Sum-of-squares meets square loss: Fast rates for agnostic tensor completion

Dylan J. Foster
Massachusetts Institute of Technology

Andrej Risteski
Massachusetts Institute of Technology

Editors: Alina Beygelzimer and Daniel Hsu

1. Introduction

Recovering structured mathematical objects from partial measurements is a fundamental task in machine learning and statistical inference. One important example, which has been a mainstay of modern research in machine learning and high-dimensional statistics, is matrix completion. Here, we receive \( n \) entries from an unknown \( d \times d \) matrix, and the goal is to complete the remaining entries when \( n \) is as small as possible. The key structural assumption that enables recovery when \( n \ll d^2 \) is that the underlying matrix is low-rank. A celebrated line of research on matrix completion (Srebro and Shraibman, 2005; Candès and Recht, 2009; Candès and Tao, 2010; Keshavan et al., 2010; Gross, 2011; Recht, 2011) has culminated in the following guarantee: to exactly recover an incoherent rank-\( r \) matrix, \( n = \tilde{O}(rd) \) uniformly sampled entries suffice.

While low-rank matrix completion has seen successful application across many problem domains (most famously in the context of the Netflix Problem), for many tasks it is natural to consider not
just pairwise interactions but higher-order interactions, leading to the problem of \textit{tensor completion}. Tensor completion poses significant computational hurdles compared to the matrix case, but in an impressive recent work, Potechin and Steurer (2017) provide an efficient algorithm based on the sum-of-squares hierarchy that exactly recovers a \(d \times d \times d\) tensor of rank-\(r\) with incoherent and orthogonal components from \(\tilde{O}(rd^{8/3})\) measurements. While this undershoots the optimal statistical rate of \(\tilde{O}(rd)\), there is evidence that this is optimal amongst polynomial time algorithms under certain average-case hardness assumptions (Barak and Moitra, 2016).

In real-world applications, the assumption that the underlying tensor is truly low-rank may be too strong, and model misspecification is unavoidable. This is the main thrust of \textit{agnostic learning} (Haussler, 1992; Kearns et al., 1994) which, rather than attempting to recover an unknown model in some class, attempts to \textit{predict} as well as the best model. The aim of this paper is to develop guarantees for \textit{agnostic tensor completion}: predicting as well as the best low rank tensor, even when exact recovery is impossible.

We work in the following \textit{agnostic tensor regression} model, which captures tensor completion as a special case: we receive examples \((X_1, Y_1), \ldots, (X_n, Y_n)\) i.i.d. from an unknown distribution \(\mathcal{D}\), where each instance \(X_i\) is a \(d \times d \times d\) tensor and \(Y_i\) is a real-valued response. Letting \((\cdot, \cdot)\) denote the usual inner product, we measure predictive performance of a given \(d \times d \times d\) tensor \(T\) via its \textit{square loss risk} \(L_{\mathcal{D}}(T) = \mathbb{E}_{(X,Y) \sim \mathcal{D}}[(T, X) - Y]^2\). Our goal is to use the samples to produce a predictor \(\hat{T}_n\) that enjoy low \textit{excess risk}:

\[
L_{\mathcal{D}}(\hat{T}_n) = \inf_{T: \text{rank} \leq r} L_{\mathcal{D}}(T) \leq \varepsilon(n, r, d),
\]

where the bound \(\varepsilon(n, r, d)\) converges to zero as \(n \to \infty\). When the observations \(X\) are uniformly distributed indicators (i.e. \(X = e_i \otimes e_j \otimes e_k\), where \((i, j, k)\) is uniform) this recovers the usual measurement model for tensor completion. If the model is well-specified in the sense that \(Y = \langle X, T^* \rangle + \xi\), where \(T^*\) is a low-rank tensor and \(\mathbb{E}[\xi | X] = 0\), then low excess risk implies approximate recovery of \(T^*\). In general, the guarantee (1) is interesting because it implies non-trivial predictive performance even in the presence of severe model misspecification.

In the matrix case, agnostic excess risk guarantees of the type in (1) were characterized by Koltchinskii et al. (2011). There it was shown that to compete with the best rank-\(r\) matrix with bounded entries, empirical risk minimization with nuclear norm penalization obtains a \textit{fast rate} of the form \(\varepsilon(n, r, d) = O(rd \log d/n)\), and also showed that this is optimal. The nomenclature fast rate is intended to contrast the \(1/n\) dependency on \(n\) with \textit{slow rates}, which have a \(1/\sqrt{n}\) dependency on \(n\), and which are typically much simpler to obtain (see Gaiffas and Lecué (2011) for slow rates in the matrix setting).

The results of Koltchinskii et al. (2011) leverage strong understanding of the (matrix) nuclear norm, namely matrix concentration and decomposability/subgradient properties. In this paper, we tackle the following questions:

- Can we give similar guarantees for agnostic tensor completion?

- Can we obtain fast rates while at the same time relaxing the strong statistical assumptions (low rank observations, incoherence) needed for exact recovery?

- What are the best rates we can obtain subject to employing a polynomial-time algorithm?

\footnote{Interestingly, their results also show that incoherence—usually taken as necessary in matrix completion for positive results—is not necessary to obtain prediction error bounds; boundedness suffices.}
Our main result is to give the first polynomial time algorithm with a fast $O(1/n)$-type rate for agnostic completion of third order tensors that improves over the rate obtained by the natural reduction to matrix completion. Our main theorem gives excess risk bounds relative to low-rank orthogonal tensors, i.e. tensors of the form
\[ T = \sum_{i=1}^{r} \lambda_i \cdot u_i \otimes v_i \otimes w_i, \]
where $\|u_i\|_2 = \|v_i\|_2 = \|w_i\|_2 = 1$ and $\{u_i\}$ are orthogonal, as are $\{v_i\}$ and $\{w_i\}$. The result is as follows.

**Theorem 1 (informal)** There is a convex set of $d \times d \times d$ tensors $T$ derived from a sum-of-squares hierarchy relaxation of the tensor nuclear norm, for which the empirical risk minimizer $\hat{T}_n = \arg\min_{T \in T} \frac{1}{n} \sum_{i=1}^{n} (\langle T, X_i \rangle - Y_i)^2$ can be computed in polynomial time, and guarantees that with probability at least $1 - o(1)$,
\[ L_D(\hat{T}_n) - \inf_{T : \text{ rank-}r, \text{ orthogonal}} L_D(T) \leq \tilde{O}\left( \frac{r^2 d^{3/2}}{n} \right). \]

The guarantee applies to both random indicator measurements (tensor completion) and Gaussian measurements (tensor compressed sensing).

The full version of the theorem is stated in Section 3. The result be thought of as a generalization of the agnostic matrix completion results of Koltchinskii et al. (2011) to higher-order tensors. We also achieve a more general exact oracle inequality that trades off approximation and estimation error. This takes the form
\[ L_D(\hat{T}_n) \leq \inf_{T : \text{ orthogonal}} \left\{ L_D(T) + \tilde{O}\left( \frac{r^2(T) d^{3/2}}{n} \right) \right\}, \]
where $r(T)$ denotes the rank of $T$.\(^2\)

**Algorithm.** Our algorithm is based on the sum-of-squares (SoS) hierarchy of convex relaxations (Shor, 1987; Parrilo, 2000; Lasserre, 2001), applied to the tensor nuclear norm. For $k \in \mathbb{N}$, the sum-of-squares hierarchy defines a sequence of outer convex relaxations $\tilde{\|\cdot\|_{\text{nuc},k}}$ that give an increasingly tight approximation to the (hard to approximate (Hillar and Lim, 2013)) tensor nuclear norm $\|\cdot\|_{\text{nuc}}$ as $k$ is increased. The SoS nuclear norm was previously used in the work of Barak and Moitra (2016), who gave $O(1/\sqrt{n})$ rates for tensor completion in the agnostic model with absolute loss. Their main contribution was to show how certain spectral bounds arising in 3-SAT refutation (Coja-Oghlan et al., 2004) bound the Rademacher complexity for the unit ball of the degree-six SoS nuclear norm, thereby controlling the usual empirical process via uniform convergence. Our guarantees are based on empirical risk minimization over the (scaled) ball in this norm, specifically the degree-six relaxation $\|\cdot\|_{\text{nuc},6}$, and our analysis builds on their Rademacher complexity bounds.

\(^2\) Here the word "exact" refers to the fact that the leading constant in front of the loss on the right-hand-side is 1; such guarantees generally do not easily follow from the usual machinery used to analyze well-specified models.
Restricted eigenvalue and subgradient lemmas. While the $O(1/\sqrt{n})$-type rates provided in Barak and Moitra (2016) are optimal (in terms of $n$ dependence) for generic Lipschitz losses, it is not immediately obvious whether their result can be used to provide $O(1/n)$-type fast rates for strongly convex losses like the square loss. It has long been recognized that to obtain fast rates for prediction with strongly convex losses, more refined control of the empirical process is necessary. In particular, it is well-known that local Rademacher complexities and related fixed point complexities characterize the rates for empirical risk minimization in a data-dependent fashion. To bound such localized complexities under convex relaxations, the typical approach is to establish a restricted eigenvalue property for the empirical design matrix (Negahban et al., 2012; Bartlett et al., 2012; Lecué and Mendelson, 2017, 2018). Establishing such guarantees for regularizers such as the $\ell_1$ norm, nuclear norm, and so on is usually done by appealing to properties of the subgradient of the norm and proving that the norm (approximately) decomposes across certain subspaces (Negahban et al., 2012). The main tool we establish here, which allows us to unlock the full power of localized complexities and establish rates fast rates, is a new guarantee of this type for the SoS nuclear norm.

Informally, we show:

**Theorem 2 (informal)** Let $T^*$ be a fixed orthogonal tensor, and let $\mathcal{T} = \{T : \|T\|_{\text{nuc}} \leq \|T^*\|_{\text{nuc}}\}$. Then:

$$\|T - T^*\|_{\text{nuc}} \leq O(r(T^*)) \cdot \|T - T^*\|_F \quad \forall T \in \mathcal{T}.$$ 

This theorem is a consequence of a more general result we prove in Section 2, which gives a characterization of the subgradient of $\|\cdot\|_{\text{nuc}}$ at any orthogonal tensor. The basic idea is to show that a characterization for the subgradient of the tensor nuclear norm given in Yuan and Zhang (2016) can be captured in the sum-of-squares proof system. The final result (Theorem 5) is fairly user-friendly, and packages the complexity of SoS into a self-contained statement about geometric properties of the norm $\|\cdot\|_{\text{nuc}}$. We hope that this will find use in other applications. As one concrete example, in Section 3 we show how the subgradient characterization can also be used to give agnostic fast rates for the problem of low-rank tensor sensing with gaussian measurements. Here we also obtain $\tilde{O}(r^2d^{3/2}/n)$-type fast rates.

**Lower bounds.** Lastly, we prove that our results can’t be strengthened significantly while still obtaining computationally efficient algorithms. It is straightforward to show that there is an inefficient predictor that obtains $O(rd/n)$ excess risk, whereas the dimension scaling in Theorem 1 is $O(d^{3/2})$. This scaling is shared by the other results based on the SoS nuclear norm (Barak and Moitra, 2016; Potechin and Steurer, 2017), and Barak and Moitra (2016) show that finding relaxations for which the Rademacher complexity grows as $o(d^{3/2})$ is at least as hard as refuting random instances of 3-XOR. In Section 4 we give a computational lower bound for agnostic learning that shows that obtaining square loss excess risk scaling as $o(d^{3/2})$ is at least as hard as a certain distinguishing problem for random 3-XOR, sometimes called learning sparse parities with noise.

1.2. Related work

Early algorithms for computationally efficient tensor completion relied on unfolding: reshaping the tensor into a matrix and applying a matrix completion algorithm (Tomioka et al., 2011; Tomioka and Suzuki, 2013; Romera-Paredes and Pontil, 2013; Mu et al., 2014; Jain and Oh, 2014). This approach yields suboptimal results for third-order tensors and other odd-order tensors. For example, for a third-order tensor in $\mathbb{R}^d \otimes^3$, the most “balanced” unfolding of the tensor is a $d \times d^2$ matrix, and so
directly reducing to an algorithm for agnostic matrix completion (e.g. Koltchinskii et al. (2011)) would yield suboptimal $O(d^2)$-type sample complexity.

Recent results of Montanari and Sun (2016), Potechin and Steurer (2017), and Xia et al. (2017) all give sub-$O(d^2)$ type rates, but apply only to noiseless or well-specified models, and do not obviously extend to the agnostic setting. Potechin and Steurer (2017) give exact completion in the noiseless case after $O(rd^{3/2})$ entries are observed. Montanari and Sun (2016) show that a refined spectral approach based on unfolding can obtain sub-$O(d^2)$ rates for prediction error in the noiseless setting, but they obtain a rate of $O(1/n^{2/3})$ that falls short of the $O(1/n)$-type rate we provide. Finally, Xia et al. (2017) recently showed that an algorithm based on power iteration provides $O(d/n)$-type rates once $n = O(d^{3/2})$, but their result only applies to well-specified models, and it seems unlikely that this algorithmic approach succeeds in the fully agnostic setting.

Our results build on the seminal work of Barak and Moitra (2016) and Potechin and Steurer (2017), both of which use sum-of-squares to give $O(d^{3/2})$-type guarantees that improve on unfolding. The former was the first paper to study the agnostic setting, and gave slow excess risk guarantees. A natural candidate to develop efficient algorithms for tensor completion is the Kronecker product, which obeys the relation $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$. Given matrices $A_1, A_2, A_3$ and tensor $T = \sum_i u_i \otimes v_i \otimes w_i$, we define $(A_1 \otimes A_2 \otimes A_3)^T = \sum_i (A_1u_i) \otimes (A_2v_i) \otimes (A_3w_i)$. Whenever $T$ is an orthogonal tensor of the form (2), we let $r(T)$ denote the rank.

2. Subgradient of the sum-of-squares nuclear norm

2.1. Tensor nuclear norm and sum-of-squares relaxation

In the classical results on matrix completion (Candès and Recht, 2009; Candès and Tao, 2010; Recht, 2011; Koltchinskii et al., 2011), a central object is the nuclear norm, which arises as a convex relaxation of the rank. A natural candidate to develop efficient algorithms for tensor completion is...
the tensor nuclear norm (e.g. Hillar and Lim (2013); Friedland and Lim (2018)), which may be defined via
\[
\|T\|_{\text{nuc}} = \inf \left\{ \sum_{i=1}^r |\lambda_i| : W = \sum_{i=1}^r \lambda_i u_i \otimes v_i \otimes w_i, \quad \|u_i\|_2 = \|v_i\|_2 = \|w_i\|_2 = 1, \quad r \in \mathbb{N} \right\},
\]
and whose dual is the injective tensor norm \(\|X\|_{\text{inj}} = \sup_{\|x\|_2 = \|y\|_2 = \|z\|_2 = 1} \{X, x \otimes y \otimes z\}\). Unfortunately, the optimization problem here—maximizing a degree-three polynomial over the sphere—is intractable in general.

The approach we take, following Barak and Moitra (2016) and Potechin and Steurer (2017), is to employ the sum-of-squares hierarchy of convex relaxations (Shor, 1987; Parrilo, 2000; Lasserre, 2001) which provides an increasingly tight sequence of relaxations of the optimization problem in (4). To describe the relaxations, we require the notion of a pseudodistribution.

**Definition 3 (Pseudodistribution (Barak and Steurer (2016)))** Let \(\mu : \mathbb{R}^d \to \mathbb{R}\) be a finitely supported function and let \(\mu f = \sum_{x \in \text{supp}(\mu)} \mu(x) f(x)\). \(\mu\) is said to be a degree-\(k\) pseudodistribution if \(\sum_{i=1}^r \mu_i = 1\) and \(\mu f^2 \geq 0\) for all polynomials \(f\) of degree at most \(k/2\).

Given a degree-\(s\) pseudodistribution \(\mu\) and system of polynomial inequalities \(A = \{f_1 \geq 0, \ldots, f_m \geq 0\}\) \(\cup\) \(\{g_1 = 0, \ldots, g_{m'} = 0\}\), we write \(\mu = A\) if for all \(S \subseteq [m]\) and all sum-of-squares polynomials \(h\) such that \(\deg h + \sum_{i \in S} \deg f_i \leq s\),
\[
\begin{split}
\mu_h \left[ \prod_{i \in S} f_i \right] \geq 0,
\end{split}
\]
and \(\mu g_i = 0\) for all \(i \in [m']\) and all polynomials \(q\) such that \(\deg (g_i q) \leq \ell\). With the pseudodistribution formalism, we define the degree-\(k\) SoS injective norm as follows:
\[
\|X\|_{\mu} = \sup_{\mu \text{ degree-}k} \{X, \mu(x) \otimes y \otimes z\}.
\]
The degree-\(k\) SoS nuclear norm is simply defined as the dual: \(\|T\|_{\text{nuc}} := \sup_{\|X\|_{\mu} \leq 1} \{X, T\}\). It can equivalently be expressed by defining
\[
\mathcal{K}_k = \left\{ T \in \mathbb{R}^{d \times d \times d} \mid \exists \mu \text{ degree-}k : \mu(x) \otimes y \otimes z = T, \quad \mu \|x\|_2 = \mu \|y\|_2 = \mu \|z\|_2 = 1 \right\},
\]
and then
\[
\|T\|_{\text{nuc}} = \inf\{\alpha : T \in \mathcal{K}_k / \alpha\}.
\]
The SoS nuclear norm and injective norm can be evaluated in \(O^{O(k)}\) time (Grötschel et al., 1981; Barak and Steurer, 2016). Moreover, the norms obey the ordering \(\|T\|_{\text{nuc}} \geq \ldots \geq \|T\|_{\text{inj}} \geq \ldots \geq \|T\|_{\text{nuc}} \geq \|T\|_{\text{inj}} \), and likewise \(\|X\|_{\text{inj}} \leq \ldots \leq \|X\|_{\text{nuc}} \leq \ldots \leq \|X\|_{\text{inj}} \leq \|X\|_{\text{nuc}}\).

---

3. There are minor technical differences (e.g., scaling) between the SoS injective/nuclear norm definitions we use and those of Barak and Moitra (2016); Potechin and Steurer (2017).

4. This equivalence is proven in Appendix A for completeness.
2.2. Subgradient and norm compatibility

Our algorithms are based on empirical risk minimization with SoS nuclear norm constraints, i.e., algorithms that minimize the empirical loss over the set \( \mathcal{T} = \{ T \in (\mathbb{R}^d)^{\otimes 3} | \| T \|_{\text{nuc},k} \leq \tau \} \) for appropriate choice of \( \tau \) and \( k \). The technical challenge to analyzing this type of relaxation is that even if measurements are realized by a rank-\( r \) tensor, there is nothing that guarantees a-priori that the tensor \( \hat{T}_n \), output by the algorithm is itself low rank. Letting \( T \) be a rank-\( r \) orthogonal tensor, our main technical result here shows that if \( \tau = \| T \|_{\text{nuc},k} \), then for all elements of \( T' \in \mathcal{T} \), the error \( \Delta = T' - T \) is “approximately low rank” in a sense that suffices to guarantee good generalization performance.

\[ \| \Delta \|_{\text{nuc},k} \leq 68r \cdot \| \Delta \|_F. \]  

(9)

In the analysis of nuclear norm regularization for matrix completion—and more broadly, throughout high-dimensional statistics—the key tool used to establish guarantees along the lines of (9) is a characterization for the subgradient for the nuclear norm, and the related notion of decomposability (Negahban et al., 2012; Negahban and Wainwright, 2012). It is known classically (Watson, 1992) that for any matrix \( W \) with singular value decomposition \( W = UV^\top \),

\[ \partial \| W \|_{\text{nuc}} = \{ UTV + X | U^\top X = XV = 0, \| X \|_{\text{op}} \leq 1 \}. \]  

(10)

Our approach in the remainder of this section is to establish a similar result for the subgradient of the SoS nuclear norm \( \| \cdot \|_{\text{nuc},k} \) at any orthogonal tensor \( T \). From here Theorem 4 will quickly follow. As a first step, we need to define certain subspaces and projection operators associated with \( T \).

**Subspaces** For the remainder of the section we let \( T \) be a rank-\( r \) orthogonal tensor as in (2). Define \( U = \text{span}\{ u_i \} \), \( V = \text{span}\{ v_i \} \), and \( W = \text{span}\{ w_i \} \), and note that each subspace has dimension at most \( r \). Let \( P_U : \mathbb{R}^d \to \mathbb{R}^d \) and \( P_V : \mathbb{R}^d \to \mathbb{R}^d \) and \( P_W : \mathbb{R}^d \to \mathbb{R}^d \) be orthogonal projections onto these subspaces and \( P_{U,1}, P_{V,1} \) and \( P_{W,1} \) be the projections onto the respective orthogonal complements. We define projection operators from \((\mathbb{R}^d)^{\otimes 3}\) to \((\mathbb{R}^d)^{\otimes 3}\) for all \( 2^3 \) combinations of subspaces:

\[
\begin{align*}
Q^0_{T,1} &= P_{U} \otimes P_{V} \otimes P_{W}, & Q^0_{T,1} &= P_{U,1} \otimes P_{V,1} \otimes P_{W,1}, \\
Q^1_{T,1} &= P_{U,1} \otimes P_{V} \otimes P_{W}, & Q^1_{T,1} &= P_{U} \otimes P_{V,1} \otimes P_{W,1}, \\
Q^2_{T,1} &= P_{U} \otimes P_{V,1} \otimes P_{W}, & Q^2_{T,1} &= P_{U,1} \otimes P_{V} \otimes P_{W,1}, \\
Q^3_{T,1} &= P_{U,1} \otimes P_{V,1} \otimes P_{W}, & Q^3_{T,1} &= P_{U} \otimes P_{V,1} \otimes P_{W}. \\
\end{align*}
\]

(11)

Lastly, we define two subspaces that play a central role in our analysis:

\[
\begin{align*}
Q_{T,1} &= Q^0_{T,1} + Q^1_{T,1} + Q^2_{T,1} + Q^3_{T,1}, & Q_{T,1} &= Q^0_{T,1} + Q^1_{T,1} + Q^2_{T,1} + Q^3_{T,1}. \\
\end{align*}
\]

(12)

One can verify via multilinearity that \( X = Q_{T,1}(X) + Q_{T,1}(X) \) for all \( X \in (\mathbb{R}^d)^{\otimes 3} \), and that any tensor in the range of \( Q_{T,1} \) spans at most \( r \) dimensions along at least two modes. We can now state our main theorem for the subgradient.

---

5. Following the convention in Section 1.3, if \( X = \sum_i a_i \otimes b_i \otimes c_i \), then \( (P_U \otimes P_V \otimes P_W)X = \sum_i (P_U a_i) \otimes (P_V b_i) \otimes (P_W c_i) \). It is also useful to note that for any \( x, y, z \) we have \( ((P_U \otimes P_V \otimes P_W) x \otimes y \otimes z) = (X, (P_U x) \otimes (P_V y) \otimes (P_W z)) \).
Then for all \( n \)

We now state our main learning results, which use the SoS nuclear norm to give efficient algorithms/parallel:alt1 which implies that

In other words, Theorem 5 states that

which we may view as a generalization of the matrix subgradient characterization (10). Yuan and Zhang (2016) proved a similar result for the (exact) tensor nuclear norm. Our proof of Theorem 5 shows that the essence of their proof can be captured by a low-degree sum-of-squares proof. It builds on the approach introduced in Potechin and Steurer (2017) to provide dual certificates for exact tensor completion.

With the subgradient lemma in hand, the path to the “approximately low rank” result of Theorem 4 is clear. Suppose that \( \|T'\|_{\text{nuc}k+2} \leq \|T\|_{\text{nuc}k+2} \), and let \( \Delta = T' - T \). By appropriately choosing the dual tensor \( X \) in (13), we can show that

which implies that \( \|\Delta\|_{\text{nuc}k} \leq 64 \|Q_{T'}(\Delta)\|_{\text{nuc}k} \leq r\|\Delta\|_\tau\), which is a consequence of the earlier remark that all the projections used to define \( Q_{T'} \) in (12) project into \( r \) dimensions along at least two modes. This full argument is in Appendix B.

3. Agnostic tensor completion

We now state our main learning results, which use the SoS nuclear norm to give efficient algorithms with fast rates for agnostic tensor completion and tensor sensing. For both results we receive observations \( (X_1, Y_1), \ldots, (X_n, Y_n) \) i.i.d. according to an unknown distribution \( D \), where \( (X_t, Y_t) \in (\mathbb{R}^d)^{\otimes 3} \times \mathbb{R} \), and the goal is to obtain low square loss excess risk in the sense of equation (1). We let \( \bar{D}_n \) denote the empirical expectation, which is uniform over the examples \( \{(X_t, Y_t)\}_{t=1}^n \).

3.1. Tensor completion

In the tensor completion model we take observations \( X_t \) to be of the form \( X_t = e_{i_t} \otimes e_{j_t} \otimes e_{k_t} \), where \( (i_t, j_t, k_t) \in [d]^3 \) is selected uniformly at random.\(^6\) In the noiseless or well-specified setting, this corresponds to observing a single entry of an unknown tensor, but we make no assumption on the responses \( Y_t \) other than boundedness. The main theorem is as follows.

**Theorem 6 (Formal version of Theorem 1)** Let \( \tau > 0 \) be fixed. Suppose that \( |Y| \leq R \) almost surely, and let \( \hat{T}_n \) be the empirical risk minimizer over the class \( \mathcal{T} = \{ T \in (\mathbb{R}^d)^{\otimes 3} \mid \|T\|_{\text{nuc}k} \leq \tau, \|T\|_\infty \leq R \} \). Then for all \( n \leq d^3 \), with probability at least \( 1 - \delta \),

\[
L_D(\hat{T}_n) - L_D(T^*) \leq O \left( \frac{R^2 r^2 (T^*)^{3/2} \log^6 d}{n} + \frac{R^2 \log (1/\delta)}{n} \right)
\]

for all orthogonal tensors \( T^* \in (\mathbb{R}^d)^{\otimes 3} \) with \( \|T^*\|_{\text{nuc}k} = \tau \) and \( \|T^*\|_\infty \leq R \).

---

\(^6\) Note that we sample entries with replacement, whereas related works use without-replacement sampling (Barak and Moitra, 2016; Potechin and Steurer, 2017).
Let us spend a moment interpreting the theorem. First, let $T^* = \arg\min_{T, \text{rank-}r} L_D(T)$. Then, by setting $\tau = \|T^*\|_{\text{nuc}}$, we are guaranteed that with probability at least $1 - \delta$,

$$L_D(\hat{T}_n) - \inf_{T, \text{rank-}r} L_D(T) \leq O\left(\frac{R^2 r^2 d^{3/2} \log^6 d}{n} + \frac{R^2 \log(1/\delta)}{n}\right).$$

(16)

More generally, (15) implies an exact oracle inequality (Koltchinskii et al., 2011; Gaiffas and Lecué, 2011): With probability at least $1 - \delta$, we have

$$L_D(\hat{T}_n) \leq \inf_{T\|T\|_{\text{nuc}} = \tau} \left\{ L_D(T) + O\left(\frac{R^2 r^2(T) d^{3/2} \log^6 d}{n} + \frac{R^2 \log(1/\delta)}{n}\right)\right\}.$$

Let us compare the result in detail with Barak and Moitra (2016), which is the only other polynomial time agnostic tensor completion result with sub-$O(d^2)$ sample complexity. For general noise distributions, their analysis gives an excess risk bound that scales as $\tilde{O}\left(\sqrt{r^2 d^{3/2}}\right)$. The bound in (15) matches this dependence on all the parameters, but is squared, and is thus always tighter. The result has excess risk against arbitrary tensors, however, while our bound requires orthogonality of the benchmark. We do not know whether this restriction can be removed.

Interestingly, while Barak and Moitra (2016) give excess risk bounds against incoherent tensors, we do not require incoherence. This is because we control the complexity of the benchmark through the $\ell_\infty$ norm of the entries rather than through the Frobenius norm; this parallels the situation in the matrix setting (Koltchinskii et al., 2011; Gaiffas and Lecué, 2011). Applying the spectral bounds of Barak and Moitra (2016) without incoherence requires slightly tightening the analysis.

Lastly, we remark that the guarantee (16) requires setting the parameter $\tau$ based on the norm of the unknown benchmark $T^*$. It is likely that this can be relaxed by appealing to penalized empirical risk minimization rather than empirical risk minimization as in Koltchinskii et al. (2011), but we leave this for future work.

3.2. Tensor sensing

In this setting we give agnostic learning guarantees for a setting we call tensor sensing, which generalizes the matrix compressed sensing setup studied in Negahban and Wainwright (2011). We assume that observations $X \in \mathbb{R}^{d \times 3}$ have independent entries from $\mathcal{N}(0; 1)$ and—as in the tensor completion setting—allow $Y \in \mathbb{R}$ to be arbitrary. For each tensor $T$, define $R(T)$ to be the smallest almost-sure bound on $|\langle T, X \rangle - Y|$. As in the tensor completion setup, the main result is a fast rate with $\tilde{O}\left(\frac{r^2 d^{3/2}}{n}\right)$-type scaling.

**Theorem 7** Let $\tau > 0$ be fixed, and let $\hat{T}_n$ be the empirical risk minimizer over the tensor class $T = \{ T \in (\mathbb{R}^d)^{\otimes 3} \mid \|T\|_{\text{nuc}} \leq \tau \}$. Then with probability at least $1 - \delta$,

$$L_D(\hat{T}_n) - L_D(T^*) \leq O\left(\frac{R^2(T^*) r^2(T^*) d^{3/2} \log^3(d/\delta)}{n}\right).$$

(17)

for all orthogonal tensors $T^* \in (\mathbb{R}^d)^{\otimes 3}$ for which $n = \Omega(r^2(T^*) d^{3/2} \log^{1/2} d + \log(1/\delta))$ and $\|T^*\|_{\text{nuc}} = \tau$. 


Overview of analysis We now sketch how the subgradient theorem can be combined with empirical process arguments to prove Theorem 6 and Theorem 7. We follow a generic recipe given in Appendix C.1—specifically, Theorem 20—which shows that to control the generalization error of empirical risk minimization, it suffices to bound a certain “offset” or “shifted” empirical process. For any fixed benchmark $T^*$, the excess risk relative to $T^*$ is bounded as

$$L_D(T_n) - L_D(T^*) \leq \sup_{\Delta \in \mathcal{T} - T^*} \left\{ \text{E} \left[ \frac{1}{n} \left( \text{E}\left[ X \right] - \text{E}\left[ X \right] \right) \right] + \text{E} \left[ \left( \text{E}\left[ X \right] - \text{E}\left[ X \right] \right)^2 - 2\text{E}\left[ X \right] \right] \right\}.$$

(18)

The offset process on the right-hand side was used to obtain high-probability fast rates for misspecified models by Liang et al. (2015), and its analysis is closely related to that of Lecué and Mendelson (2013); Mendelson (2014). To bound the process, it suffices to establish a type of lower isometry/restricted eigenvalue property, which we state here for the case of tensor regression: let $X_n : (\mathbb{R}^d)^{\otimes 3} \to \mathbb{R}^n$ be the data operator, which maps any tensor $T$ to the sequence $(T, X_1), \ldots, (T, X_n)$, and let $\Sigma = \text{E}_X [XX^T] \in \mathbb{R}^{d^3 \times d^3}$ be the covariance matrix for the vectorized measurements. Then it suffices to show that with high probability, the following restricted eigenvalue bound holds:

$$\frac{1}{\sqrt{n}} \|X_n(\Delta)\|_2 \geq c \|\Sigma^{1/2} \Delta\|_F \quad \forall \Delta \in \mathcal{T} - T^*,$$

where $c > 1/\sqrt{2}$ is a sufficiently large constant.

Our starting point to establish the guarantee is to borrow a bound from Hopkins et al. (2015), which states that $\text{E} \|X\|_{\text{inj}} = O(d^{3/4} \log^{1/4} d)$ under gaussian measurements, and suffices to bound the Rademacher complexity of our tensor class. Using this bound in conjunction with standard gaussian concentration arguments and the “peeling” method (e.g. (Negahban and Wainwright, 2011)), we prove Theorem 25, which states that with high probability,

$$\frac{1}{\sqrt{n}} \|X_n(\Delta)\|_2 \geq 0.79 \cdot \|\Delta\|_F - \frac{C d^{3/4} \log^{1/4} d}{\sqrt{n}} \cdot \|\Delta\|_{\text{inj}} \quad \forall \Delta \in (\mathbb{R}^d)^{\otimes 3}.$$

Combined with Theorem 4, which asserts that all $\Delta \in \mathcal{T} - T^*$ have $\|\Delta\|_{\text{inj}} \leq O(r(T^*)) \cdot \|\Delta\|_2$, we have the following consequence: once $n = \Omega(r(T^*) d^{3/2} \log^{1/2} d)$, with high probability,

$$\frac{1}{\sqrt{n}} \|X_n(\Delta)\|_2 \geq (0.79 - o(1)) \cdot \|\Delta\|_F \quad \forall \Delta \in \mathcal{T} - T^*.$$

To establish the analogous bound in the tensor completion model we use the SoS Rademacher bound from Barak and Moitra (2016), but utilize somewhat different concentration arguments. Indeed, due to the sparse nature of the measurement distribution one cannot hope to exactly establish the restricted eigenvalue property for $X_n$, and must instead show that it holds up to a small additive error.

4. Computational lower bounds

In the rank-one case, the excess risk bound of Theorem 6 scales as $\tilde{O}(d^{3/2} / n)$, while the excess risk attained by the natural inefficient algorithm scales as $\tilde{O}(d/n)$. It is natural to ask whether this $O(d^{1/2})$ gap can be improved or whether it poses a fundamental barrier. In the slow rate regime, Barak and Moitra (2016) gave a computational lower bound showing that finding efficiently computable classes of tensors for which the Rademacher complexity grows as $o(\sqrt{d^{3/2} / n})$ is at least
as hard as refuting random instances of 3-XOR with \( o(d^{3/2}) \) clauses. In this section we show that this computational hardness is also present in the fast rate regime: Under plausible average-case hardness assumptions, no polynomial time algorithm can obtain a fast rate for square loss scaling as \( O(d^{3/2-\varepsilon}/n) \) for any \( \varepsilon > 0 \).

Our improper learning lower bound applies to any algorithm that obtains low excess risk in the sense of (15), and states that under conjectured hardness of a certain distinguishing problem for 3-XOR it is not possible to improve the \( O(d^{3/2}) \) dependence on dimension.

We reduce from the 3-XOR problem over variables \( x \in \{\pm 1\}^d \). A 3-XOR instance consists of a sequence of \( m \) clauses of the form

\[
x_i \cdot x_j \cdot x_k = z_{ijk},
\]

where \( z_{ijk} \in \{\pm 1\} \) is a target. We consider two families of instances:

- **Planted.** Fix an arbitrary assignment \( a \in \{\pm 1\}^d \). Select \( m \) triples \((i,j,k)\) uniformly at random with replacement.\(^7\) For each such triple \((i,j,k)\), include a clause

\[
x_i \cdot x_j \cdot x_k = z_{ijk} := \begin{cases} a_i \cdot a_j \cdot a_k, & \text{with probability } 1 - \eta \\ -a_i \cdot a_j \cdot a_k, & \text{with probability } \eta. \end{cases}
\]

Note, we sample the value \( z_{i,j,k} \) for a triple \((i,j,k)\) only once: if the triple is sampled multiple times, the value \( z_{i,j,k} \) will be the same.

- **Random.** Select \( m \) triples \((i,j,k)\) uniformly at random with replacement, and take each clause to be \( x_i \cdot x_j \cdot x_k = z_{ijk} \), where \( z_{ijk} \) is drawn from \( \{\pm 1\} \) uniformly at random. Again, we sample the value \( z_{i,j,k} \) for a triple \((i,j,k)\) only once.

An algorithm for the distinguishing problem takes \( m \) clauses as input and outputs either “Planted” or “Random”. The algorithm is said to succeed if it outputs “Planted” for planted instances and “Random” for random instances with probability at least \( 1 - o(1) \) over the draw of the instance. Note that the problem becomes easier as \( \eta \) gets smaller, and in particular when \( \eta = 0 \) the problem can be solved in polynomial time using Gaussian elimination.

**Conjecture 8** \( \exists \) some constant \( \eta < 1/4 \) such that no algorithm that succeeds for the 3-XOR distinguishing problem with \( m = o(d^{3/2}) \) runs in polynomial time.

All known polynomial time algorithms for distinguishing require \( m = \Omega(d^{3/2}) \) clauses, and conjectured hardness of the closely related problem of strong refutation for random 3-XOR for with \( o(d^{3/2}) \) clauses has been used as a basis to establish hardness of other learning problems (Daniely, 2016; Raghavendra et al., 2017; Kothari et al., 2017; Feldman et al., 2018).

**Theorem 9** Let \( \varepsilon > 0 \) be fixed. Assuming the 3-XOR distinguishing conjecture, there is no polynomial time algorithm for agnostic tensor completion that guarantees that for any distribution \( D \), with probability at least \( 1 - o(1) \),

\[
L_D(\hat{T}_n) - \inf_{T^*: \text{rank-1}} L_D(T^*) = O\left(d^{3/2-\varepsilon}/n\right). \tag{19}
\]
5. Conclusion

Our results demonstrate the power of the sum-of-squares hierarchy for agnostic statistical learning, and show for the first time that sum-of-squares algorithms can obtain fast rates for prediction. We hope our work will serve as a starting point for applying sum-of-squares to obtain polynomial time algorithms with fast rates in statistical learning for broader classes of models.

A few immediate technical questions emerge. Can the dependence on rank in our results be improved? Can the subgradient results be extended to the general undercomplete or even overcomplete case? Can similar agnostic learning results be obtained with a more practical algorithm that does not rely on solving the full sum-of-squares SDP?

Acknowledgements  We thank Sasha Rakhlin and Ankur Moitra for helpful discussions and thank Matthew J. Telgarsky for being a constant source of inspiration.

References


Appendix A. Preliminaries

A.1. Sum-of-squares proof system

Let \( \mathcal{A} = \{ f_1 \geq 0, \ldots, f_m \geq 0 \} \cup \{ g_1 = 0, \ldots, g_{m'} = 0 \} \) be a system of polynomial constraints. A degree-\( \ell \) sum-of-squares proof that \( \mathcal{A} \) implies a constraint \( \{ h \geq 0 \} \) is a set of polynomials \( (q_i)_{i \in [m']} \) and sum-of-squares polynomials \( (p_S)_{S \subseteq [m]} \) such that

\[
  h = \sum_{S \subseteq [m]} p_S \prod_{i \in S} f_i + \sum_{i \in [m']} q_i g_i,
\]

and where and \( \deg(p_S \prod_{i \in S} f_i) \leq \ell \) for all \( S \) and \( \deg(q_i g_i) \leq \ell \) for all \( i \).\(^8\) We write \( \mathcal{A} \vdash \{ h \geq 0 \} \) whenever such a proof exists. Our proofs going forward utilize the well-known duality of SoS proofs and pseudodistributions. See O’Donnell and Zhou (2013) and Barak and Steurer (2016) for further discussion, as well as inference rules for the SoS proof system.

We note the following well-known, but useful lemma:

\(^8\) We use the convention \( \prod_{i \notin \emptyset} f_i = 1 \), so that if \( \vdash \ell \) then \( h \) itself is a degree-\( \ell \) sum of squares.
Lemma 10 (Pseudo Cauchy-Schwarz) Let \( f \) and \( g \) be polynomials and let \( \ell = 2 \max \{ \deg f, \deg g \} \). Then for any \( \eta > 0 \),
\[
\sup \{ f g \leq \eta f^2 + \frac{1}{2\eta} g^2 \}. \tag{20}
\]
As a consequence, if \( \mu \) is a degree-\( s \) pseudodistribution with \( s \geq \ell \), then
\[
\overline{E}_\mu[f g] \leq \sqrt{\overline{E}_\mu[f^2] \cdot \overline{E}_\mu[g^2]} . \tag{21}
\]

A.2. Basic technical results

A.2.1. Sum-of-squares norms

We state a few lemmas capturing useful properties of the sum-of-squares norms.

Proposition 11 The SoS nuclear norm and SoS injective norm are dual: \( \| \cdot \|_{\text{nuc}}^* = \| \cdot \|_{\text{inj}} \) and \( \| \cdot \|_{\text{inj}}^* = \| \cdot \|_{\text{nuc}}^* \).

Proof of Proposition 11 It is immediate from the norm definitions that
\[
\| X \|_{\text{nuc}}^* = \sup_{W \in K_r} \langle W, X \rangle = \sup_{\mu \text{ degree-} r} \sup_{\| x \|^2 \leq E_\mu, \| y \|^2 \leq E_\mu, \| z \|^2 = 1} \langle W, X \rangle = \| X \|_{\text{inj}}^* .
\]
The other direction is a consequence of the standard duality theory for finite-dimensional Banach spaces. See, e.g., Theorem 15.4 in Rockafellar (1970).

Lemma 12 Let \( T = \sum_{i=1}^r \lambda_i \cdot u_i \otimes v_i \otimes w_i \) be an orthogonal rank-\( r \) tensor. Then for all \( k \geq 4 \),
\[
\| T \|_{\text{nuc}} = \| T \|_k = \sum_{i=1}^r |\lambda_i|, \quad \text{and} \quad \| T \|_{\text{inj}} = \| T \|_{\text{inj}} = \max_{i \leq r} |\lambda_i| . \tag{22}
\]

Lemma 12 states that the SoS relaxations of the nuclear norm and injective norm are essentially “integral” for orthogonal tensors. Note, this should not be a huge surprise since it is well-known that polynomial time methods such as power iteration succeed at decomposing orthogonal tensors (Kolda and Bader, 2009). We should mention it doesn’t seem possible to directly apply such results to give agnostic learning guarantees along the lines of our main theorem. While our benchmark is an orthogonal tensor, the data itself may have no orthogonal structure, and thus there is no clear object to which one might apply such a decomposition.

Proof of Lemma 12 We may assume \( \lambda_i \geq 0 \) without loss of generality. We first prove equality for the injective norms. Let \( \iota^* = \arg \max_{i \leq r} \lambda_i \). As a starting point, for any \( k \) we have \( \| T \|_{\text{inj}} \geq \| T \|_{\text{inj}} \geq \| T \|_{\text{inj}} \geq \| T \|_{\text{inj}} \geq \lambda_i \) by exhibiting \( u_i^* \otimes v_i^* \otimes w_i^* \) as a feasible solution to the supremum in \( \| T \|_{\text{inj}} = \sup_{\| x \|^2 \leq E_\mu, \| y \| = 1} \langle T, x \otimes y \otimes z \rangle . \)

For the upper bound, let \( x, y, z \) be indeterminates and—exploiting orthogonality—let us change coordinates such that \( u_i = v_i = w_i = e_i \). Then we have
\[
\langle T, x \otimes y \otimes z \rangle = \sum_{i=1}^r \lambda_i x_i y_i z_i
\]
From equation (20), we have
\[ t_A x_i y_i z_i \leq \frac{1}{2} x_i^2 + \frac{1}{2} y_i^2 z_i^2. \]
We also have \( t_A \sum_{i=1}^{r} y_i^2 z_i^2 \leq (\sum_{i=1}^{r} y_i^2)(\sum_{i=1}^{r} z_i^2). \) By the additivity of SoS proofs, and since we have assumed \( \lambda_i \geq 0, \) this implies
\[ t_A (T, x \otimes y \otimes z) \leq \max_{i \leq r} \lambda_i \cdot \left( \frac{1}{2} x_2^2 + \frac{1}{2} y_2^2 z_2^2 \right). \]

Now let \( A = \{ \|x\|_2^2 = 1, \|y\|_2^2 = 1, \|z\|_2^2 = 1 \}. \) We claim \( \|y\|_2^2 z_2^2 = 1. \) To see this, write
\[ 1 - \|y\|_2^2 z_2^2 = (1 - \|y\|_2^2)(1 + \|z\|_2^2) + (1 - \|z\|_2^2) + \|y\|_2^2 - 1 \text{ and use that } \deg((1 - \|y\|_2^2)(1 + \|z\|_2^2)) = 4. \]

Putting everything together, we see that \( \|T, x \otimes y \otimes z\| \leq \max_{i \leq r} \lambda_i. \) Thus, since the \( \ell_2 \) norm is preserved under change of basis, it follows that if \( \mu \) is any feasible degree-4 pseudodistribution for the maximization problem (6), we must have \( \overline{E}_\mu (T, x \otimes y \otimes z) \leq \max_i \lambda_i, \) and so \( \|T\|_{\tilde{m}_k} \leq \max_i \lambda_i. \)

We now establish equality for the nuclear norms. We trivially have \( \|T\|_{\tilde{m}_k} \leq \|T\|_{\text{nuc}} \leq \sum_{i=1}^{r} \lambda_i \) by exhibiting the decomposition \( T = \sum_{i=1}^{r} \lambda_i \cdot u_i \otimes v_i \otimes w_i \) as a feasible solution to the minimization problem in (4). For the other direction, define \( X^* = \sum_{i=1}^{r} u_i \otimes v_i \otimes w_i, \) and observe that the equality we just established for the injective norm implies \( \|X^*\|_{\tilde{m}_k} \leq 1. \) Thus, using the duality of the SoS nuclear norm and injective norm from Proposition 11, we have
\[ \|T\|_{\tilde{m}_k} = \sup_{X \in (\mathbb{R}^d)^{\otimes 3}: \|X\|_{\tilde{m}_k} \leq 1} \langle X, T \rangle \geq \langle X^*, T \rangle = \sum_{i=1}^{r} \lambda_i, \]
where the last equality uses that \( \{u_i\}, \{v_i\}, \) and \( \{w_i\} \) are all orthogonal. \( \blacksquare \)

**Proposition 13** Let \( k \geq 4. \) For any degree-\( k \) pseudodistribution \( \mu, \)
\[ \langle T, \overline{E}_\mu (x \otimes y \otimes z) \rangle \leq \|T\|_{\tilde{m}_k} \sqrt{\overline{E}_\mu \|x\|_2^2 \cdot \overline{E}_\mu \|y\|_2^2 \cdot \overline{E}_\mu \|z\|_2^2}. \] (23)
Furthermore, the following statements hold:
\[ \|z\|_2^2 \leq 1 \implies t_k \langle T, x \otimes y \otimes z \rangle \leq \|T\|_{\tilde{m}_k} \cdot \left( \frac{1}{2} x_2^2 + \frac{1}{2} y_2^2 z_2^2 \right), \]
\[ \|y\|_2^2 \leq 1 \implies t_k \langle T, x \otimes y \otimes z \rangle \leq \|T\|_{\tilde{m}_k} \cdot \left( \frac{1}{2} x_2^2 + \frac{1}{2} z_2^2 \right), \]
\[ \|x\|_2^2 \leq 1 \implies t_k \langle T, x \otimes y \otimes z \rangle \leq \|T\|_{\tilde{m}_k} \cdot \left( \frac{1}{2} y_2^2 + \frac{1}{2} z_2^2 \right). \] (24)

**Proof of Proposition 13** Equation (23) follows by rescaling a given pseudodistribution \( \mu \) with \( x' = x/\sqrt{\overline{E}_\mu \|x\|_2^2} \) and so forth so that the pseudodistribution is feasible for the maximization problem (6).

For (24), let \( \mu \) be a degree-\( k \) pseudodistribution with \( \mu = \{ \|z\|_2^2 \leq 1 \}. \) Then, using (23) and the AM-GM inequality we get \( \langle T, \overline{E}_\mu (x \otimes y \otimes z) \rangle \leq \frac{\|T\|_{\tilde{m}_k}}{\left( \overline{E}_\mu \|x\|_2^2 + \overline{E}_\mu \|y\|_2^2 \right)}. \) Using linearity of the pseudoexpectation operator we have \( \overline{E}_\mu \left[ \frac{\|T\|_{\tilde{m}_k}}{\left( \overline{E}_\mu \|x\|_2^2 + \overline{E}_\mu \|y\|_2^2 \right)} - \langle T, x \otimes y \otimes z \rangle \right] \geq 0, \) and so (24) follows from the duality of pseudoexpectations and sum-of-squares proofs. The remaining statements follow by symmetry. \( \blacksquare \)
A.2.2. Projections

Here we state some basic results regarding the projection operators defined in Section 2.2.

**Proposition 14** Let \( x, y, z \in \mathbb{R}^d \) be given. If at least two of the follow conditions hold:

1) \( x \in \mathbb{U} \), 2) \( y \in \mathbb{V} \), 3) \( z \in \mathbb{W} \),

then \( Q_{\mathbb{T}^i}(x \otimes y \otimes z) = 0 \).

**Proof of Proposition 14** Suppose that \( x \in \mathbb{U} \) and \( y \in \mathbb{V} \). Then \( P_{\mathbb{U}^i}(x) = 0 \), and so \( Q_{\mathbb{T}^i}^0(x \otimes y \otimes z) = Q_{\mathbb{T}^i}^2(x \otimes y \otimes z) = 0 \). We also have \( P_{\mathbb{V}^i}(y) = 0 \), and so \( Q_{\mathbb{T}^i}^1(x \otimes y \otimes z) = 0 \). The remaining cases follow by symmetry.

**Lemma 15** Let \( k \geq 4 \). For any tensor \( X \in (\mathbb{R}^d)^{\otimes 3} \), and any subspaces \( \mathbb{U}, \mathbb{V}, \mathbb{W} \) we have

\[
\|(P_{\mathbb{U}} \otimes P_{\mathbb{V}} \otimes P_{\mathbb{W}})X\|_{\mathbb{U}_{\mathbb{V}_{\mathbb{W}}}} \leq \|X\|_{\mathbb{U}_{\mathbb{V}_{\mathbb{W}}}},
\]

and in particular \( \|Q_{\mathbb{T}^i}(X)\|_{\mathbb{U}_{\mathbb{V}_{\mathbb{W}}}} \leq 4\|X\|_{\mathbb{U}_{\mathbb{V}_{\mathbb{W}}}} \) and \( \|Q_{\mathbb{T}^i}(X)\|_{\mathbb{U}_{\mathbb{V}_{\mathbb{W}}}} \leq 4\|X\|_{\mathbb{U}_{\mathbb{V}_{\mathbb{W}}}} \).

**Proof of Lemma 15** For any degree-4 pseudodistribution \( \mu \) over indeterminates \( x, y, z \) we have

\[
\|\mathbb{P}_{\mu}((P_{\mathbb{U}} \otimes P_{\mathbb{V}} \otimes P_{\mathbb{W}})X, x \otimes y \otimes z)\| = \mathbb{P}_{\mu}(X, (P_{\mathbb{U}}x) \otimes (P_{\mathbb{V}}y) \otimes (P_{\mathbb{W}}z)) \\
\leq \sqrt{\mathbb{P}_{\mu}\|P_{\mathbb{U}}x\|_2^2 \cdot \mathbb{P}_{\mu}\|P_{\mathbb{V}}y\|_2^2 \cdot \mathbb{P}_{\mu}\|P_{\mathbb{W}}z\|_2^2} \\
\leq \sqrt{\mathbb{P}_{\mu}\|x\|_2^2 \cdot \mathbb{P}_{\mu}\|y\|_2^2 \cdot \mathbb{P}_{\mu}\|z\|_2^2},
\]

where the first inequality uses Proposition 13 and the second uses that \( \|x\|_2 \leq \|x\|_2 \) for any subspace \( X \). This establishes the first result. Now observe from (11) that \( \|Q_{\mathbb{T}^i}(T)\|_{\mathbb{U}_{\mathbb{V}_{\mathbb{W}}}} \leq \sum_{i=1}^4 \|Q_{\mathbb{T}^i}(T)\|_{\mathbb{U}_{\mathbb{V}_{\mathbb{W}}}} \) and \( \|Q_{\mathbb{T}^i}(T)\|_{\mathbb{U}_{\mathbb{V}_{\mathbb{W}}}} \leq \sum_{i=1}^4 \|Q_{\mathbb{T}^i}(T)\|_{\mathbb{U}_{\mathbb{V}_{\mathbb{W}}}} \). We thus obtain the second result by applying the first to each of the summands.

A.2.3. Flattening

The multilinear rank of a tensor \( T \in (\mathbb{R}^d)^{\otimes 3} \) is the triple \( (r_1, r_2, r_3) \), where

\[
r_1(T) = \dim \text{span}\{T_{i,j,k} \mid j, k \in [d]\},
\]

is the dimension of the space spanned by the mode-1 fibers and \( r_2(T) \) and \( r_3(T) \) are defined likewise for the second and third mode.

We define the \( i \)th flattening map \( b_i : (\mathbb{R}^d)^{\otimes 3} \rightarrow \mathbb{R}^{d \times d^2} \) via

\[
b_1(T)_{i,(j,k)} = b_2(T)_{j,(i,k)} = b_3(T)_{k,(i,j)} = T_{i,j,k}.
\]

A standard result is that \( \text{rank}(b_i(T)) = r_i(T) \) (Friedland and Lim, 2018). We also have the following comparison between the nuclear norm of the tensor and its flattenings.

**Lemma 16 (Friedland and Lim (2018), Theorem 9.4)** For any tensor \( T \in (\mathbb{R}^d)^{\otimes 3} \),

\[
\|T\|_{\text{nuc}} \leq \min\{\sqrt{\min\{r_2, r_3\}} \|b_1(T)\|_{\text{nuc}}, \sqrt{\min\{r_1, r_3\}} \|b_2(T)\|_{\text{nuc}}, \sqrt{\min\{r_1, r_2\}} \|b_3(T)\|_{\text{nuc}}\}.
\]
A.2.4. Concentration

**Lemma 17 (Talagrand-type concentration for supremum of empirical process)** Let $\mathcal{F}$ be a class of functions of the form $f : \mathcal{Z} \to \mathbb{R}$. Let $z_1, \ldots, z_n$ be sampled i.i.d. from a distribution $\mathcal{D}$ over $\mathcal{Z}$ that satisfies $\mathbb{E}[f(z)] = 0$ and has $|f(z)| \leq c$ almost surely. Let $\sigma^2 = \sup_{f \in \mathcal{F}} \mathbb{E}[f^2(z)]$. Then for any $\delta > 0$, with probability at least $1 - 2\delta$ over the i.i.d. draw of $z_1, \ldots, z_n$,

$$
\mathbb{E}_{z_1, n} \mathbb{E}_{\epsilon} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(z_i) \leq 4 \mathbb{E}_{z_1, n} \mathbb{E}_{\epsilon} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(z_i) + \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}} + \frac{2c \log(1/\delta)}{n}.
$$

**Proof of Lemma 17** Follows from Theorem A.1 of Bartlett et al. (2005) applied to the classes $\mathcal{F}$ and $-\mathcal{F}$ separately, along with the standard in-expectation symmetrization lemma for uniform convergence. ■

Appendix B. Proofs from Section 2

The main result in this section is to prove Theorem 5, then use this result to prove Theorem 4. Before proceeding to the main proofs we state an intermediate result.

**Lemma 18 (Potechin and Steurer (2017))** Let $x, y, z \in \mathbb{R}^d$ be indeterminates. Let $A = \{\|y\|_2^2 = 1\}$. Then for any $r \in [d]$,

$$
A \vdash 6 \sum_{i=1}^{r} x_i y_i z_i \leq \frac{1}{2} \|x\|_2^2 + \frac{1}{2} \|z\|_2^2 - \frac{1}{4} \sum_{i=r+1}^{d} x_i^2 + z_i^2 - \frac{1}{8} \sum_{i=1}^{d} \sum_{j \neq i} y_i^2 (x_j^2 + z_j^2).
$$

**Corollary 19** Let $A = \{\|x\|_2^2 = 1, \|z\|_2^2 = 1\}$. Then for any $r \in [d]$,

$$
A \vdash 6 \sum_{i=1}^{r} x_i y_i z_i \leq 1 - \frac{1}{4} \sum_{i=r+1}^{d} x_i^2 + z_i^2 - \frac{1}{8} \sum_{i=1}^{d} \sum_{j \neq i} y_i^2 (x_j^2 + y_j^2 + z_j^2).
$$

**Proof of Corollary 19** We will show that $A \vdash 6 - \sum_{i=1}^{d} \sum_{j \neq i} y_i^2 y_j^2 \|x\|_2^2 = - \sum_{i=1}^{d} \sum_{j \neq i} y_i^2 y_j^2$; the term involving $\|z\|_2^2$ follows from the same reasoning. The desired inequality is equivalent to $\sum_{i=1}^{d} \sum_{j \neq i} y_i^2 y_j^2 (\|x\|_2^2 - 1) = 0$, which is clearly the product of a degree-4 polynomial and the equality constraint $\{\|x\|_2^2 - 1 = 0\}$. ■

**Proof of Theorem 5** Preliminaries. We first claim that the following equalities hold:

- $Q_{T_1}^0(X^*) = X^*$.
- $\|X^*\|_{\tilde{m}_k} = \|X^*\|_{\tilde{m}_{k+2}} = \|X^*\|_{m_k} = 1$.
- $(X^*, T) = \|T\|_{\tilde{m}_{k}^{\star}} = \|T\|_{\tilde{m}_{k+2}^{\star}} = \|T\|_{\text{nuc}}$.

Indeed, it is immediate from the definition of $X^*$ that $Q_{T_1}^0(X^*) = X^*$, and it follows from Lemma 12 that $\|X^*\|_{\tilde{m}_k} = 1$ and $(X^*, T) = \|T\|_{\tilde{m}_{k}^{\star}}$ for all $k \geq 4$. 19
Bounding dual norm is sufficient. To establish the inequality (13), we reduce to a simpler problem. The claim is as follows: Fix a constant $\alpha > 0$. If for all $X \in \mathbb{R}^d$ with $\|X\|_{\tilde{m}_k} \leq \alpha$, we have

$$\|X^* + Q_{T^*}(X)\|_{\tilde{m}_{k+2}} \leq 1,$$  

(27)

then (13) holds for all $X$ with $\|X\|_{\tilde{m}_k} \leq \alpha$. To see that this is the case, observe that for any $T'$ and all such $X$ we have

$$\{X^* + Q_{T^*}(X), T' - T\} \leq \{X^* + Q_{T^*}(X)\|_{\tilde{m}_{k+2}} \|T'\|_{\tilde{nuc}_{k+2}}, \{X^* + Q_{T^*}(X), T\} \leq \|T'\|_{\tilde{nuc}_{k+2}} - \{X^* + Q_{T^*}(X), T\} = \|T'\|_{\tilde{nuc}_{k+2}} - \|T\|_{\tilde{nuc}_{k+2}},$$

where the first inequality uses Proposition 11, the second inequality uses (27), and the final equality uses the definition of $X^*$ and that $\{Q_{T^*}(X), T\} = \{X, Q_{T^*}(T)\} = 0$. Rearranging the inequality yields (13).

Bounding the dual norm. The remainder of the proof establishes that (27) holds for $\alpha = 1/64$.

Let $x, y, z \in \mathbb{R}^d$ be indeterminates. We will provide a degree-$(k + 2)$ SoS upper bound on the polynomial $\{X^* + Q_{T^*}(X), x \otimes y \otimes z\}$, which will suffice to establish (27).

Let $\{u_i\}_{i=1}^r, \{v_i\}_{i=1}^r, \{w_i\}_{i=1}^r$ be as in (2). Then, let $U = \text{span}(\{u_i\}_{i=1}^r), \mathcal{V} = \text{span}(\{v_i\}_{i=1}^r), \mathcal{W} = \text{span}(\{w_i\}_{i=1}^r)$, so that $\{u_i\}_{i=1}^r$ is a basis for $U$, and likewise for the other modes. Let $\{u_i\}_{i=r+1}^d$ be an arbitrary orthonormal basis for $U^\perp$ and likewise with $\{v_i\}_{i=r+1}^d$ for $\mathcal{V}^\perp$, and $\{w_i\}_{i=r+1}^d$ for $\mathcal{W}^\perp$.

We perform a change of basis and let $u_i = v_i = w_i = e_i$, where $e_i$ is the $i$th standard basis vector. Then with $x, y, z$ expressed in the new basis we can write

$$\{X^* + Q_{T^*}(X), x \otimes y \otimes z\} = \sum_{i=1}^r x_i y_i z_i + \{Q_{T^*}(X), x \otimes y \otimes z\}. \quad (28)$$

Let $A = \{\|x\|^2 = 1, \|y\|^2 = 1, \|z\|^2 = 1\}$. From Corollary 19, we have

$$A \mathcal{A}_{\mathcal{A}} \sum_{i=1}^r x_i y_i z_i \leq \frac{1}{2} \|x\|^2 + \frac{1}{2} \|z\|^2 - \frac{1}{4} \sum_{i=r+1}^d x_i^2 + z_i^2 - \frac{1}{8} \sum_{i=1}^d y_i^2 (x_i^2 + z_i^2 + y_i^2).$$

We now handle the second term in (28). We will establish that

$$A \mathcal{A}_{\mathcal{A}} \{Q_{T^*}(X), x \otimes y \otimes z\} \leq \mathcal{A} + \mathcal{O}(\alpha) \cdot \sum_{i=r+1}^d x_i^2 + z_i^2 + \mathcal{O}(\alpha) \cdot \sum_{i=1}^d y_i^2 (x_i^2 + z_i^2 + y_i^2). \quad (29)$$

under the assumption that $\|X\|_{\tilde{m}_k} \leq \alpha$. To do this it suffices show that for each $i$ individually,

$$A \mathcal{A}_{\mathcal{A}} y_i^2 \{Q_{T^*}(X), x \otimes y \otimes z\} \leq \mathcal{O}(\alpha) \cdot y_i^2 \mathcal{A} \sum_{j=1}^d (x_j^2 + z_j^2 + y_j^2),$$

with an extra additive factor of $\mathcal{O}(\alpha) \cdot y_i^2 (x_i^2 + z_i^2)$ when $i > r$. Let $1 \leq i \leq d$ be fixed and let $x' = x - x_i e_i, y' = y - y_i e_i$ and $z' = z - z_i e_i$. We write

$$\{Q_{T^*}(X), x \otimes y \otimes z\} = \{X, Q_{T^*}((x_i e_i + x') \otimes (y_i e_i + y') \otimes (z_i e_i + z'))\}. \quad (30)$$
Case: $i \leq r$. Observe that with our change of basis we have $e_i \in \mathbb{U}$ along the first mode, $e_i \in \mathbb{V}$ along the second mode, and $e_i \in \mathbb{W}$ along the third mode. In view of Proposition 14, this means we have

$$0 = \mathcal{Q}_T((x_ie_i) \otimes (y_ie_i) \otimes (z_ie_i))$$

$$= \mathcal{Q}_T((x_ie_i) \otimes (y_ie_i) \otimes z')$$

$$= \mathcal{Q}_T(x' \otimes (y_ie_i) \otimes (z_ie_i))$$

$$= \mathcal{Q}_T((x_ie_i) \otimes y' \otimes (z_ie_i)).$$

Consequently, using multilinearity we can write

$$(\mathcal{Q}_T(X), x \otimes y \otimes z) = (\mathcal{Q}_T(X), x' \otimes y' \otimes z') + (\mathcal{Q}_T(X), x' \otimes y' \otimes (z_ie_i))$$

$$+ (\mathcal{Q}_T(X), (x_ie_i) \otimes y' \otimes z') + (\mathcal{Q}_T(X), x' \otimes (y_ie_i) \otimes z').$$

(31)

In what follows we will repeatedly invoke that $\mathcal{A} \vdash_2 \|x_ie_i\|_2^2 \leq \|x\|_2^2 \leq 1$, $\mathcal{A} \vdash_2 ||x'||_2^2 \leq \|x\|_2^2 \leq 1$, and so forth. We will also use that if $\|X\|_{\mathcal{M}_k} \leq \alpha$ then—via Proposition 13 and Lemma 15—for any indeterminates $a, b, c$,

$$\{ \|x\|_2^2 \leq 1 \} \vdash_k (\mathcal{Q}_T(X), a \otimes b \otimes c) \leq 2\alpha(\|b\|_2^2 + \|c\|_2^2),$$

$$\{ \|b\|_2^2 \leq 1 \} \vdash_k (\mathcal{Q}_T(X), a \otimes b \otimes c) \leq 2\alpha(\|a\|_2^2 + \|c\|_2^2),$$

$$\{ \|c\|_2^2 \leq 1 \} \vdash_k (\mathcal{Q}_T(X), a \otimes b \otimes c) \leq 2\alpha(\|b\|_2^2 + \|b\|_2^2).$$

These inequalities allow us to bound the terms in (31) as follows:

$$\{ \|x\|_2^2 \leq 1 \} \vdash_k (\mathcal{Q}_T(X), x' \otimes y' \otimes z') \leq 2\alpha(\|x'||_2^2 + \|z'||_2^2).$$

$$\{ \|z\|_2^2 \leq 1 \} \vdash_k (\mathcal{Q}_T(X), x' \otimes y' \otimes (z_ie_i)) \leq 2\alpha(\|y'||_2^2 + \|z'||_2^2).$$

$$\{ \|x\|_2^2 \leq 1 \} \vdash_k (\mathcal{Q}_T(X), (x_ie_i) \otimes y' \otimes z') \leq 2\alpha(\|y'||_2^2 + \|z'||_2^2).$$

$$\{ \|y\|_2^2 \leq 1 \} \vdash_k (\mathcal{Q}_T(X), x' \otimes (y_ie_i) \otimes z') \leq 2\alpha(\|x'||_2^2 + \|z'||_2^2).$$

Adding these inequalities, we get

$$\mathcal{A} \vdash_k (\mathcal{Q}_T(X), x \otimes y \otimes z) \leq 6\alpha \sum_{j \neq i}(x_j^2 + y_j^2 + z_j^2),$$

and by the multiplication rule for SoS proofs,

$$\mathcal{A} \vdash_{k+2} y_j^2 (\mathcal{Q}_T(X), x \otimes y \otimes z) \leq 6\alpha \sum_{j \neq i}y_j^2(x_j^2 + y_j^2 + z_j^2).$$

Case: $i > r$. As in the previous case, we split the expression in (30) using multilinearity, however we can no longer argue that four of the eight terms vanish. As a starting point, for the four terms that appeared in the $i \leq r$ case we can adopt the same upper bound to get

$$\mathcal{A} \vdash_k (\mathcal{Q}_T(X), x \otimes y \otimes z) \leq \mathcal{Q}_T((x_ie_i) \otimes (y_ie_i) \otimes (z_ie_i)) + \mathcal{Q}_T((x_i e_i) \otimes y' \otimes (z_i e_i))$$

$$+ \mathcal{Q}_T((x_i e_i) \otimes (y_i e_i) \otimes z') + \mathcal{Q}_T(x' \otimes (y_i e_i) \otimes (z_i e_i))$$

$$+ 6\alpha \sum_{j \neq i}(x_j^2 + y_j^2 + z_j^2).$$
We bound the four remaining terms as follows:

\[
\begin{align*}
\{\|y_i\|^2 \leq 1\} \implies & k Q_{T^i}((x_i e_i) \otimes (y_i e_i) \otimes (z_i e_i)) \leq 2\alpha(x_i^2 + z_i^2), \\
\{\|y_i\|^2 \leq 1\} \implies & k Q_{T^i}((x_i e_i) \otimes (y_i e_i) \otimes z') \leq 2\alpha(x_i^2 + z'^2), \\
\{\|y_i\|^2 \leq 1\} \implies & k Q_{T^i}(x' \otimes (y_i e_i) \otimes (z_i e_i)) \leq 2\alpha(\|x'\|^2 + z_i^2), \\
\{\|y_i\|^2 \leq 1\} \implies & k Q_{T^i}((x_i e_i) \otimes y' \otimes (z_i e_i)) \leq 2\alpha(x_i^2 + z_i^2).
\end{align*}
\]

Adding together all of these inequalities, we get

\[
A \overset{k}{\implies} Q_{T^i} (X), x \otimes y \otimes z \leq 6\alpha(x_i^2 + z_i^2) + 8\alpha \sum_{j \neq i} (x_j^2 + y_j^2 + z_j^2),
\]

and

\[
A \overset{k+2}{\implies} \|y\|^2 \cdot Q_{T^i}(X), x \otimes y \otimes z \leq 6\alpha y_i^2(x_i^2 + z_i^2) + 8\alpha \sum_{j \neq i} y_j^2(x_j^2 + y_j^2 + z_j^2).
\]

**Putting everything together.** Taking the inequalities we proved for the individual coordinates $i$ and summing them up, we have

\[
A \overset{k+2}{\implies} \|y\|^2 \cdot Q_{T^i}(X), x \otimes y \otimes z \leq 6\alpha \sum_{i=r+1}^d y_i^2(x_i^2 + z_i^2) + 8\alpha \sum_{i=1}^d \sum_{j \neq i} y_j^2(x_j^2 + y_j^2 + z_j^2).
\]

Since \(\{\|y\|^2 = 1\} \subset A\), we get

\[
A \overset{k+2}{\implies} Q_{T^i} (X), x \otimes y \otimes z \leq 6\alpha \sum_{i=r+1}^d (x_i^2 + z_i^2) + 8\alpha \sum_{i=1}^d \sum_{j \neq i} y_j^2(x_j^2 + y_j^2 + z_j^2).
\]

Returning to (28), this inequality plus the earlier bound from Corollary 19 imply

\[
A \overset{k+2}{\implies} X^* + Q_{T^i} (X), x \otimes y \otimes z \\
\leq 1 + (6\alpha - 1/4) \sum_{i=r+1}^d (x_i^2 + z_i^2) + (8\alpha - 1/8) \sum_{i=1}^d \sum_{j \neq i} y_j^2(x_j^2 + y_j^2 + z_j^2).
\]

\[
\leq 1, \quad \text{for } \alpha \leq 1/64.
\]

By the duality of SoS proofs and pseudodistributions, we have \(\|X^* + Q_{T^i} (X)\|_{\bar{m}_{k+2}^h} \leq 1\) as desired.

**Proof of Theorem 4** We first establish equation (14). We combine the assumption that \(\|T'\|_{\bar{m}_{k+2}^h} \leq \|T\|_{\bar{m}_{k+2}^h}\) with equation (13) to get

\[
\|T\|_{\bar{m}_{k+2}^h} \geq \|T'\|_{\bar{m}_{k+2}^h} \geq \|T\|_{\bar{m}_{k+2}^h} + \langle X^* + Q_{T^i} (X), \Delta \rangle.
\]

for all \(X\) with \(\|X\|_{\bar{m}_{k+2}^h} \leq 1/64\) and \(X^*\) as in Theorem 5. Rearranging, this yields

\[
\langle X, Q_{T^i} (\Delta) \rangle = \langle Q_{T^i} (X), \Delta \rangle \leq - \langle X^*, \Delta \rangle = -\langle X^*, Q_{T^i} (\Delta) \rangle.
\]
We now establish equation (9). Observe that we can write

\[ \langle X, Q_T(\Delta) \rangle = \frac{1}{64} \cdot \| Q_T(\Delta) \|_{\text{nuc}} \leq \| Q_T^0(\Delta) \|_{\text{nuc}}. \]

We now establish equation (9). Observe that we can write

\[ \| \Delta \|_{\text{nuc}} = \| Q_T(\Delta) + Q_T(\Delta) \|_{\text{nuc}} \leq \| Q_T(\Delta) \|_{\text{nuc}} + \| Q_T(\Delta) \|_{\text{nuc}}. \]

Combining this with (14), we get

\[ \| \Delta \|_{\text{nuc}} \leq \| Q_T(\Delta) \|_{\text{nuc}} + 64 \| Q_T^0(\Delta) \|_{\text{nuc}}. \]

Using the triangle inequality, we upper bound the first term as

\[ \| Q_T(\Delta) \|_{\text{nuc}} \leq \sum_{i=1}^{4} \| Q_T^i(\Delta) \|_{\text{nuc}} \leq \sum_{i=1}^{4} \| Q_T^i(\Delta) \|_{\text{nuc}}. \]

To proceed, we flatten each tensor in the summation above into a matrix and use Lemma 16 to show that the nuclear norm of the flattening leads to an upper bound. Let \( r_1 = \dim U, r_2 = \dim V, \) and \( r_3 = \dim W. \) Then the following inequalities hold

\[ \| Q_T^1(\Delta) \|_{\text{nuc}} \leq \sqrt{r_2} \| b_2(Q_T^1(\Delta)) \|_{\text{nuc}} \leq \sqrt{r_2} \| Q_T^1(\Delta) \|_F. \]
\[ \| Q_T^2(\Delta) \|_{\text{nuc}} \leq \sqrt{r_3} \| b_2(Q_T^2(\Delta)) \|_{\text{nuc}} \leq \sqrt{r_3} \| Q_T^2(\Delta) \|_F. \]
\[ \| Q_T^3(\Delta) \|_{\text{nuc}} \leq \sqrt{r_1} \| b_2(Q_T^3(\Delta)) \|_{\text{nuc}} \leq \sqrt{r_1} \| Q_T^3(\Delta) \|_F. \]

The first inequality in each line above follows from Lemma 16 and the definitions in (11). The second follows from the fact that \( \| A \|_{\text{nuc}} \leq \sqrt{\text{rank}(A)} \| A \|_F \) for any matrix \( A, \) along with the fact that \( \text{rank}(b_i(T)) = r_i(T) \) for any tensor, and that flattening does not change the entrywise \( \ell_2 \) norm. We have \( \| Q_T^0(\Delta) \|_{\text{nuc}} \leq \sqrt{r_2} \| Q_T^0(\Delta) \|_F \) as well by the same argument, though the choice of \( r_1/r_2/r_3 \) in this case is arbitrary.

To combine all the bounds, we use that \( r_1, r_2, r_3 \leq r \) and that orthogonal projection decreases the \( \ell_2 \) norm, which yields

\[ \| \Delta \|_{\text{nuc}} \leq 68r \| \Delta \|_F. \]

Appendix C. Proofs from Section 3

This section of the appendix is structured as follows. First, in Appendix C.1, we provide we provide a generalization bound for general classes of tensors and measurement models, from which all of our main statistical results will follow as special cases. This bound assumes that a restricted eigenvalue-type property holds for the tensor class and measurement model under consideration. In Appendix C.2 and Appendix C.3 we establish that this restricted eigenvalue property holds for the measurement models in Section 3. In Appendix C.4 we combine these results to prove the main results of that section.
C.1. Agnostic generalization bounds for tensor classes

In this section we given generalization guarantees for empirical risk minimization in a general learning setup. We receive a set of examples $S := \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ i.i.d. from a distribution $\mathcal{D}$ over $(\mathbb{R}^d)^ \otimes 3 \times \mathbb{R}$. We assume that a convex class of tensors $\mathcal{T} \subseteq (\mathbb{R}^d)^ \otimes 3$ is given, and that our goal is to achieve excess risk against an unknown benchmark $T^* \in \mathcal{T}$. We analyze the performance of empirical risk minimization over $\mathcal{T}$:

$$\hat{T}_n = \arg \min_{T \in \mathcal{T}} \hat{L}_n(T),$$

where $\hat{L}_n$ is the empirical square loss. The main result from this section is Theorem 20, which bounds the performance of ERM under various assumptions on the data distribution, the class $\mathcal{T}$, and the benchmark $T^*$. To state the result, recall from Section 3 that $\Sigma$ is the population correlation matrix and $\mathcal{X}_n$ is the empirical design operator.

**Theorem 20** Let a benchmark $T^* \in \mathcal{T}$ be fixed, and let $\xi_t = (Y_t - \langle T^*, X_t \rangle)$. Suppose there exist a pair of dual norms $\| \cdot \|_p$ and $\| \cdot \|_q$, for which the following conditions hold:

1. There are constants $0 \leq c < 2$, $\gamma_n > 0$, and $\delta_0 \geq 0$ such that with probability at least $1 - \delta_0$,

$$\| \Sigma^{1/2} \mathcal{X} \|_p^p \leq \frac{c}{n} \| \mathcal{X}_n(\mathcal{X}) \|_F^p + \gamma_n \quad \forall \Delta \in \mathcal{T} - T^* \quad \text{(Property 1)}. $$

2. There are constants $M \geq 0$ and $\delta_1 \geq 0$ such that with probability at least $1 - \delta_1$,

$$\left\| \sum_{t=1}^{n} \xi_t X_t - \mathbb{E}[\xi X] \right\| \leq M \cdot \sqrt{n} \quad \text{(Property 2)}. $$

3. There is a constant $\kappa \geq 0$ such that

$$\| \Delta \|_F^2 \leq \kappa^2 \cdot \| \Sigma^{1/2} \mathcal{X} \|_F^2 \quad \forall \Delta \in \mathcal{T} - T^* \quad \text{(Property 3)}. $$

Then with probability at least $1 - (\delta_0 + \delta_1)$,

$$L_D(\hat{T}_n) - L_D(T^*) \leq \frac{2\kappa^2 M^2}{c'n} + 2\gamma_n. \quad (32)$$

**Proof of Theorem 20** Since $\mathcal{T}$ is convex, and since $\hat{T}_n$ minimizes the (strongly convex) empirical risk, we have

$$\hat{L}_n(T^*) - \hat{L}_n(\hat{T}_n) \geq \langle \nabla \hat{L}_n(\hat{T}_n), T^* - \hat{T}_n \rangle + \mathbb{E}_n(\hat{T}_n - T^*, X)^2 \geq \mathbb{E}_n(\hat{T}_n - T^*, X)^2. $$

By rearranging and expanding the definition of $\hat{L}_n$, this implies

$$\mathbb{E}_n((T^*, X) - Y)^2 - \mathbb{E}_n((\hat{T}_n, X) - Y)^2 - \mathbb{E}_n(\hat{T}_n - T^*, X)^2 \geq 0.$$
Since the left-hand side is non-negative, we can add it to the population excess risk, which implies
\[ \mathbb{E}(\langle \hat{T}_n, X \rangle - Y)^2 - \mathbb{E}(\langle T^*, X \rangle - Y)^2 \]
\[ \leq \mathbb{E}(\langle \hat{T}_n, X \rangle - Y)^2 - \mathbb{E}(\langle T^*, X \rangle - Y)^2 \]
\[ + \mathbb{E}_n((T^*, X) - Y)^2 - \mathbb{E}_n((\hat{T}_n, X) - Y)^2 - \mathbb{E}_n(\hat{T}_n - T^*, X)^2. \]
Rearranging, this is equal to
\[ = (\mathbb{E} - \mathbb{E}_n)[2(\hat{T}_n - T^*, X)(\langle T^*, X \rangle - Y)] + \mathbb{E}(\hat{T}_n - T^*, X)^2 - 2\mathbb{E}_n(\hat{T}_n - T^*, X)^2. \]
Since \( \hat{T}_n - T^* \) is an element of \( \mathcal{T} - T^* \), we move to an upper bound by taking a supremum over elements of this set.
\[ \leq \sup_{\Delta \in \mathcal{T} - T^*} \{ (\mathbb{E} - \mathbb{E}_n)[2(\Delta, X)(\langle T^*, X \rangle - Y)] + \mathbb{E}(\Delta, X)^2 - 2\mathbb{E}_n(\Delta, X)^2 \} + \gamma_n. \]
This establishes inequality (18). It remains to use the assumptions in the theorem statement to bound the process. Property 1 states that with probability at least \( 1 - \delta_0 \), \( \mathbb{E}(\Delta, X)^2 \leq c \cdot \mathbb{E}_n(\Delta, X)^2 + \gamma_n \) for all \( \Delta \in \mathcal{T} - T^* \). Define \( c' = 2 - c > 0 \), so that conditioned on this event we have
\[ \sup_{\Delta \in \mathcal{T} - T^*} \{ (\mathbb{E} - \mathbb{E}_n)[2(\Delta, X)(\langle T^*, X \rangle - Y)] + \mathbb{E}(\Delta, X)^2 - 2\mathbb{E}_n(\Delta, X)^2 \} \]
\[ \leq \sup_{\Delta \in \mathcal{T} - T^*} \{ (\mathbb{E} - \mathbb{E}_n)[2(\Delta, X)(\langle T^*, X \rangle - Y)] - c' \cdot \mathbb{E}_n(\Delta, X)^2 \} + \gamma_n. \]
We expand the first term in the supremum as
\[ (\mathbb{E} - \mathbb{E}_n)[2(\Delta, X)(\langle T^*, X \rangle - Y)] = -2 \frac{1}{n} \sum_{t=1}^{n} \xi_t(\Delta, X_t) - \mathbb{E}[\xi(\Delta, X)] \]
\[ = -2 \left\{ \frac{1}{n} \sum_{t=1}^{n} \xi_t X_t - \mathbb{E}[\xi X], \Delta \right\}. \]
It follows from Hölder’s inequality that this expression is bounded as
\[ 2 \left\| \frac{1}{n} \sum_{t=1}^{n} \xi_t X_t - \mathbb{E}[\xi X] \right\| \cdot \| \Delta \|. \]
Defining \( \psi_n = \left\| \frac{1}{n} \sum_{t=1}^{n} \xi_t X_t - \mathbb{E}[\xi X] \right\| \), the development so far states that with probability at least \( 1 - \delta_0 \),
\[ \mathbb{E}(\langle \hat{T}_n, X \rangle - Y)^2 - \mathbb{E}(\langle T^*, X \rangle - Y)^2 \leq \sup_{\Delta \in \mathcal{T} - T^*} \left\{ 2\psi_n \cdot \| \Delta \| \cdot c' \cdot \frac{1}{n} \| \chi_n(\Delta) \|_2^2 \right\} + \gamma_n. \]
Using Property 3 we upper bound the leading term by
\[ \sup_{\Delta \in \mathcal{T} - T^*} \left\{ 2\kappa \psi_n \cdot \| \Sigma^{1/2} \Delta \|_F - c' \cdot \frac{1}{n} \| \chi_n(\Delta) \|_2^2 \right\}, \]
and the event we already conditioned on implies that this is at most

$$\sup_{\Delta \in T^*} \left\{ 4\kappa \psi_n \sqrt{\frac{1}{n}} \|X_n(\Delta)\|_2 - c' \cdot \frac{1}{n} \|X_n(\Delta)\|_2 \right\} + 2\kappa \psi_n \sqrt{\gamma_n} \leq \frac{\kappa^2 \psi_n^2}{c'} + 2\kappa \psi_n \sqrt{\gamma_n},$$

where the second inequality follows from AM-GM. Finally, by Property 2 we have that with probability at least $1 - \delta_1$, $\psi_n \leq M/\sqrt{n}$, which leads to the final bound of

$$\frac{\kappa^2 M^2}{c'n} + 2\kappa M \sqrt{\gamma_n/n} + \gamma_n \leq \frac{2\kappa^2 M^2}{c'n} + 2\gamma_n.$$

\(\blacksquare\)

### C.2. Restricted eigenvalue for tensor completion

The main result in this section is Theorem 21, which relates the empirical covariance and population covariance for all tensors with bounded entries and bounded SoS nuclear norm under sampling model for tensor completion. This result is then used in the proof of Theorem 6 to establish a restricted eigenvalue guarantee.

**Theorem 21** Let $d^{3/2} \leq n \leq d^3$, and let $\varepsilon \in (0, 1/2)$. Suppose observed entries are drawn with replacement. Then for any $\delta > 0$, with probability at least $1 - \delta$, all $T \in (\mathbb{R}^d)^{\otimes 3}$ with $\|T\|_\infty \leq R$ satisfy

$$\|\Sigma^{1/2} T\|_F^2 \leq \frac{1 + 2\varepsilon}{n} \|X_n(T)\|_2^2 + O\left( \frac{\log d}{nd^{3/2}} + \frac{R^2 \log(\log(d^{3/2}/\sqrt{n})/\delta)}{\varepsilon n} \right).$$

The proofs in this section pass back and forth between the with-replacement sampling model for tensor completion used in the main body of the paper and a without-replacement model, in which each entry is only observed a single time for tensor completion. To distinguish between the models, we use $S \sim D^n_{w/1}$ to refer to the draw of the dataset under the with-replacement sampling model and $S \sim D^n_{w/o}$ to refer to the draw under the without-replacement model. $\Omega \subseteq [d]^3$ will denote the support of the set of observed entries.

Before proving Theorem 21, we state and prove a number of auxiliary lemmas, from which the main result will follow.

**Proposition 22** Let $N_{j,k} = \{|t| \in [n] \mid j_t = j, k_t = k\}$. Then

$$\mathbb{P}_{S \sim D^n_{w/o}} \left( N_{j,k} \geq \frac{2n}{d^2} + \frac{10 \log(1/\delta)}{3} \right) \leq \delta.$$ 

**Proof of Proposition 22** This is an immediate consequence of Bernstein’s inequality for without-replacement sampling. See Bardenet et al. (2015), Proposition 1.4. \(\blacksquare\)
Lemma 23  The Rademacher complexity of $\| \cdot \|_{\hat{m}_{ij}}$ under without-replacement sampling is bounded as

$$\mathbb{E}_{S \sim \mathcal{D}_{w/o}} \mathbb{E}_r \left\| \sum_{t=1}^n \epsilon_t X_t \right\|_{\hat{m}_{ij}} \leq O \left( \sqrt{\frac{n \log^4 d}{d^{3/2}}} + \sqrt{\log d} \right).$$

(33)

Additionally, if $n \leq d^3$, then the Rademacher complexity under with-replacement sampling is bounded as

$$\mathbb{E}_{S \sim \mathcal{D}_{w/o}} \mathbb{E}_r \left\| \sum_{t=1}^n \epsilon_t X_t \right\|_{\hat{m}_{ij}} \leq O \left( \sqrt{\frac{n \log^6 d}{d^{3/2}}} + \sqrt{\log^3 d} \right).$$

(34)

Proof of Lemma 23  We first bound the Rademacher complexity in the without-replacement sampling case, then handle the with-replacement case by reduction. This analysis follows Barak and Moitra (2016), except that we handle certain “diagonal” terms that arise in the analysis slightly more carefully so as to get the right scaling for our setup, which differs from theirs in that it does not assume incoherence.

Consider a fixed draw of $\epsilon_{1:n}$ and $X_{1:n}$, and let $Z = \sum_{t=1}^n \epsilon_t X_t$. Letting $x, y, z$ be indeterminates, we will give a degree-four SoS upper bound on the polynomial $(Z, x \otimes y \otimes z)$. This will imply that the pseudoexpectation of $(Z, x \otimes y \otimes z)$ is bounded for any feasible pseudodistribution in the maximization problem defining $\| \cdot \|_{\hat{m}_{ij}}$.

To begin, for any fixed constant $\eta > 0$, Lemma 10 implies that

$$\imath_{\preceq 4} \langle Z, x \otimes y \otimes z \rangle = \sum_{i,j,k} Z_{i,j,k} x_i y_j z_k \leq \frac{1}{2\eta} \sum_{i=1}^d x_i^2 + \frac{\eta}{2} \sum_{i,j,k} \left( \sum_{j,k} Z_{i,j,k} y_j z_k \right)^2,$$

and so

$$\{ \| x \|_2^2 = 1 \} \imath_{\preceq 4} \langle Z, x \otimes y \otimes z \rangle \leq \frac{1}{2\eta} + \frac{\eta}{2} \sum_{i,j,k} \left( \sum_{j,k} Z_{i,j,k} y_j z_k \right)^2.$$

Define a matrix $A \in \mathbb{R}^{d^2 \times d^2}$ via $A_{j,k',j',k} = \sum_{i=1}^d Z_{i,j,k} Z_{i,j',k'}$. Define additional matrices $D, B \in \mathbb{R}^{d^2 \times d^2}$ via

$$D_{j,k,j',k'} = \begin{cases} A_{j,k,j',k}, & \text{if } (j, k) = (j', k'), \\ 0, & \text{otherwise,} \end{cases}$$

and $B = A - D$. Then we have

$$\sum_i \left( \sum_{j,k} Z_{i,j,k} y_j z_k \right)^2 = \{ A, (y \otimes z)(y \otimes z)^\top \} = \{ B, (y \otimes z)(y \otimes z)^\top \} + \{ D, (y \otimes z)(y \otimes z)^\top \}.$$
and so
\[ \{ \| y \|_2^2 = 1, \| z \|_2^2 = 1 \} \preceq_4 \{ D, (y \otimes z)(y \otimes z)^T \} \leq \| R \|_\infty. \]

By the duality of sum-of-squares proofs and pseudodistributions, this implies that any degree-four pseudodistribution \( \mu \) with \( \mu = \{ \| x \|_2^2 = 1, \| y \|_2^2 = 1, \| z \|_2^2 = 1 \} \) has \( \mathbb{E}(Z, x \otimes y \otimes z) \leq \frac{4}{2}(\| B \|_{\text{op}} + \| R \|_\infty) + \frac{1}{2\eta} \). Optimizing over \( \eta \), we conclude that
\[ \| Z \|_{\hat{m}_j^1} \leq \sqrt{\| B \|_{\text{op}} + \| R \|_\infty}. \]

The bound for the matrix \( B \) is taken care of Theorem 4.4 of Barak and Moitra (2016), which implies that\( \frac{\sqrt{\mathbb{E}_{S \sim D_{w_o}^n} \| B \|_2^2}}{\mathbb{E}_x \| B \|_{\text{op}}} \leq O\left( \frac{n \log d^4}{d^{3/2}} \right) \). For the matrix \( R \), observe that under the without-replacement sampling model, for any \( j, k \) we have
\[ R_{j,k} = \sum_{i=1}^d Z_{i,j,k}^2 = \sum_{i=1}^d 1 \{ (i, j, k) \in \Omega \} = | \{ t \in [n] : j_t = j, k_t = k \} |. \]

Proposition 22 thus implies that for any fixed \( j, k \), with probability at least \( 1 - \delta \),
\[ R_{j,k} \leq O\left( n/d^2 + \log(1/\delta) \right), \]
and so, by taking a union bound and integrating out the tail, we have
\[ \sqrt{\mathbb{E}_{S \sim D_{w_o}^n} \| R \|_\infty^2} \leq O\left( n/d^2 + \log d \right). \]

Combining the bounds on \( B \) and \( R \) and using Jensen’s inequality yields
\[ \mathbb{E}_{S \sim D_{w_o}^n} \mathbb{E}_x \| Z \|_{\hat{m}_j^1} \leq \sqrt{\mathbb{E}_{S \sim D_{w_o}^n} \mathbb{E}_x \| Z \|_{\hat{m}_j^1}^2} \leq O\left( \sqrt{\frac{n \log d}{d^{3/2}}} + \sqrt{\log d} \right). \tag{35} \]

This completes the bound for the without-replacement case. For the with-replacement case we reduce to the bound on the second moment above. Condition on the draw of \( S \), and let \( T_{i,j,k} = \{ t \in [n] : i_t = i, j_t = j, k_t = k \} \). Then we have
\[ \mathbb{E}_t \left\| \sum_{i=1}^n \epsilon_t X_i \right\|_{\hat{m}_j^1} = \mathbb{E}_t \left\| \sum_{(i,j,k) \in \Omega} \left( \sum_{t \in T_{i,j,k}} \epsilon_t \right) e_i \otimes e_j \otimes e_k \right\|_{\hat{m}_j^1}. \]

Introduce a new sequence of Rademacher random variables \( \sigma \in \{ \pm 1 \}^{d^3} \). Then the right-hand side above is equal to
\[ \mathbb{E}_t \mathbb{E}_\sigma \left\| \sum_{(i,j,k) \in \Omega} \left( \sum_{t \in T_{i,j,k}} \epsilon_t \right) \sigma_{i,j,k} \cdot e_i \otimes e_j \otimes e_k \right\|_{\hat{m}_j^1} \]
\[ \leq \mathbb{E}_t \max_{i,j,k} \left\| \sum_{t \in T_{i,j,k}} \epsilon_t \right\| \cdot \mathbb{E}_\sigma \left\| \sum_{(i,j,k) \in \Omega} \sigma_{i,j,k} \cdot e_i \otimes e_j \otimes e_k \right\|_{\hat{m}_j^1}, \]

28
where the inequality follows from the standard Lipschitz contraction lemma for Rademacher complexity (Ledoux and Talagrand, 1991). By the Hoeffding bound, we have that for any fixed index \((i, j, k)\) with probability at least \(1 - \delta\),

\[
\left| \sum_{t \in T_{i,j,k}} \epsilon_t \right| \leq O \left( \max_{i,j,k} \sqrt{|T_{i,j,k}| \log (1/\delta)} \right).
\]

Taking a union bound and integrating out the tail, we have

\[
\mathbb{E}_{\epsilon} \max_{i,j,k} \left| \sum_{t \in T_{i,j,k}} \epsilon_t \right| \leq O \left( \max_{i,j,k} \sqrt{|T_{i,j,k}| \log d} \right).
\]

We now move to the final bound by taking the expectation over \(S\). Using Cauchy-Schwarz, the development above implies

\[
\mathbb{E}_{\mathcal{S}} \mathbb{E}_{\mathcal{D}^m_{wr}} \mathbb{E}_{\mathcal{E}} \left\| \sum_{t=1}^{n} \epsilon_t X_t \right\|_{\infty} \leq \left( \sqrt{\mathbb{E}_{\mathcal{S}} \mathbb{E}_{\mathcal{D}^m_{wr}} \mathbb{E}_{\mathcal{E}} \left| T_{i,j,k} \right| \log d} \right) \cdot \sqrt{\mathbb{E}_{\mathcal{S}} \mathbb{E}_{\mathcal{D}^m_{wr}} \mathbb{E}_{\mathcal{E}} \left( \sum_{(i,j,k) \in \Omega} \sigma_{i,j,k} \cdot e_i \otimes e_j \otimes e_k \right)^2}.
\]

Bernstein’s inequality and the union bound imply that

\[
\sqrt{\mathbb{E}_{\mathcal{S}} \mathbb{E}_{\mathcal{D}^m_{wr}} \mathbb{E}_{\mathcal{E}} \left| T_{i,j,k} \right| \log d} \leq O \left( \sqrt{\log d (n/d^3 + \log d)} \right) \leq O(\log d),
\]

where the second inequality uses that \(n \leq d^3\). For the second term, we have

\[
\mathbb{E}_{\mathcal{S}} \mathbb{E}_{\mathcal{D}^m_{wr}} \mathbb{E}_{\mathcal{E}} \left\| \sum_{(i,j,k) \in \Omega} \sigma_{i,j,k} \cdot e_i \otimes e_j \otimes e_k \right\|^2_{\infty} \\
= \sum_{m=1}^{n} \mathbb{P}_{\mathcal{S} \mathcal{D}^m_{wr}} (|\Omega| = m) \cdot \mathbb{E}_{\mathcal{S}} \mathbb{E}_{\mathcal{D}^m_{wr}} \mathbb{E}_{\mathcal{E}} \left\| \sum_{(i,j,k) \in \Omega} \sigma_{i,j,k} \cdot e_i \otimes e_j \otimes e_k \right\|^2_{\infty} \\
\leq \sum_{m=1}^{n} \mathbb{P}_{\mathcal{S} \mathcal{D}^m_{wr}} (|\Omega| = m) \cdot O \left( \frac{m \log^4 d}{d^{3/2}} + \log d \right),
\]

where the second inequality uses the bound we proved for the without-replacement setting and the second inequality uses \(m \leq n\) and that the probabilities sum to one. Combining these two bounds completes the proof.
Lemma 24  Define \( \mathcal{T} = \{ T \in (\mathbb{R}^d)^{\otimes 3} \mid \|T\|_{\text{nc}_4} \leq \tau, \|\Sigma^{1/2} T\|_F \leq \beta, \|T\|_\infty \leq R \} \). Under the with-replacement sampling model, when \( d^{3/2} \leq n \leq d^3 \), we have that for any \( \delta > 0 \), with probability at least \( 1 - \delta \),
\[
\frac{1}{n} \|X_n(T)\|_F^2 - \|\Sigma^{1/2} T\|_F^2 \leq O \left( R\tau \sqrt{\frac{\log^6 d}{nd^{3/2}}} + \sqrt{\frac{2R^2 \beta^2 \log(1/\delta)}{n}} + \frac{2R^2 \log(1/\delta)}{n} \right) \quad \forall T \in \mathcal{T}.
\]

Proof of Lemma 24  To begin, we write
\[
\frac{1}{n} \|X_n(T)\|_F^2 - \|\Sigma^{1/2} T\|_F^2 = \|\mathbb{E}_n(T, X)^2 - \mathbb{E}(T, X)^2\|.
\]

Using this representation, since entries are drawn i.i.d. with replacement, we can apply Lemma 17 with the function class \( \mathcal{F} = \{ X \mapsto (T, X)^2 - \mathbb{E}(T, X)^2 \mid T \in \mathcal{T} \} \). In particular, note that for all \( T \in \mathcal{T} \) we have \( \|(T, X_t)\| = |T_{i_t, j_t, k_t}| \leq R \), and furthermore
\[
\sup_{T \in \mathcal{T}} \mathbb{E}( (T, X)^2 - \mathbb{E}(T, X)^2 )^2 \leq \sup_{T \in \mathcal{T}} \mathbb{E}(T, X)^4 \leq R^2 \sup_{T \in \mathcal{T}} \mathbb{E}(T, X)^2 = R^2 \sup_{T \in \mathcal{T}} \|\Sigma^{1/2} T\|_F^2 \leq R^2 \beta^2.
\]

Consequently, Lemma 17 implies that with probability at least \( 1 - \delta \), for all \( T \in \mathcal{T} \),
\[
\frac{1}{n} \|X_n(T)\|_F^2 - \|\Sigma^{1/2} T\|_F^2 \leq 4 \mathbb{E}_S \mathbb{E}_e \sup_{T \in \mathcal{T}} \frac{1}{n} \sum_{t=1}^n \epsilon_t ((T, X_t)^2 - \mathbb{E}(T, X)^2) + \sqrt{\frac{2R^2 \beta^2 \log(1/\delta)}{n}} + \frac{2R^2 \log(1/\delta)}{n}.
\]

Using Jensen’s inequality and splitting the supremum, we have
\[
\mathbb{E}_S \mathbb{E}_e \sup_{T \in \mathcal{T}} \frac{1}{n} \sum_{t=1}^n \epsilon_t ((T, X_t)^2 - \mathbb{E}(T, X)^2) \leq 2 \mathbb{E}_S \mathbb{E}_e \sup_{T \in \mathcal{T}} \frac{1}{n} \sum_{t=1}^n \epsilon_t (T, X_t)^2.
\]

Using the Lipschitz contraction lemma for Rademacher complexity, we remove the square
\[
\mathbb{E}_S \mathbb{E}_e \sup_{T \in \mathcal{T}} \left[ \frac{1}{n} \sum_{t=1}^n \epsilon_t (T, X_t)^2 \right] \leq 2R \mathbb{E}_S \mathbb{E}_e \sup_{T \in \mathcal{T}} \left[ \frac{1}{n} \sum_{t=1}^n \epsilon_t (T, X_t) \right] \leq \frac{2R\tau}{n} \mathbb{E}_S \mathbb{E}_e \left\| \frac{1}{n} \sum_{t=1}^n \epsilon_t X_t \right\|_{\tilde{m}_{4,4}}.
\]

Finally, using Lemma 23, we have
\[
\mathbb{E}_S \mathbb{E}_e \left\| \frac{1}{n} \sum_{t=1}^n \epsilon_t X_t \right\|_{\tilde{m}_{4,4}} \leq O \left( \sqrt{\frac{\log^6 d}{nd^{3/2}}} + \frac{\log^{3/2} d}{n} \right) \leq O \left( \sqrt{\frac{\log^6 d}{nd^{3/2}}} + \frac{\log^{3/2} d}{n} \right).
\]

The final bound follows by using \( n \geq d^{3/2} \) to simplify this expression.

Proof of Theorem 21  Let \( \mathcal{T} = \{ T \in (\mathbb{R}^d)^{\otimes 3} \mid \|T\|_\infty \leq R \} \). Recall that for tensor completion, the population correlation matrix \( \Sigma \) under with-replacement sampling is equal to \( \frac{1}{d^3} I \).
Let \( \tau_{\text{max}} = Rd^3 \), \( \beta_{\text{max}} = Rd^3/2 \), \( \tau_{\text{min}} = Rd^{3/2}/\sqrt{n} \), \( \beta_{\text{min}} = R/\sqrt{n} \), and let \( N = \lceil \log(\tau_{\text{max}}/\tau_{\text{min}}) \rceil + 1 \) and \( M = \lceil \log(\beta_{\text{max}}/\beta_{\text{min}}) \rceil + 1 \). For each \( i \in [N] \) and \( j \in [M] \) define \( \tau_i = \tau_{\text{max}}e^{1-i} \) and \( \beta_j = \beta_{\text{max}}e^{1-j} \). Define

\[
\mathcal{T}_{i,j} = \{ T \in \mathcal{T} \mid \tau_{i+1} \leq \| T \|_{\text{nuc}} \leq \tau_i, \quad \beta_{j+1} \leq \| \Sigma^{1/2}T \|_F \leq \beta_j \}.
\]

Using Lemma 24 and a union bound, we get that with probability at least \( 1 - \delta \), for all \( i, j \) simultaneously, for all \( T \in \mathcal{T}_{i,j} \),

\[
\| \Sigma^{1/2}T \|_F^2 \leq \frac{1}{n} \| X_n(T) \|_2^2 + O \left( R\tau_i \sqrt{\frac{\log d}{nd^{3/2}}} + \| \Sigma^{1/2}T \|_F \sqrt{\frac{R^2\beta_j^2 \log (MN/\delta)}{n}} + \frac{R^2 \log (MN/\delta)}{n} \right).
\]

Now consider a fixed tensor \( T \in \mathcal{T} \). There are two cases. First, if \( \| T \|_{\text{nuc}} \geq \tau_{\text{min}} \) and \( \| \Sigma^{1/2}T \|_F \geq \beta_{\text{min}} \), then there must be indices \( i \) and \( j \) for which \( \tau_{i+1} \leq \| T \|_{\text{nuc}} \leq \tau_i \), \( \beta_{j+1} \leq \| \Sigma^{1/2}T \|_F \leq \beta_j \). Consequently, the uniform bound above implies

\[
\| \Sigma^{1/2}T \|_F^2 \leq \frac{1}{n} \| X_n(T) \|_2^2 + O \left( R\| T \|_{\text{nuc}} \sqrt{\frac{\log d}{nd^{3/2}}} + \| \Sigma^{1/2}T \|_F \sqrt{\frac{R^2 \log (MN/\delta)}{n}} + \frac{R^2 \log (MN/\delta)}{n} \right).
\]

On the other hand, if either \( \| T \|_{\text{nuc}} \leq \tau_{\text{min}} \) or \( \| \Sigma^{1/2}T \|_F \leq \beta_{\text{min}} \), we trivially have

\[
\| \Sigma^{1/2}T \|_F^2 \leq \frac{R^2}{n} \leq \frac{1}{n} \| X_n(T) \|_2^2 + \frac{R^2}{n}.
\]

Combining these cases, and using the values for \( N \) and \( M \), we get that with probability at least \( 1 - \delta \), for all \( T \in \mathcal{T} \),

\[
\| \Sigma^{1/2}T \|_F^2 \leq \frac{1}{n} \| X_n(T) \|_2^2 + O \left( R\| T \|_{\text{nuc}} \sqrt{\frac{\log d}{nd^{3/2}}} + \| \Sigma^{1/2}T \|_F \sqrt{\frac{R^2 \log (2d^{3/2}/\sqrt{n})/\delta)}{n}} + \frac{R^2 \log (2d^{3/2}/\sqrt{n})/\delta)}{n} \right).
\]

Using the AM-GM inequality on the second-to-last term and rearranging, we have that for any \( \varepsilon > 0 \),

\[
(1 - \varepsilon)\| \Sigma^{1/2}T \|_F^2 \leq \frac{1}{n} \| X_n(T) \|_2^2 + O \left( R\| T \|_{\text{nuc}} \sqrt{\frac{\log d}{nd^{3/2}}} + \frac{R^2 \log (2d^{3/2}/\sqrt{n})/\delta)}{\varepsilon n} \right).
\]

When \( \varepsilon \in (0, 1/2) \), this implies

\[
\| \Sigma^{1/2}T \|_F^2 \leq \frac{(1 + 2\varepsilon)}{n} \| X_n(T) \|_2^2 + O \left( R\| T \|_{\text{nuc}} \sqrt{\frac{\log d}{nd^{3/2}}} + \frac{R^2 \log (2d^{3/2}/\sqrt{n})/\delta)}{\varepsilon n} \right).
\]
C.3. Restricted eigenvalue for tensor sensing

In this section we prove the main technical result used to establish restricted eigenvalue guarantees for tensor sensing, which is as follows.

**Theorem 25** Suppose $X \sim \mathcal{N}(0 ; I)$. There is some universal constant $C > 0$ such that for any $\varepsilon < \sqrt{2/\pi}$, with probability at least $1 - 2e^{-\frac{n\varepsilon^2}{8}}/(1 - e^{-\frac{n\varepsilon^2}{8}})$,

$$\frac{1}{\sqrt{n}} \left\| X_n(\Delta) \right\|_2 \geq \left( \sqrt{2/\pi} - \varepsilon \right) \left\| \Delta \right\|_F - \frac{C\beta^{1/4} \log^{1/4} d}{\sqrt{n}} \cdot \Delta \langle \psi \rangle_{\tilde{m}_4} \quad \forall \Delta \in (\mathbb{R}^d)^{\otimes 3}. \quad (36)$$

To prove Theorem 25 we require two key technical lemmas.

**Lemma 26** Let $\left\| \cdot \right\|$ be any norm, let $\left\| \cdot \right\|_\ast$ be the dual. Suppose that the rows of $X_n$ are formed by drawing $X_1, \ldots, X_n$ i.i.d. from $\mathcal{N}(0 ; I_{p \times p})$, and let $\mathbb{E}_{X \sim \mathcal{N}(0 ; I_{p \times p})} \left| X \right| \leq \psi$. Then for any $\varepsilon < \sqrt{2/\pi}$, with probability at least $1 - 2e^{-\frac{n\varepsilon^2}{8}}/(1 - e^{-\frac{n\varepsilon^2}{8}})$,

$$\frac{1}{\sqrt{n}} \left\| X_n(\Delta) \right\|_2 \geq \left( \sqrt{2/\pi} - \varepsilon \right) \left\| \Delta \right\|_F - \frac{4\psi}{\sqrt{n}} \left\| \Delta \right\|_\ast \quad \forall \Delta \in \mathbb{R}^p. \quad (37)$$

**Lemma 27 (Corollary of Hopkins et al. (2015), Theorem 3.3)** Let $X \in (\mathbb{R}^d)^{\otimes 3}$ have entries drawn i.i.d. from $\mathcal{N}(0 ; 1)$. Then

$$\mathbb{E} \left\| X \right\|_{\tilde{m}_4} \leq O(d^{3/4} \log^{1/4} d). \quad (38)$$

**Proof of Theorem 25** This is an immediate consequence of Lemma 26 and Lemma 27, along with the duality of $\left\| \cdot \right\|_{\tilde{m}_4}$ and $\left\| \cdot \right\|_{\tilde{m}_4}$.

In the remainder of the section we prove Lemma 26. The result follows from fairly standard techniques (e.g. Wainwright (2019)), but we include the proof for completeness. We first restate some basic results on gaussian concentration.

**Lemma 28 (Concentration for Lipschitz functions (Milman and Schechtman, 1986))** Let $Z \in \mathbb{R}^n$ have entries drawn i.i.d. from $\mathcal{N}(0 ; 1)$, and let $f : \mathbb{R}^n \to \mathbb{R}$ be $L$-Lipschitz with respect to the $\ell_2$ norm. Then for all $t \geq 0$,

$$\mathbb{P}(\left| f(Z) - \mathbb{E} f(Z) \right| \geq t) \leq 2e^{-\frac{t^2}{2L^2}}. \quad (39)$$

**Lemma 29 (Gordon’s Inequality (Davidson and Szarek, 2001))** Let $\{Z_{a,b}\}$ and $\{Y_{a,b}\}$ be zero-mean gaussian processes indexed by $A \times B$. Suppose that

$$\mathbb{E} (Z_{a,b} - Z_{a',b'})^2 \leq \mathbb{E} (Y_{a,b} - Y_{a',b'})^2 \quad \text{for all } (a, b), (a', b') \in A \times B$$

and

$$\mathbb{E} (Z_{a,b} - Z_{a,b'})^2 = \mathbb{E} (Y_{a,b} - Y_{a,b'})^2 \quad \text{for all } a \in A, b, b' \in B.$$

Then

$$\mathbb{E} \sup_{a \in A} \inf_{b \in B} Z_{a,b} \leq \mathbb{E} \sup_{a \in A} \inf_{b \in B} Y_{a,b}. \quad (40)$$

32
Proof of Lemma 26

Part 1: Bound at a single scale. Define $B(\tau) = \{\Delta \in \mathbb{R}^p \mid \|\Delta\|_F = 1, \|\Delta\|_* \leq \tau\}$. We will prove that with probability at least $1 - 2e^{-\frac{\tau^2}{2\pi}}$,

$$\frac{1}{\sqrt{n}}\|\mathcal{X}_n(\Delta)\|_2 \geq \sqrt{\frac{2}{\pi}} - \frac{\psi \tau}{\sqrt{n}} - t, \quad \forall \Delta \in B(\tau).$$

We will prove a lower bound on the random variable $\min_{\Delta \in B(\tau)} \frac{1}{\sqrt{n}}\|\mathcal{X}_n(\Delta)\|_2$, which is equivalent to providing an upper bound on the random variable

$$S_n(\tau) = -\inf_{\Delta \in B(\tau)} \frac{1}{\sqrt{n}}\|\mathcal{X}_n(\Delta)\|_2 = -\inf_{\Delta \in B(\tau)} \sup_{\|u\|_2 = 1} \frac{1}{\sqrt{n}}(u, \mathcal{X}_n(\Delta)) = \sup_{\Delta \in B(\tau)} \inf_{\|u\|_2 = 1} \frac{1}{\sqrt{n}}(u, \mathcal{X}_n(\Delta)).$$

As a starting point, we show how to upper bound the expectation $\mathbb{E}[S_n(\tau)]$. Define $Z_{\Delta,u} = \frac{1}{\sqrt{n}}(u, \mathcal{X}_n(\Delta))$, so that $S_n(\tau) = \sup_{\Delta \in B(\tau)} \inf_{\|u\|_2 = 1} Z_{\Delta,u}$. Note that $Z_{\Delta,u}$ is a gaussian process with variance $n^{-1}$, since $\|\Delta\|_F = 1$. We now define a new gaussian process that will serve as an upper bound through Gordon’s inequality. Let $g \in \mathbb{R}^p$ and $h \in \mathbb{R}^n$ be standard gaussian random variables, and define

$$Y_{\Delta,u} = \frac{1}{\sqrt{n}}(g, \Delta) + \frac{1}{\sqrt{n}}(h, u).$$

Note that for any $(\Delta, u)$ and $(\Delta', u')$ we have

$$\mathbb{E}(Y_{\Delta,u} - Y_{\Delta',u'})^2 = \frac{1}{n}\|\Delta - \Delta'\|_F^2 + \frac{1}{n}\|u - u'\|_2^2.$$ Interpreting $\Delta$ and $\Delta'$ as vectors in $\mathbb{R}^{d^3}$, we also have

$$\mathbb{E}(Z_{\Delta,u} - Z_{\Delta',u'})^2 = \frac{1}{n}\|u\Delta^T - u'\Delta'^T\|_F^2$$

$$= \frac{1}{n}\|\Delta - \Delta'\|_F^2 + \frac{1}{n}\|u - u'\|_2^2 + \frac{1}{n}\{(u - u')\Delta^1, u'(\Delta - \Delta')\}$$

$$= \frac{1}{n}\|\Delta - \Delta'\|_F^2 + \frac{1}{n}\|u - u'\|_2^2 + \frac{1}{n}\{(u, u') - 1)(1 - \langle \Delta, \Delta' \rangle)$$

$$\leq \frac{1}{n}\|\Delta - \Delta'\|_F^2 + \frac{1}{n}\|u - u'\|_2^2$$

where we have used that $\|u\|_2 = \|u'\|_2 = \|\Delta\|_F = \|\Delta'\|_F = 1$. It is also easily seen from the representation above that for any triple $(\Delta, u, u')$ we have equality: $\mathbb{E}(Z_{\Delta,u} - Z_{\Delta,u'})^2 = \mathbb{E}(Y_{\Delta,u} - Y_{\Delta,u'})^2$. This means that the preconditions of Lemma 29 are satisfied, and so

$$\mathbb{E}[S_n(\tau)] \leq \mathbb{E} \sup_{\Delta \in B(\tau)} \inf_{\|u\|_2 = 1} Y_{\Delta,u}$$

$$= \frac{1}{\sqrt{n}} \mathbb{E}_g \sup_{\Delta \in B(\tau)} \langle g, \Delta \rangle + \mathbb{E}_h \frac{1}{\sqrt{n}} \min_{\|u\|_2 = 1} \langle h, u \rangle$$

$$= \frac{\tau}{\sqrt{n}} \mathbb{E}_g \|g\| - \frac{1}{\sqrt{n}} \mathbb{E}_h \|h\|_2$$

$$\leq \frac{\psi \tau}{\sqrt{n}} - \sqrt{\frac{2}{\pi}}.$$
The provides the desired upper bound in expectation. To establish the high probability result we appeal to gaussian concentration for Lipschitz functions. Let $X \in \mathbb{R}^{n \times d^3}$ denote the sequence of measurements $X_1, \ldots, X_n$, interpreted as a matrix with vectorized measurements as rows. Define $f : \mathbb{R}^{n \times d^3} \to \mathbb{R}$ via $f(X) = \min_{\Delta \in \mathcal{B}(\tau)} \frac{1}{\sqrt{n}} \|X\Delta\|_2$. Observe that we have

$$\left| f(X) - f(X') \right| \leq \sup_{\Delta \in \mathcal{B}(\tau)} \frac{1}{\sqrt{n}} \| (X - X') \Delta \|_2 \leq \frac{1}{\sqrt{n}} \| X - X' \|_{\text{op}} \leq \frac{1}{\sqrt{n}} \| X - X' \|_F,$$

and so $f$ is $n^{-\frac{1}{2}}$-Lipschitz with respect to $\ell_2$. Lemma 28 therefore implies that

$$\mathbb{P}( |S_n(\tau) - E S_n(\tau)| \geq t ) \leq 2e^{-\frac{nt^2}{2}}.$$

In other words, (39) holds.

**Part 2: Bound at all scales.** We will show that for any $\varepsilon < \sqrt{2/\pi}$, with probability at least $1 - 2e^{-\frac{n}{32}}/(1 - e^{-\frac{n^2}{8\pi}})$,

$$\frac{1}{\sqrt{n}} \| X_n(\Delta) \|_2 \geq \sqrt{\frac{2}{\pi}} - \varepsilon - \frac{4\psi}{\sqrt{n}} \| \Delta \|_*, \quad \forall \Delta : \| \Delta \|_F = 1. \tag{40}$$

Define $\mathcal{B}(\tau_\ell, \tau_u) = \{ \Delta \in \mathbb{R}^p : \| \Delta \|_F \leq \tau, \tau \leq \frac{\psi}{\sqrt{n}} \| \Delta \| \leq \tau_u \}$. Set $\mu = \varepsilon/2$ and consider the classes $\mathcal{B}(0, \mu)$ and $\mathcal{B}(2^{i-1} \mu, 2^i \mu)$ for $i \in \mathbb{N}$. Note that if equation (40) fails to hold and $\Delta \in \mathcal{B}(0, \mu)$ then

$$\frac{1}{\sqrt{n}} \| X_n(\Delta) \|_2 \leq \sqrt{\frac{2}{\pi}} - \varepsilon - \frac{4\psi}{\sqrt{n}} \| \Delta \|_* \leq \sqrt{\frac{2}{\pi}} - \varepsilon = \sqrt{\frac{2}{\pi}} - 2\mu.$$

Furthermore, if equation (40) fails to hold and $\Delta \in \mathcal{B}(2^{i-1} \mu, 2^i \mu)$, then

$$\frac{1}{\sqrt{n}} \| X_n(\Delta) \|_2 \leq \sqrt{\frac{2}{\pi}} - \varepsilon - \frac{4\psi}{\sqrt{n}} \| \Delta \|_* \leq \sqrt{\frac{2}{\pi}} - 2\cdot2^i \mu.$$

Our development in part 1 of the proof implies that for any fixed $\tau_\ell \leq \tau_u$, with probability at least $1 - 2e^{-\frac{n}{32}} e^{-\frac{n^2}{8\pi}}$,

$$\frac{1}{\sqrt{n}} \| X_n(\Delta) \|_2 \geq \sqrt{\frac{2}{\pi}} - 2\tau_u, \quad \forall \Delta \in \mathcal{B}(\tau_\ell, \tau_u). \tag{41}$$

Thus, by a union bound, we get that with probability at least $1 - 2e^{-\frac{n}{32}} \sum_{i=0}^{\infty} e^{-\frac{2^{2i+1} \mu^2}{8\pi}}$, or conservatively at least $1 - 2e^{-\frac{n}{32}}/(1 - e^{-\frac{n^2}{8\pi}})$,

$$\frac{1}{\sqrt{n}} \| X_n(\Delta) \|_2 \geq \sqrt{\frac{2}{\pi}} - 2\cdot2^i \mu, \quad \forall \Delta \in \mathcal{B}(2^{i-1} \mu, 2^i \mu), \quad \forall i \in \mathbb{N},$$

and

$$\frac{1}{\sqrt{n}} \| X_n(\Delta) \|_2 \geq \sqrt{\frac{2}{\pi}} - 2\mu, \quad \forall \Delta \in \mathcal{B}(0, \mu),$$

or in other words, equation (40) holds.
We extend the guarantee in equation (40) to arbitrary \( \Delta \in \mathbb{R}^p \) by rescaling so that \( \| \Delta \|_F = 1 \) and then exploiting homogeneity to get

\[
\frac{1}{\sqrt{n}} \| \chi_n(\Delta) \|_2 \geq \left( \sqrt{\frac{2}{\pi}} - \varepsilon \right) \| \Delta \|_F - \frac{4\psi}{\sqrt{n}} \| \Delta \|_\infty, \quad \forall \Delta \in \mathbb{R}^p.
\]

\[\blacksquare\]

**C.4. Proofs of main results**

**Proof of Theorem 6** We prove the theorem by appealing to the generic result of Theorem 20 using norms \( \| \cdot \| = \| \cdot \|_{\tilde{\mu}_3} \) and \( \| \cdot \|_\infty = \| \cdot \|_{\tilde{\mu}_{\infty}} \). Let \( T^* \) be an arbitrary rank-\( r \) orthogonal tensor with \( \| T^* \|_{\tilde{\mu}_{\infty}} = \tau \) and \( \| T^* \|_\infty \leq R \). First, observe that all elements \( T \in \mathcal{T} \) have \( \| T \|_{\tilde{\mu}_3} \leq \| T^* \|_{\tilde{\mu}_{\infty}} \). Consequently, Theorem 4 implies that all \( \Delta \in \mathcal{T} - T^* \) satisfy

\[
\| \Delta \|_{\tilde{\mu}_3} \leq 68r \cdot \| \Delta \|_F,
\]

which establishes Property 3 with \( \kappa = O(r \cdot d^{3/2}) \). To establish Property 2 we appeal to Theorem 21, which implies that for any \( \varepsilon < 1/2 \), with probability at least \( 1 - \delta \), all \( \Delta \in \mathcal{T} - T^* \) satisfy

\[
\left\| \Sigma^{1/2} \Delta \right\|_F^2 \leq \frac{1 + 2\varepsilon}{n} \left\| \chi_n(\Delta) \right\|_2^2 + O \left( R \left\| \Delta \right\|_{\tilde{\mu}_3} \sqrt{\frac{r^2 \log^6 d}{nd^{3/2}}} + \frac{R^2 \log^2 (d^{3/2} \sqrt{n})/\delta}{\varepsilon n} \right).
\]

Using equation (42) this is upper bounded by

\[
\left\| \Sigma^{1/2} \Delta \right\|_F^2 \leq \frac{1 + 2\varepsilon}{n} \left\| \chi_n(\Delta) \right\|_2^2 + O \left( R \left\| \Delta \right\|_F \sqrt{\frac{r^2 \log^6 d}{nd^{3/2}}} + \frac{R^2 \log^2 (d^{3/2} \sqrt{n})/\delta}{\varepsilon n} \right)

= \frac{1 + 2\varepsilon}{n} \left\| \chi_n(\Delta) \right\|_2^2 + O \left( R \left\| \Sigma^{1/2} \Delta \right\|_F \sqrt{\frac{r^2 d^{3/2} \log^6 d}{n}} + \frac{R^2 \log^2 (d^{3/2} \sqrt{n})/\delta}{\varepsilon n} \right).
\]

Using the AM-GM inequality, this is further upper bounded by

\[
\left\| \Sigma^{1/2} \Delta \right\|_F^2 \leq \frac{1 + 2\varepsilon}{n} \left\| \chi_n(\Delta) \right\|_2^2 + O \left( R^2 \frac{r^2 d^{3/2} \log^6 d}{n} + \frac{R^2 \log^2 (d^{3/2} \sqrt{n})/\delta}{\varepsilon n} \right).
\]

Rearranging, this is equivalent to

\[
(1 - \varepsilon) \left\| \Sigma^{1/2} \Delta \right\|_F^2 \leq \frac{1 + 2\varepsilon}{n} \left\| \chi_n(\Delta) \right\|_2^2 + O \left( R^2 \frac{r^2 d^{3/2} \log^6 d}{n} + \frac{R^2 \log^2 (d^{3/2} \sqrt{n})/\delta}{\varepsilon n} \right),
\]

and since \( \varepsilon < 1/2 \) this implies

\[
\left\| \Sigma^{1/2} \Delta \right\|_F^2 \leq \frac{(1 + 2\varepsilon)^2}{n} \left\| \chi_n(\Delta) \right\|_2^2 + O \left( R^2 \frac{r^2 d^{3/2} \log^6 d}{n} + \frac{R^2 \log^2 (d^{3/2} \sqrt{n})/\delta}{\varepsilon n} \right).
\]

35
Making the somewhat arbitrary choice of $\varepsilon = 1/100$, and simplifying the right-hand-side, we get
\[
\left\| \Sigma^{1/2} \Delta \right\|_F^2 \leq \frac{1.1}{n} \left\| X_n(\Delta) \right\|_2^2 + O\left( R^2 r^2 d^{3/2} \log^6 d + \log(1/\delta) \right).
\]
So we can take $c = 1.1$ and $\gamma_n = O\left( R^2 r^2 d^{3/2} \log^6 d + \log(1/\delta) \right)$. Finally, we establish Property 1. Lemma 17 establishes that with probability at least $1 - \delta$,
\[
\left\| \sum_{t=1}^n \xi_t X_t - \mathbb{E} [\xi X] \right\|_{\tilde{m}_4} \leq 4 \mathbb{E} \mathbb{E}' \left\| \sum_{t=1}^n \epsilon_t (T^*, X_t) - Y_t \right\|_{\tilde{m}_4} + \sqrt{2 \sigma^2 n \log(1/\delta)} + 4 R \log(1/\delta),
\]
where $\sigma^2 = \sup_{|T|_{m_{4,4}} \leq 1} \mathbb{E} [(T, X)^2 (T^*, X_t) - Y_t^2] \leq \sup_{|T|_{m_{4,4}} \leq 1} 4 \mathbb{E} [|T|^2] \leq \frac{4 R^2}{d^2}$. Using Lemma 23 with the standard in-expectation Lipschitz contraction lemma (e.g., Ledoux and Talagrand, (1991)), we have
\[
4 \mathbb{E} \mathbb{E}' \left\| \sum_{t=1}^n \epsilon_t (T^*, X_t) - Y_t \right\|_{\tilde{m}_4} \leq 8 R \mathbb{E} \mathbb{E}' \left\| \sum_{t=1}^n \epsilon_t X_t \right\|_{\tilde{m}_4} \leq O\left( R \sqrt{\frac{n \log^6 d}{d^{3/2}}} + R \sqrt{\log^3 d} \right).
\]
Since $n = \Omega(d^{3/2})$, these bounds together imply that we can take $M = O\left( R \sqrt{\log^6 d + \log(1/\delta)} \right)$. Theorem 20 now implies the claimed result.

**Proof of Theorem 7** As in the tensor completion case, we prove the theorem by appealing to Theorem 20 using norms $\| \cdot \| = \| \cdot \|_{\tilde{m}_4}$ and $\| \cdot \|_* = \| \cdot \|_{\tilde{m}_{4,4}}$.

Let $T^*$ be an arbitrary orthogonal tensor with $\| T^* \|_{\tilde{m}_{4,4}} = \tau$, $R(T^*) = R$, and $r(T^*) = r$. Theorem 4 implies that all $\Delta \in \mathcal{T} - T^*$ satisfy
\[
\| \Delta \|_{\tilde{m}_{4,4}} \leq 68 r \cdot \| \Delta \|_F \quad (43)
\]
and thus, since $\Sigma = I$, we may take $\kappa = 68 r$ to establish Property 3. To establish Property 2 we appeal to Theorem 25. This implies that for any $\varepsilon < \sqrt{2/\pi}$, with probability at least $1 - 2 e^{-\frac{\varepsilon^2}{2}}/(1 - e^{-\varepsilon^2/2})$,
\[
\frac{1}{\sqrt{n}} \left\| X_n(\Delta) \right\|_2 \geq (\sqrt{2/\pi} - \varepsilon) \| \Delta \|_F - \frac{C d^{3/4} \log^{1/4} d}{\sqrt{n}} \cdot \| \Delta \|_{\tilde{m}_{4,4}},
\]
where $C > 0$ is some absolute constant. Using equation (43), we have that for all $\Delta \in \mathcal{T} - T^*$, conditioned on the event above,
\[
\frac{1}{\sqrt{n}} \left\| X_n(\Delta) \right\|_2 \geq (\sqrt{2/\pi} - \varepsilon) \| \Delta \|_F - \frac{C' r d^{3/4} \log^{1/4} d}{\sqrt{n}} \cdot \| \Delta \|_F.
\]
This means that when $n = \Omega(r^2 (T^*) d^{3/2} \log^{1/2} d/\varepsilon)$, we have
\[
\| \Delta \|_F^2 \leq \frac{1}{(\sqrt{2/\pi} - 2\varepsilon)^2} \frac{1}{n} \| X_n(\Delta) \|_2^2.
\]
It suffices to set $\varepsilon = 1/40$ to get
\[
\|\Delta\|_F^2 \leq \frac{1.8}{n} \|X_n(\Delta)\|_2^2.
\]

So, simplifying, Property 2 is satisfied with $c = 1.8$ and $\gamma_n = 0$ with probability at least $1 - e^{-\frac{n}{64}}$ when $n = \Omega(r^2(T^*)^{d^3/2} \log^{1/2} d)$. To establish Property 1 we use Lemma 27, but some care needs to be taken to establish that this applies with high probability. Pick a constant $\tau \geq 0$, and observe via Lemma 28 that
\[
\mathbb{P}\left(\|X\|_F \geq d^{3/2} + \tau\right) \leq 2e^{-\frac{\tau^2}{3}}.
\]

Define a truncated sequence $X_t' = X_t1\{\|X_t\|_F \leq d^{3/2} + \tau\}$, and set $\delta_0 = 2ne^{-\frac{\tau^2}{2}}$. Observe that with probability at least $1 - \delta_0$, $X_t = X_t'$ for all $t$, and so
\[
\left\|\sum_{t=1}^{\tilde{n}} \xi_t X_t - \mathbb{E} [\xi X] \right\|_{\tilde{m}_{ij}^4} \leq \left\|\sum_{t=1}^{\tilde{n}} \xi_t X_t' - \mathbb{E} [\xi X'] \right\|_{\tilde{m}_{ij}^4}.
\]

We apply Lemma 17 to the truncated complexity, which implies that with probability at least $1 - (\delta + \delta_0)$,
\[
\left\|\sum_{t=1}^{\tilde{n}} \xi_t X_t' - \mathbb{E} [\xi X'] \right\|_{\tilde{m}_{ij}^4} \\
\leq 4 \mathbb{E}_{X_{1:n}} \mathbb{E}_\varepsilon \left\|\sum_{t=1}^{\tilde{n}} \varepsilon_t (\langle T^*, X_t \rangle - Y_t) X_t' \right\|_{\tilde{m}_{ij}^4} + O\left(\sqrt{\frac{1}{2} \log(1/\delta)} + R(d^{3/2} + \tau) \cdot \log(1/\delta)\right),
\]

where $\sigma^2 = \sup_{\|T\|_{\tilde{m}_{ij}^4} \leq 1} \mathbb{E} \left[\langle T, X' \rangle^2 - \langle T^*, X_t \rangle^2 \right] \leq \sup_{\|T\|_{\tilde{m}_{ij}^4} \leq 1} \|T\|_F^2 \cdot 4R^2 \leq 4R^2$. We now apply Lipschitz contraction to bound the in-expectation Rademacher complexity as
\[
4 \mathbb{E}_{X_{1:n}} \mathbb{E}_\varepsilon \left\|\sum_{t=1}^{\tilde{n}} \varepsilon_t (\langle T^*, X_t \rangle - Y_t) X_t' \right\|_{\tilde{m}_{ij}^4} \leq 8R \mathbb{E}_{X_{1:n}} \mathbb{E}_\varepsilon \left\|\sum_{t=1}^{\tilde{n}} \varepsilon_t X_t' \right\|_{\tilde{m}_{ij}^4}.
\]

Integrating out the tail, we can bound the error due to truncation as
\[
\mathbb{E}_{X_{1:n}} \mathbb{E}_\varepsilon \left\|\sum_{t=1}^{\tilde{n}} \varepsilon_t X_t' \right\|_{\tilde{m}_{ij}^4} \leq \mathbb{E}_{X_{1:n}} \mathbb{E}_\varepsilon \left\|\sum_{t=1}^{\tilde{n}} \varepsilon_t X_t \right\|_{\tilde{m}_{ij}^4} + n \cdot \int_{d^{3/2} + \tau}^{\infty} \mathbb{P}(\|X\|_F > t) dt.
\]

We have
\[
n \cdot \int_{d^{3/2} + \tau}^{\infty} \mathbb{P}(\|X\|_F > t) dt = n \cdot \int_{d^{3/2} + \tau}^{\infty} e^{-\frac{t^2}{2}} dt \leq n \cdot \exp\left(-C(d^3 + \tau^2)\right),
\]

for some absolute constant $C$. We pick $\tau = (\sqrt{(d^{3/2}n + \log(1/\delta))/C})$, so that the quantity above is $o(1)$ and $\delta_0 \leq \delta$. Lastly, since $X_{1:n}$ are gaussian, Lemma 27 implies
\[
\mathbb{E}_{X_{1:n}} \mathbb{E}_\varepsilon \left\|\sum_{t=1}^{\tilde{n}} \varepsilon_t X_t \right\|_{\tilde{m}_{ij}^4} = \sqrt{n} \cdot \mathbb{E}_X \|X\|_{\tilde{m}_{ij}^4} \leq O(\sqrt{d^{3/4} \log^{1/4} d \sqrt{n}}).
When $n \geq d^{3/2}$, we have $d^{3/2} \leq d^{3/4}\sqrt{n}$, and so we conclude that with probability at least $1 - \delta$, 
\[
\left\| \frac{1}{n} \sum_{i=1}^{n} \xi_{i}X_{i} - \mathbb{E}[\xi X] \right\|_{\mathbb{F}_{1}} \leq O(Rd^{3/4}\log^{3/2}(d/\delta)\sqrt{n}),
\]
so $M = O(Rd^{3/4}\log^{3/2}(d/\delta))$.

\section*{Appendix D. Proofs from Section 4}

\textbf{Proof of Theorem 9} Let $\varepsilon > 0$ be fixed and let $m = d^{3/2-\varepsilon/2}$.

We will generate an instance of tensor completion from the 3-XOR instance by treating the indices $(i, j, k)$ in each clause as an observed entry. Precisely, for each clause, we set the corresponding $X = e_{i} \otimes e_{j} \otimes e_{k}$, and the corresponding $Y = z_{i,j,k}$. The induced risk for each tensor $T$ in this setting is simply $L_{D_{Y}}(T) = \frac{1}{m} \sum_{i,j,k} (T_{i,j,k} - Z_{i,j,k})^{2}$, where the tensor $Z \in \{\pm 1\}^{d}$ is defined as follows: in the random case, the entries of $Z$ are selected uniformly at random from $\{\pm 1\}$; in the planted case with planted assignment $a \in \{\pm 1\}^{d}$, we start off with the tensor $a \otimes a \otimes a$; then get $Z$ by flipping each coordinate independently with probability $\eta$.

If we set $n = O(m)$, then we are guaranteed to receive $\tilde{O}(m)$ unique clauses under with-replacement sampling, since there are at most logarithmically many repeats with inverse polynomial probability for $m = o(d^{3/2})$. With this observation, it is clear that any algorithm that takes as input the examples $S$ and outputs “Planted” with probability at least $1 - o(1)$ when $Z$ comes from the planted distribution and “Random” with probability at least $1 - o(1)$ when $Z$ is from the random distribution is a successful distinguisher for the planted 3-XOR problem with $\tilde{O}(m)$ clauses. Going forward we assume $n = O(d)$, since this is the interesting regime for distinguishing and refutation.

Now, suppose we have an algorithm that enjoys the excess risk bound (19) for any $n$, and let $\hat{T}_{S}$ denote its output given input dataset $S$. Since $Z \in \{\pm 1\}^{d}$, we can take $\|\hat{T}_{S}\|_{\mathbb{F}} \leq 1$ without loss of generality. We will turn $\hat{T}_{S}$ into a successful distinguishing algorithm for the planted 3-XOR problem as follows: Define $\gamma = 1 - 4\eta$, and note the assumption that $\eta < 1/4$ implies that $\gamma > 0$. Split the sample set $S$ into halves $S_{1}$ and $S_{2}$, and let $\hat{T}_{S_{1}}$ and $\hat{T}_{S_{2}}$ denote the empirical risk on the respective halves. Let $\tilde{T}_{S_{1}}$ be the output of the assumed algorithm on $S_{1}$. If $\tilde{T}_{S_{2}}(\hat{T}_{S_{1}}) \geq 1 - \gamma^{4}/2$ return “Random”, else return “Planted”.

We will show that the algorithm succeeds in both the planted and random case by analyzing the value of $\tilde{T}_{S_{2}}(\hat{T}_{S_{1}})$ in each case.

\textbf{Random case.} Let $S'_{2}$ be the result of removing all entries that appear in $S_{1}$ from $S_{2}$. The number of repeats is at most $O(n^{2}/d^{3} + \log(1/\delta))$ with probability at least $1 - \delta$. It follows that with probability at least, say, $1 - O(d^{-1})$,
\[
\frac{|S'_{2}|}{|S_{2}|} \geq 1 - o(1), \quad \text{and} \quad \tilde{T}_{S_{2}}(\hat{T}_{S_{1}}) \geq \tilde{T}_{S'_{2}}(\hat{T}_{S_{1}}) - o(1).
\]

Observe that for every remaining example $X = e_{i} \otimes e_{j} \otimes e_{k}$ in $S'_{2}$, the value of $\tilde{T}_{S_{1}}$ is statistically independent of the value of $Z_{i,j,k}$. Abbreviating $\hat{T}_{S_{1}}$ to $\hat{T}$, this means that we have
\[
\mathbb{E}[Z_{X}(\tilde{T}_{S_{1}})] = \frac{1}{|S'_{2}|} \sum_{X = e_{i} \otimes e_{j} \otimes e_{k} \in S'_{2}} \mathbb{E}[Z_{i,j,k}(\hat{T}_{i,j,k} - Z_{i,j,k})^{2}] = \frac{1}{|S'_{2}|} \sum_{X = e_{i} \otimes e_{j} \otimes e_{k} \in S'_{2}} (\hat{T}_{i,j,k})^{2} + 1 \geq 1.
\]
Condition on $S'_2$. Since $\tilde{T}_{S_1}$ and $Z$ have bounded entries, it follows from Hoeffding’s inequality that with probability at least $1 - \delta$ over the choice of $Z$,

$$\tilde{L}_{S'_2} (\tilde{T}_{S_1}) \geq \mathbb{E}_Z [\tilde{L}_{S'_2} (\tilde{T}_{S_1})] - c \sqrt{\frac{\log(1/\delta)}{|S'_2|}},$$

for absolute constant $c > 0$. Since $|S'_2| = \tilde{\Omega}(m)$, we can use the AM-GM inequality to conclude with probability at least $1 - O(d^{-1})$,

$$\tilde{L}_{S_2} (\tilde{T}_{S_1}) \geq \tilde{L}_{S'_2} (\tilde{T}_{S_1}) - o(1).$$

It follows by union bound that

$$\tilde{L}_{S_2} (\tilde{T}_{S_1}) \geq 1 - o(1) \geq 1 - \gamma^4/100 - o(1).$$

This proves that our strategy will indeed return “Random” for random instances.

**Planted case.** For any $Z$, the assumed excess risk bound implies that for any rank-1 tensor $T^*$ with bounded entries, we have

$$L_{D_2} (\tilde{T}_{S_1}) - L_{D_2} (T^*) \leq o(1),$$

when $n = \omega(d^{3/2-\epsilon})$. We choose $T^* = a \otimes^3$, where $a \in \{\pm 1\}^d$ is the planted assignment. Taking expectation over the flips in $Z$, we have

$$\mathbb{E}_Z [L_{D_2} (a \otimes^3)] = \mathbb{E}_Z \left[ \frac{4}{d^3} \sum_{i,j,k} 1\{Z_{ijk} \text{ flipped}\} \right] = 4\eta.$$

Applying Bernstein’s inequality, we have that with probability at least $1 - O(d^{-1})$ over the choice of $Z$,

$$L_{D_2} (a \otimes^3) \leq 4(1 + \gamma)\eta + o(1) = 1 - \gamma^2 + o(1).$$

Finally, we use Bernstein once more to show that the empirical loss for $S_2$ converges to the population loss, leveraging that $S_2$ is an independent hold-out set for $\tilde{T}_{S_1}$. We have that for any choice of $S_1$, with probability at least $1 - \delta$ over the draw of $S_2$,

$$\tilde{L}_{S_2} (\tilde{T}_{S_1}) \leq L_{D_2} (\tilde{T}_{S_1}) + c_1 \sqrt{\frac{L_{D_2} (\tilde{T}_{S_1}) \log(1/\delta)}{n}} + c_2 \frac{\log(1/\delta)}{n},$$

where $c_1$ and $c_2$ are absolute constants. Applying AM-GM, this implies that $\tilde{L}_{S_2} (\tilde{T}_{S_1}) \leq (1 + \gamma^2) L_{D_2} (\tilde{T}_{S_1}) + o(1)$ with probability at least $1 - O(d^{-1})$, and so by combining this with the excess risk bound and the bound on $L_{D_2} (a \otimes^3)$, we have

$$\tilde{L}_{S_2} (\tilde{T}_{S_1}) \leq 1 - \gamma^4 + o(1).$$