Condition (a) holds. Moreover the bound on the Rademacher complexity (Theorem 1.3) immediately implies that $opt$ is $o(1)$ whenever $m = \omega(n^{3/2} \log^4 n)$ which implies Condition (b) holds too. \qed

Finally we prove Corollary 1.2:

Proof. Our goal is to lower bound the absolute value of a typical entry in $T$. Recall that each vector $a_i$ has its entries at least $1/D \sqrt{n}$ in absolute value where we can think of $D$ as polylogarithmic, or even larger if we so choose. Then each entry in $a \otimes a \otimes a$ is at least $1/D^3 n^{3/2}$. We will invoke the Littlewood-Offord Lemma, as proved by Erdös:

**Lemma 6.6.** [35] If $|v_i| \geq 1$ for all $i$, then the probability that $|\sum_{i=1}^r \sigma_i v_i| \leq 1$ is at most $1/\sqrt{r}$.

Hence, since $T = \sum_{i=1}^r \sigma_i a \otimes a \otimes a$, we have that any particular entry of $T$ has probability $1 - 1/\log^2 r$ is at least $r^{1/2}/D^3 n^{3/2} \log r$. Moreover this probability is $1 - o(1)$ since we have assumed that $r = \omega(1)$. Let $f(r, n, D) = r^{1/2}/D^3 n^{3/2} \log^2 r$. It follows from Markov’s bound that with probability $1 - 1/\log r$, all but a $1 - 1/\log r$ fraction of the entries satisfy $|T_{i,j,k}| \geq f(r, n, D)$. Let this set be $\mathcal{R} \subseteq [n] \times [n] \times [n]$ and let $\mathcal{R}' \subseteq \mathcal{R}$ be the subset of $\mathcal{R}$ where $\text{sgn}(X_{i,j,k}) \neq \text{sgn}(T_{i,j,k})$. Then we have

$$\text{err}(X) \geq \frac{|\mathcal{R}'|}{n^3} f(r, n, D)$$

We can now invoke Theorem 1.1 which guarantees that the hypothesis $X$ that results from solving $(K)$ satisfies $\text{err}(X) = f(r, n, D)/\log r$ with probability $1 - 1/\log r$ provided that $m = \Omega(n^{3/2} r \log^4 n \log^2 r)$. This bound on the error immediately implies that $|\mathcal{R}'| \leq n^3 / \log r$ and so $|\mathcal{R} \setminus \mathcal{R}'| = (1 - 1/\log r)n^3$. This completes the proof of the corollary in the noiseless case. It is easy to see that an identical argument works in the presence of noise, provided that $\max_{i,j,k} |\eta_{i,j,k}| \leq f(r, n, D)/\log^2 r$. \qed

7 Sum-of-Squares Lower Bounds

Recall that in Definition 3.4 we introduced the norm $K_k$ that results from the $k$th level sum-of-squares relaxation to the atomic norm. Here we will show strong lower bounds on the Rademacher complexity of sum-of-squares relaxations, and that the resolution norm is essentially the best we can do. Our lower bounds follow as a corollary from known lower bounds for refuting random instances of 3-XOR [47, 69]. First we need to introduce the formulation of the sum-of-squares hierarchy used in [69]: We will call a Boolean function $f$ a $k$-junta if there is set $S \subseteq [n]$ of at most $k$ variables so that $f$ is determined by the values in $S$.

**Definition 7.1.** The $k$-round Lasserre hierarchy is the following relaxation:

(a) $\|v_0\|^2 = 1$, $\|v_C\|^2 = 1$ for all $C \in C$

(b) $\langle v_f, v_g \rangle = \langle v_f', v_g' \rangle$ for all $f, g, f', g'$ that are $k$-juntas and $f \cdot g \equiv f' \cdot g'$

(c) $v_f + v_g = v_{f+g}$ for all $f, g$ that are $k$-juntas and satisfy $f \cdot g \equiv 0$
Here we define a vector $v_f$ for each $k$-junta, and $C$ is a class of constraints that must be satisfied by any Boolean solution (and are necessarily $k$-juntas themselves). See [69] for more background, but it is easy to construct a feasible solution to the above convex program given a distribution on feasible solutions for some constraint satisfaction problem. In the above relaxation, we think of functions $f$ as being $\{0,1\}$-valued. It will be more convenient to work with an intermediate relaxation where functions are $\{-1,1\}$-valued and the intuition is that $u_S$ for some set $S \subseteq [n]$ should correspond to the vector for the character $\chi_S$.

**Definition 7.2.** Alternatively, the $k$-round Lasserre hierarchy is the following relaxation:

(a) $\|u_0\|^2 = 1$, $(u_0, u_S) = (-1)^{Z_S}$ for all $(\oplus_S, Z_S) \in C$

(b) $(u_S, u_T) = (u_{S'}, u_{T'})$ for sets $S, T, S', T'$ that are size at most $k$ and satisfy $S \Delta T = S' \Delta T'$, where $\Delta$ is the symmetric difference.

Here we have explicitly made the switch to XOR-constraints — namely $(\oplus_S, Z_S)$ has $Z_S \in \{0,1\}$ and correspond to the constraint that the parity on the set $S$ is equal to $Z_S$. Now if we have a feasible solution to the constraints in Definition 7.1 where all the clauses are XOR-constraints, we can construct a feasible solution to the constraints in Definition 7.2 as follows. If $S$ is a set of size at most $k$, we define

$$u_S \equiv v_g - v_f$$

where $f$ is the parity function on $S$ and $g = 1 - f$ is its complement. Moreover let $u_0 = v_0$.

**Claim 7.3.** \{u_S\} is a feasible solution to the constraints in Definition 7.2

**Proof.** Consider Constraint (b) in Definition 7.2, and let $S, T, S', T'$ be sets of size at most $k$ that satisfy $S \oplus T = S' \oplus T'$. Then our goal is to show that

$$\left\langle v_{g_S} - v_{f_S}, v_{g_T} - v_{f_T} \right\rangle = \left\langle v_{g_{S'}} - v_{f_{S'}}, v_{g_{T'}} - v_{f_{T'}} \right\rangle$$

where $f_S$ is the parity function on $S$, and similarly for the other functions. Then we have $f_S \cdot f_T \equiv f_{S'} \cdot f_{T'}$, because $S \oplus T = S' \oplus T'$, and this implies that $\langle v_f, v_f \rangle = \langle v_{f_{S'}}, v_{f_{T'}} \rangle$. An identical argument holds for the other terms. This implies that all the Constraints (b) hold. Similarly suppose $(\oplus_S, Z_S) \in C$. Since $f_S \cdot g_S = 0$ and $f_S + g_S = 1$ it is well-known that (1) $v_{f_S}$ and $v_{g_S}$ are orthogonal (2) $v_{f_S} + v_{g_S} = v_0$ and (3) since $f_S \in C$ in Definition 7.1, we have $v_{g_S} = 0$ (see [69]). Thus

$$\langle u_0, u_S \rangle = \langle v_0, u_S \rangle - \langle v_0, v_{f_S} \rangle = -1$$

and this completes the proof. $\square$

Now, following Barak et al. [9] we can use the constraints in Definition 7.2 to define the operator $\mathcal{L}(\cdot)$. In particular, given $p \in P^k$ where $p \equiv \sum_S c_S \prod_{i \in S} Y_i$ and $p$ is multilinear, we set

$$\mathcal{L}(p) = \sum_S c_S \left( \frac{1}{n} \right)^{|S|/2} \langle u_0, u_S \rangle$$

Here we will also need to define $\mathcal{L}(p)$ when $p$ is not multilinear, and in that case if $Y_i$ appears $2d$ times we replace it with $(1/n)^d$ and if it appears $2d+1$ times we replace it by $(1/n)^d Y_i$ to get a multilinear polynomial $q$ and then set $\mathcal{L}(p) = \mathcal{L}(q)$.

**Claim 7.4.** $\mathcal{L}(\cdot)$ is a feasible solution to the constraints in Definition 3.4, and for any $(\oplus_S, Z_S) \in C$ we have $\mathcal{L}(\prod_{i \in S} Y_i) = (-1)^{2Z_S (1/n)^{|S|/2}}$.

**Proof.** Then by construction $\mathcal{L}(1) = 1$, and the proof that $\mathcal{L}(p^2) \geq 0$ is given in [9], but we repeat it here for completeness. Let $p = \sum_S c_S \prod_{i \in S} Y_i$ be multilinear where we follow the above recipe and replace terms of the form $Y_i^2$ with $(1/n)$ as needed. Then $p^2 = \sum_{S,T} c_{ST} \prod_{i \in S} Y_i \prod_{j \in T} Y_j$ and moreover

$$\mathcal{L}(p^2) = \sum_{S,T} c_{ST} \left( \frac{1}{n} \right)^{|S\Delta T|/2} \langle u_0, u_{S\Delta T} \rangle$$

$$= \sum_{S,T} c_S \left( \frac{1}{n} \right)^{|S|/2} c_T \left( \frac{1}{n} \right)^{|T|/2} \langle u_S, u_T \rangle = \left\| \sum_S c_S (1/n)^{|S|/2} u_S \right\|^2 \geq 0$$

16
as desired. Next we must verify that $\mathcal{L}(\cdot)$ satisfies the constraints \{\sum_{i=1}^{n} Y_{i}^{2} = 1\} and \{Y_{i}^{2} \leq \frac{C^{2}}{n}\} for all $i \in \{1, 2, ..., n\}$, in accordance with Definition 3.3. To that end, observe that

$$\mathcal{L}\left(\left(\sum_{i=1}^{n} Y_{i}^{2} - 1\right)q\right) = \mathcal{L}\left(\left(\sum_{i=1}^{n} \frac{1}{n} - 1\right)q\right) = 0$$

which holds for any polynomial $q \in P_{k-2}^{n}$. Finally consider

$$\mathcal{L}\left(\left(\frac{C^{2}}{n} - Y_{i}^{2}\right)q^{2}\right) = \mathcal{L}\left(\left(\frac{C^{2}}{n} - \frac{1}{n}\right)q^{2}\right) \geq 0$$

which follows because $C > 1$ and holds for any polynomial $q \in P_{\lfloor (d-d')/2\rfloor}^{n}$. This completes the proof.

Theorem 7.5. [47, 69] Let $\phi$ be a random 3-XOR formula with $m = n^{3/2-\epsilon}$ constraints. Then for any $\epsilon > 0$ and any $c < 2$, the $k = \Omega(n^{c\epsilon})$ round Lasserre hierarchy given in Definition 7.1 permits a feasible solution, with probability $1 - o(1)$.

Note that the constant in the $\Omega(\cdot)$ depends on $\epsilon$ and $c$. Then using the above reductions, we have the following as an immediate corollary:

Corollary 7.6. For any $\epsilon > 0$ and any $c < 2$ and $k = \Omega(n^{c\epsilon})$, if $m = n^{3/2-\epsilon}$ the Rademacher complexity $R^{n}(K_{k}) = \frac{1 - o(1)}{n^{3/2}}$.

In particular, for third order tensors the resolution norm has Rademacher complexity $o(1/n^{3/2})$ whenever $m = \omega(n^{3/2} \log^{4} n)$ with probability $1 - o(1)$ and in contrast the Rademacher complexity of even very strong relaxations derived from essentially $n^{2\epsilon}$ rounds of the sum-of-squares hierarchy — which would require time $2^{n^{\epsilon}}$ to optimize over — achieve no non-trivial generalization bounds for $m = n^{3/2-\epsilon}$.

Similar results hold for arbitrary order tensors. We can define the atomic norm for an $d$th order tensor analogously to Definition 3.2, in which case $\mathcal{A} \subseteq \mathbb{R}^{n^{d}}$. We can also extend Definition 3.4 to this setting, and let $K_{k}^{d}$ be the norm that results from the $k$th level of the sum-of-squares relaxation. Then the following is an immediate corollary of the more general results in [47, 69] that provide lower bounds for random $d$-XOR formulas:

Corollary 7.7. For any $d > 2$, $\epsilon > 0$ and $c < \frac{1}{r/2-1}$ and $k = \Omega(n^{c\epsilon})$, if $m = n^{d/2-\epsilon}$ the Rademacher complexity $R^{n}(K_{k}^{d}) = \frac{1 - o(1)}{n^{d/2}}$.

In particular for even values of $d$, we can perform tensor completion for an $d$-order tensor by flattening it into an $n^{d/2} \times n^{d/2}$ matrix and ignoring the tensor structure entirely. As naive as this algorithm is, it is essentially optimal in that any algorithm derived from the sum-of-squares hierarchy and that runs in polynomial time would need to use almost as many samples.

Conclusions

An important open question is whether our algorithm is optimal in the sense that there is no polynomial-time algorithm (even one not using the sum-of-squares framework) that can predict a low rank tensor from significantly fewer measurements. If it is, then this problem truly possesses a computational phase transition in the number of measurements that is distinct from the information theoretical phase transition needed for inefficient prediction. We conjecture that this is the case. In the language of 3-XOR, this amounts to the following:

Conjecture 1. Let $\phi$ be a random $d$-XOR formula on $n$ variables with $m$ clauses where $d \geq 3$. Then for any $\epsilon > 0$, there is no polynomial time algorithm that outputs a quantity $\text{val}$ that satisfies the conditions

(a) $\text{val}$ is an upper bound on the fraction of clauses that can be satisfied

(b) $\text{val}$ is $1/2 + o(1)$ with high probability, when $m = n^{d/2-\epsilon}$
We remark that our conjecture (for \( d = 3 \)) is implied\(^6\) by Conjecture 2.2 in [32] who gave sample complexity vs. computational complexity tradeoffs for a problem of learning halfspaces over sparse vectors. We believe that there are many further directions to explore about the implications of conjectures like the one above, in terms of sample complexity vs. computational complexity tradeoffs. See also [12] for a unified discussion of the average-case complexity of CSPs and (conjectured) relations to lower bounds for semidefinite programming.

It is also an interesting open question to explore how sharp the computational complexity predictions are, that are made by sum-of-squares lower bounds. There are several notable cases where it is possible to speed up to some extent semidefinite programming based algorithms (e.g. [7, 70]) and in particular in some special cases one can solve the \( k \) round relaxation in time \( C^k \text{poly}(n) \) instead of in \( n^{O(k)} \) time [13]. Thus even if the sum-of-squares algorithm is approximately optimal for a number of problems, there may still be hope to speed it up in some applications. The question in this direction most related to this work is to find out whether it is possible to speed up our tensor prediction algorithm. In particular, it would be very interesting to give provable guarantees for alternating minimization, as it is the heuristic of choice in practice. It is only known that alternating minimization works when the factors are orthonormal [56], and there are no guarantees in the much more general setting we consider here.

Acknowledgements

We would like to thank Aram Harrow for many helpful discussions, and also Ryan O’Donnell for providing us with a copy of [2].

\(^6\)It follows from Feige’s XOR principle [37] that any algorithm for strongly refuting random 3-XOR can be used to \( \epsilon \)-refute random 3-SAT with the same number of clauses. We thank Ryan O’Donnell for this observation. Note that Feige, Kim and Ofek [41] gave a non-deterministic algorithm for refuting for random 3SAT with \( m \sim n^{1.4} \) clauses by showing that there is a short but potentially hard to find certificate of unsatisfiability. However their work does not provide a \( 1 - \epsilon \) upper bound on the fraction of clauses that can be satisfied.
References

[1] M. Alekhnovich. More on average case vs approximation complexity. In Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science (FOCS), pages 298–307, 2003.

[2] S. Allen, R. O’Donnell and D. Witmer. Private Communication, 2015.

[3] A. Anandkumar, D. Foster, D. Hsu, S. Kakade, Y. Liu. A spectral algorithm for latent dirichlet allocation. In Advances in Neural Information Processing Systems 25 (NIPS), pages 926–934, 2012.

[4] A. Anandkumar, R. Ge, D. Hsu and S. Kakade. A tensor spectral approach to learning mixed membership community models. In Proceedings of the 26th Conference on Learning Theory (COLT), pages 867–881, 2013.

[5] A. Anandkumar, D. Hsu and S. Kakade. A method of moments for mixture models and hidden markov models. In Proceedings of the 25th Conference on Learning Theory (COLT), pages 1–33, 2012.

[6] B. Applebaum, B. Barak and D. Xiao. On basing lower-bounds for learning on worst-case assumptions. In Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science (FOCS), pages 211–220, 2008.

[7] S. Arora, E. Hazan and S. Kale. \(O(\sqrt{\log(n)})\) Approximation to SPARSEST CUT in \(O(n^2)\) Time. In SIAM J. Comput., 39(5), pages 1748–1771, 2010. Preliminary version in FOCS ‘04.

[8] N. Balcan. Machine Learning Theory Notes. http://www.cc.gatech.edu/~ninamf/ML11/lect1115.pdf

[9] B. Barak, F. Brandao, A. Harrow, J. Kelner, D. Steurer and Y. Zhou. Hypercontractivity, sum-of-squares proofs, and their applications. In Proceedings of the 44th ACM Symposium on Theory of Computing (STOC), pages 307–326, 2012.

[10] B. Barak, J. Kelner and D. Steurer. Rounding sum-of-squares relaxations. In Proceedings of the 46th ACM Symposium on Theory of Computing (STOC), pages 31–40, 2014.

[11] B. Barak, J. Kelner and D. Steurer. Dictionary learning and tensor decomposition via the sum-of-squares method. arXiv:1407.1543, 2014

[12] B. Barak, G. Kindler and D. Steurer. On the optimality of semidefinite relaxations for average-case and generalized constraint satisfaction. In 4th Innovations in Theoretical Computer Science, pages 197–214, 2013.

[13] B. Barak, D. Steurer and P. Raghavendra. Rounding semidefinite programming hierarchies via global correlation. In Proceedings of the 52nd Annual IEEE Symposium on Foundations of Computer Science (FOCS), pages 472–481, 2011.

[14] B. Barak and D. Steurer. Sum-of-squares proofs and the quest toward optimal algorithms/ In Proceedings of International Congress of Mathematicians (ICM), 2014, to appear.

[15] P. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. Journal of Machine Learning Research, 3:463–482, 2003.

[16] E. Ben-Sasson and A. Wigderson. Short proofs are narrow — resolution made simple. Journal of the ACM, 48(2):149–169, 2001.

[17] Q. Berthet and P. Rigollet. Computational lower bounds for sparse principal component detection. In Proceedings of the 26th Conference on Learning Theory (COLT), pages 1046–1066, 2013.

[18] A. Bhaskara, M. Charikar, A. Moitra and A. Vijayaraghavan. Smoothed analysis of tensor decompositions. In Proceedings of the 46th ACM Symposium on Theory of Computing (STOC), pages 594–603, 2014.
[19] A. Bogdanov and L. Trevisan. On worst-case to average-case reductions for \textit{NP}-problems. \textit{SIAM Journal on Computing}, 36(4):1119–1159, 2006.

[20] E. Candes, Y. Eldar, T. Strohmer and V. Voroninski. Phase retrieval via matrix completion. \textit{SIAM Journal on Imaging Sciences}, 6(1):199–225, 2013.

[21] E. Candes and C. Fernandez-Granda. Towards a mathematical theory of super-resolution. \textit{Communications on Pure and Applied Mathematics}, 67(6):906–956, 2014.

[22] E. Candes, X. Li, Y. Ma and J. Wright. Robust principal component analysis? \textit{Journal of the ACM}, 58(3):1–37, 2011.

[23] E. Candes and Y. Plan. Matrix completion with noise. \textit{Proceedings of the IEEE}, 98(6):925–936, 2010.

[24] E. Candes and B. Recht. Exact matrix completion via convex optimization. \textit{Foundations of Computational Math.}, 9(6):717–772, 2008.

[25] E. Candes, J. Romberg and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. \textit{Communications of Pure and Applied Math.}, pp. 1207–1223, 2006.

[26] E. Candes and T. Tao. Decoding by linear programming. \textit{IEEE Trans. on Information Theory}, 51(12):4203–4215, 2005.

[27] E. Candes and T. Tao. The power of convex relaxation: Near-optimal matrix completion. \textit{IEEE Transactions on Information Theory}, 56(5):2053-2080, 2010

[28] A. Carbery and J. Wright. Distributional and $\ell^p$ norm inequalities for polynomials over convex bodies in $\mathbb{R}^n$. \textit{Mathematics Research Letters}, 8(3):233–248, 2001.

[29] V. Chandrasekaran and M. Jordan. Computational and statistical tradeoffs via convex relaxation. \textit{Proceedings of the National Academy of Sciences}, 110(13):E1181–E1190, 2013.

[30] V. Chandrasekaran, B. Recht, P. Parrilo and A. Willsky. The convex geometry of linear inverse problems. \textit{Foundations of Computational Math.}, 12(6)805–849, 2012.

[31] A. Coja-Oghlan, A. Goerdt and A. Lanka. Strong refutation heuristics for random $k$-SAT. \textit{Combinatorics, Probability and Computing}, 16(1):5–28, 2007.

[32] A. Daniely, N. Linial and S. Shalev-Shwartz. More data speeds up training time in learning half spaces over sparse vectors. In \textit{Advances in Neural Information Processing Systems (NIPS)}, pages 145–153, 2013.

[33] A. Daniely, N. Linial and S. Shalev-Shwartz. From average case complexity to improper learning complexity. In \textit{Proceedings of the 46th ACM Symposium on Theory of Computing (STOC)}, pages 441–448, 2014.

[34] D. Donoho. Compressed sensing. \textit{IEEE Transactions on Information Theory}, 52(4):1289–1306, 2006.

[35] P. Erdős. On a lemma of Littlewood and Offord. \textit{Bulletin of the American Mathematical Society}, pages 898–902, 1945.

[36] M. Fazel. \textit{Matrix Rank Minimization with Applications}. PhD thesis, Stanford University, 2002.

[37] U. Feige. Relations between average case complexity and approximation complexity. In \textit{Proceedings of the 34th ACM Symposium on Theory of Computing (STOC)}, pages 534–543, 2002.

[38] U. Feige, J.H. Kim and E. Ofek. Witnesses for non-satisfiability of dense random 3CNF formulas In \textit{Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)}, pages 497–508, 2006.

[39] U. Feige and E. Ofek. Easily refutable subformulas of large random 3-CNF formulas. \textit{Theory of Computing} 3:25–43, 2007.
[40] J. Friedman, A. Goerdt and M. Krivelevich. Recognizing more unsatisfiable random $k$-SAT instances efficiently. *SIAM Journal on Computing* 35(2):408–430, 2005.

[41] U. Feige, J. H. Kim and E. Ofek. Witnesses for non-satisfiability of dense random 3CNF formulas *FOCS*, 497–508, 2006.

[42] J. Friedman, J. Kahn and E. Szemerédi. On the second eigenvalue of random regular graphs. In *Proceedings of the 21th ACM Symposium on Theory of Computing (STOC)*, pages 534–543, 1989.

[43] Z. Füredi and J. Komlós. The eigenvalues of random symmetric matrices. *Combinatorica*, 1:233–241, 1981.

[44] R. Ge and T. Ma. Private Communication, 2015.

[45] S. Gharibian. Strong $NP$-hardness of the quantum separability problem. *Quantum Information and Computation*, 10(3-4):343–360, 2010.

[46] A. Goerdt and M. Krivelevich. Efficient recognition of random unsatisfiable $k$-SAT instances by spectral methods. In *Annual Symposium on Theoretical Aspects of Computer Science*, pages 294–304, 2001.

[47] D. Grigoriev. Linear lower bound on degrees of Positivstellensatz calculus proofs for the parity. *Theoretical Computer Science* 259(1-2):613–622, 2001.

[48] L. Gurvits. Classical deterministic complexity of Edmonds’ problem and quantum entanglement. In *Proceedings of the 35th ACM Symposium on Theory of Computing (STOC)*, pages 10–19, 2003.

[49] M. Hardt. Understanding alternating minimization for matrix completion. In *Proceedings of the 55th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 651–660, 2014.

[50] A. Harrow and A. Montanaro. Testing product states, quantum merlin-arthur games and tensor optimization. *Journal of the ACM*, 60(1):1–43, 2013.

[51] C. Hillar and L-H. Lim. Most tensor problems are $NP$-hard. *Journal of the ACM*, 60(6):1–39, 2013.

[52] S. Hopkins, T. Schramm, J. Shi and D. Steurer. Private Communication, 2015.

[53] R. Horn and C. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.

[54] D. Hsu and S. Kakade. Learning mixtures of spherical gaussians: Moment methods and spectral decompositions. In *4th Annual Innovations in Theoretical Computer Science (ITCS)*, pages 11–20, 2013.

[55] P. Jain, P. Netrapalli and S. Sanghavi. Low rank matrix completion using alternating minimization. In *Proceedings of the 45th ACM Symposium on Theory of Computing (STOC)*, pages 665–674, 2013.

[56] P. Jain and S. Oh. Provable tensor factorization with missing data. In *Advances in Neural Information Processing Systems 27*, pages 1431–1439, 2014.

[57] R. Keshavan, A. Montanari and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

[58] V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics* 30(1):1–50, 2002.

[59] J. Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11(3):796–817, 2001.

[60] J. Lasserre. *Moments, Positive Polynomials and Their Applications* Imperial College Press, 2009.

[61] Y.K. Liu. *The Complexity of the Consistency and $N$ Representability Problems for Quantum States*. PhD thesis, University of California, San Diego, 2007.
A Reduction from Asymmetric to Symmetric Tensors

Here we give a general reduction, and show that any algorithm for tensor prediction that works for symmetric tensors can be used to predict the entries of an asymmetric tensor too. Hardt gave a related reduction for the cases of matrices [49] and it is instructive to first understand this reduction, before proceeding to the tensor case. Suppose we are given a matrix $M$ that is not necessarily symmetric. Then the approach of [49] is to construct the following symmetric matrix:

$$S = \begin{bmatrix} 0 & M^T \\ M & 0 \end{bmatrix}.$$ 

We have not precisely defined the notion of incoherence that is used in the matrix completion literature, but it turns out to be easy to see that $S$ is low rank and incoherent as well.

The important point is that given $m$ samples generated uniformly at random from $M$, we can generate random samples from $S$ too. It will be more convenient to think of these random samples as being generated without replacement, but this reduction works just as well without replacement too. Let $M \in \mathbb{R}^{n_1 \times n_2}$. Now for each sample from $S$, with probability $p = \frac{n_1^2 + n_2^2}{(n_1 + n_2)^2}$ we reveal a uniformly random entry in the either block of zeros. And with probability $1 - p$ we reveal a uniformly random entry from $M$. Each entry in $M$ appears exactly twice in $S$, and we choose to reveal this entry of $M$ with probability $1/2$ from the top-right block, and otherwise from the bottom-left block. Thus given $m$ samples from $M$, we can generate from $S$
(in fact we can generate even more, because some of the revealed entries will be zeros). It is easy to see that this approach works for the case of sampling without replacement to, in that \( m \) samples without replacement from \( M \) can be used to generate at least \( m \) samples without replacement from \( S \).

Now let us proceed to the tensor case. Let us introduce the following definition, for ease of notation:

**Definition A.1.** Let \( m(n, r, \delta, f, C) \) be such that, there is an algorithm that on a rank \( r \), order \( d \), size \( n \times n \times \ldots \times n \) symmetric tensor where each factor has norm at most \( C \), the algorithm returns an estimate \( X \) with \( \text{err}(X) = f \) with probability \( 1 - \delta \) when it is given \( m(n, r, \delta, f) \) samples chosen uniformly at random (and without replacement).

**Lemma A.2.** For any odd \( d \), suppose we are given \( m(\sum_{j=1}^{d} n_j, r2^{d-1}, \delta, f, \sqrt{d}) \) samples chosen uniformly at random (and without replacement) from an \( n_1 \times n_2 \times \ldots \times n_d \) tensor

\[
T = \sum_{i=1}^{r} a_i^1 \otimes a_i^2 \otimes \ldots \otimes a_i^d
\]

where each factor is unit norm. There is an algorithm that with probability at least \( 1 - \delta \) returns an estimate \( Y \) with

\[
\text{err}(Y) \leq \frac{(\sum_{j=1}^{d} n_j)^d}{d!2^{d-1} \prod_{j=1}^{d} n_j}
\]

**Proof.** Our goal is to symmetrize an asymmetric tensor, and in such a way that each entry in the symmetrized tensor is either zero or else correspond to an entry in the original tensor. Our reduction will work for any odd order \( d \) tensor. In particular let

\[
T = \sum_{i=1}^{r} a_i^1 \otimes a_i^2 \otimes \ldots \otimes a_i^d
\]

be an order \( d \) tensor where the dimension of \( a_i^j \) is \( n_j \). Also let \( n = \sum_{j=1}^{d} n_j \). Then we will construct a symmetric, order \( d \) tensor as follows. Let \( \sigma_1, \sigma_2, \ldots \sigma_d \) be a collection of \( d \) random \( \pm \) variables that are chosen uniformly at random from the \( 2^{d-1} \) configurations where \( \prod_{j=1}^{d} \sigma_j = 1 \). Then we consider the following random vector

\[
a_i(\sigma_1, \sigma_2, \ldots \sigma_d) = [\sigma_1 a_i^1, \sigma_2 a_i^2, \ldots, \sigma_d a_i^d]
\]

Here \( a_i(\sigma_1, \sigma_2, \ldots \sigma_d) \) is an \( n \)-dimensional vector that results from concatenating the vectors \( a_i^1, a_i^2, \ldots, a_i^d \) but after flipping some of their signs according to \( \sigma_1, \sigma_2, \ldots \sigma_d \). Then we set

\[
S = \mathbb{E}_{\sigma_1, \sigma_2, \ldots \sigma_d} \left[ \sum_{i=1}^{r} a_i(\sigma_1, \sigma_2, \ldots \sigma_d) \right]^{\otimes d}
\]

It is immediate that \( S \) is symmetric and has rank at most \( 2^{d-1}r \) by expanding out the expectation into a sum over the valid sign configurations. Moreover each rank one term in the decomposition is of the form \( a^{\otimes d} \) where \( \|a\|_2^2 = d \) because it is the concatenation of \( d \) unit vectors.

If \( \sigma_1, \sigma_2, \ldots \sigma_d \) is fixed, then each entry in \( S \) is itself a degree \( d \) polynomial in the \( \sigma_j \) variables. By our construction of the \( \sigma_j \) variables, and because \( d \) is odd so there are no terms where every variable appears to an even power, it follows that all the terms vanish in expectation except for the terms which have a factor of \( \prod_{j=1}^{d} \sigma_j \), and these are exactly terms that correspond to some permutation \( \pi : [d] \rightarrow [d] \), and a term of the form

\[
\sum_{i=1}^{d} a_i^{\pi(1)} \otimes a_i^{\pi(2)} \otimes \ldots \otimes a_i^{\pi(d)}
\]

Hence all of the entries in \( S \) are either zero or are \( 2^{d-1} \) times an entry in \( T \). As before, we can generate \( m \) uniformly random samples from \( S \) given \( m \) uniformly random samples from \( T \), by simply choosing to sample an entry from one of the blocks of zeros with the appropriate probability, or else revealing an entry of \( T \) and choosing where in \( S \) to reveal this entry uniformly at random. Hence:

\[
\frac{1}{(\sum_{j=1}^{d} n_j)^d} \sum_{(i_1, i_2, \ldots, i_d) \in \Gamma} |Y_{i_1, i_2, \ldots, i_d} - S_{i_1, i_2, \ldots, i_d}| \leq \frac{1}{(\sum_{j=1}^{d} n_j)^d} \sum_{i_1, i_2, \ldots, i_d} |Y_{i_1, i_2, \ldots, i_d} - S_{i_1, i_2, \ldots, i_d}|
\]
where \( \Gamma \) represents the locations in \( S \) where an entry of \( T \) appears. The right hand side above is at most \( f \) with probability \( 1 - \delta \). Moreover each entry in \( T \) appears in exactly \( d! \) locations in \( S \). And when it does appear, it is scaled by \( 2^{d-1} \). And hence if we multiply the left hand side by

\[
\left( \sum_{j=1}^{d} n_j \right)^d / d! 2^{d-1} \prod_{j=1}^{d} n_j
\]

we obtain \( \text{err}(Y) \). This completes the reduction.

Note that in the case where \( n_1 = n_2 = n_3 = \ldots = n_d \), the error and the rank in this reduction increase only by at most an \( e^d \) and \( 2^d \) factor respectively.

## B Extensions to Higher Order Tensors

Our techniques immediately extend to the order \( d \) tensor prediction problem, for all odd \( d \). We have chosen to focus on the case of \( d = 3 \) to simplify the exposition, but a purely syntactic change in our arguments works more generally. Specifically, for odd \( d \geq 3 \) we can obtain the following bound on the Rademacher complexity:

\[
R^m(K_{2d}) \leq O\left( \sqrt{\log^4 n / mn^{d/2}} \right)
\]

The idea is if we are given a \( 5 \)-XOR constraint \( x_{i_1} \cdot x_{i_2} \cdot x_{i_3} \cdot x_{i_4} \cdot x_{i_5} = Z_{i_1,i_2,i_3,i_4,i_5} \) then we can view this instead as a \( 3 \)-XOR constraint \( x_i \cdot x_j \cdot x_k = Z_{i,j,k} \) where \( i = i_1, j = (i_2,i_3) \) and \( k = (i_4,i_5) \). Thus \( i \) takes on values from the set \([n]\) and \( j \) and \( k \) both take on values from the set \([n^2]\). More generally, we can map an \( d \)-XOR constraint to a \( 3 \)-XOR constraint where the variable \( i \) takes on values from the set \([n]\) and both \( j \) and \( k \) to take on values from the set \([n^{(d-1)/2}]\). Then the key ingredient is the following asymmetric variant of Definition 3.4:

**Definition B.1** (\( \text{SOS}_k \) norm). We let \( K_k \) be the set of all \( X \in \mathbb{R}^{n^3} \) such that there exists a degree \( k \) pseudo-expectation operator on \( P^i_k \) satisfying the constraints

(a1) \( \{ \sum_{i=1}^{n} (Y_i^{(1)})^2 = 1 \} \)

(a2) \( \{ \sum_{i=1}^{n^{(d-1)/2}} (Y_i^{(a)})^2 = 1 \} \) for \( a = 2, 3 \)

(b1) \( \{ (Y_i^{(1)})^2 \leq C^2 / n \} \) for all \( i \in \{1, 2, \ldots, n\} \)

(b2) \( \{ (Y_i^{(a)})^2 \leq C^{d-1} / n^{(d-1)/2} \} \) for all \( i \in \{1, 2, \ldots, n\} \) and \( a = 2, 3 \)

(c) \( X_{i,j,k} = \overline{E}[Y_i^{(1)} Y_j^{(2)} Y_k^{(3)}] \) for all \( i,j,k \).

The \( \text{SOS}_k \) norm of \( X \in \mathbb{R}^{n^3} \) is the infimum over \( \alpha \) such that \( X/\alpha \in K_k \).

Then if we set \( Q_{i,z}(Y) = \sum_{j,k} Z_{i,j,k} Y_j^{(2)} Y_k^{(3)} \) we again use Cauchy-Schwartz and properties of the pseudo-expectation operator (namely that \( (\overline{E}[p])^2 \leq \overline{E}[p^2] \)) as we did in Section 5, to obtain:

\[
\left( \langle Z, X \rangle \right)^2 \leq n \sum_i \overline{E} \left[ (Y_i Q_{i,z}(Y))^2 \right]
\]

Also it is easy to check that the following variant of Claim 5.4 holds, in the setting of Definition B.1:

**Claim B.2.** \( \overline{E}[Y_j^{(1)} Y_k^{(2)}] \leq C^{d-2} / n^{d-1} \)

And finally, because the variables \( i, j \) and \( k \) are now asymmetric and \( i \) takes on values in the domain \([n]\) and \( j \) and \( k \) take on values in the domain \([n^{(d-1)/2}]\), when we bound the number of possible encodings (for each term that contributes to the trace in Section 6) we obtain the following lemma:
Lemma B.3. $E[\text{Tr}(BB^TBB^T...BB^T)] \leq n^{d/2}p^f(t)^{3t+3}$

The proofs of these ingredients are identical and just involve accounting for the asymmetry in $i, j$ and $k$, so we do not repeat them here. These tools together yield the desired bound on the Rademacher complexity, and in fact all of the implications of a bound on the Rademacher complexity (but now for order $d$ tensors) such as Theorem 1.1, Theorem 1.4 and Corollary 1.2 go through. In particular, there is a polynomial time algorithm for order $d$ tensor prediction that achieves error $o(1/n^{d/2})$ with $m = \omega(n^{d/2} \log^4 n)$ observations. There is a polynomial time algorithm for recovering a $1 - o(1)$ fraction of the entries of a random order $d$ tensor up to a $1 \pm o(1)$ multiplicative factor with $m = \omega(n^{d/2} \log^4 n)$ observations. And finally there is a polynomial time algorithm for strongly refuting random $d$-XOR that succeeds with high probability with $m = \omega(n^{d/2} \log^4 n)$ clauses. We emphasize that for even $d$, bounds of this form hold by a direct reduction to the case of $d = 2$. But here we complete the picture for all order $d$. 

25