Near-Isometric Properties of Kronecker-Structured Random Tensor Embeddings

Qijia Jiang
Lawrence Berkeley National Laboratory
qjiang@lbl.gov

Abstract

We give uniform concentration inequality for random tensors acting on rank-1 Kronecker structured signals, which parallels a Gordon-type inequality for this class of tensor structured data. Two variants of the random embedding are considered, where the embedding dimension depends on explicit quantities characterizing the complexity of the signal. As applications of the tools developed herein, we illustrate with examples from signal recovery and optimization.

1 Introduction

It is hardly an overstatement to proclaim that underpins most of the analysis for high-dimensional statistics and structured signal recovery is the heavy hammer made possible by the machinery of Gaussian process, and in particular Gordon-type inequality that gives tight characterization of the suprema of the empirical process with geometric properties of the underlying index set. In this paper, we put Kronecker-structured random tensors into scrutiny and ask for analog of Gordon's inequality for correspondingly tensor-structured signals. We embark with a brief reminder of the classics.

1.1 Gordon's inequality for Gaussian random matrix

For signal $u \in T \subset \mathbb{R}^n$ a vector, it is known for $S \in \mathbb{R}^{m \times n}$ random i.i.d standard Gaussian matrix,

$$\mathbb{E}[\min_{u \in T} \|Su\|] \geq a_m - w(T) \quad \text{and} \quad \mathbb{E}[\max_{u \in T} \|Su\|] \leq a_m + w(T)$$

for $a_m = \mathbb{E}[\|g_m\|] \approx \sqrt{m}$ where $g_m \sim \mathcal{N}(0, I_m)$ and $w(T) = \mathbb{E}[\max_{x \in T} g^\top x]$ the Gaussian width for set $T \subset \mathbb{S}^{n-1}$, a subset of the unit sphere. This statement hinges on the Gaussian min-max comparison lemma (i.e., Fernique-Slepian theorem), which implies for g, h independent standard Gaussian vectors,

$$\mathbb{E}_{g,h}[\min_{u \in T} \max_{v \in \mathbb{S}^{m-1}} g^\top v + h^\top u] \leq \mathbb{E}_S[\min_{u \in T} \max_{v \in \mathbb{S}^{m-1}} v^\top Su]. \tag{1}$$

This trades the quadratic form for a more innocuous separable process, from which one can see that the LHS evaluates to the first part of the previous display. The other side is essentially similar. For this expectation bound to justify the attention it deserves, one needs to recognize that $\min_{u \in T} \|Su\|$ (analogously for $\max$) is a Lipschitz function in the Gaussian random matrix S, from which (dimension-free) concentration inequality, alongside the bound on the expectation derived above, conspire to deliver a uniform concentration bound as stated below.

Theorem 1 (Gordon's escape through the mesh [12])**.** *For all $u \in T \subset \mathbb{R}^n$, where T is a (not necessarily convex) cone, with probability at least $1 - 2\exp(-\delta^2/2)$ for S entrywise i.i.d standard Gaussian,*

$$(1 - \epsilon)\|u\| \leq \frac{1}{a_m} \|Su\| \leq (1 + \epsilon)\|u\|$$

when $m \geq \frac{(w(T)+\delta)^2}{\epsilon^2}$.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Later work of CGMT [17] showed that the reduction of (1) is essentially tight for convex sets, which has surprising consequences for analyzing the risk of various statistical estimators in a high-dimensional asymptotic regime. This elegant analysis, nevertheless, cannot be carried out beyond the Gaussian case due to the lack of comparison lemma (1) (even for subgaussian), but gives that for example, the extreme singular values of a Gaussian random matrix $1/\sqrt{m} \cdot S$ scales as $1 \pm \sqrt{n/m}$ by picking $T = \mathbb{S}^{n-1}$. It also recovers the familiar Johnson-Lindenstrauss lemma for distance-preserving random projection of finite point set when $w = \sqrt{\log(|T|)}$, where $|T|$ is the cardinality of the set.

Seemingly a natural obsession for probabilists for its mathematical allure, results of this flavor have found unexpectedly number of applications across many areas in numerical linear algebra, signal processing, theoretical computer science, among others. Such uniform convergence result is frequently encountered for deriving tight sample complexity bounds for recovery problems, where the problem boils down to characterizing the probability that a random subspace (i.e., null space of Gaussian measurement matrix) distributed uniformly misses the tangent cone of a regularizer. Nonconvex gradient-based optimization heavily leans on these tools for characterizing restricted singular value for deriving convergence with Empirical Risk Minimization. Sketching-based least-squares optimization $\min_x \|SAx - Sb\|_2^2$ also crucially rely on such results, where $w(U \cap \mathbb{S}^{n-1}) = \sqrt{\dim(U)}$ for $U = \text{colspan}([A, b])$ for the subspace embedding property.

1.2 Contributions

We aim to generalize Gordon's uniform concentration result for tensor-structured signal $x = u^1 \otimes \cdots \otimes u^d$ while insisting on efficient computation of the embedding operation. More concretely, we consider Kronecker-structured random rank-1 tensor, which when acting on rank-1 tensor-structured signals, can be performed without explicitly forming the $n \times n \times \cdots \times n$ tensor since it can be done factor-by-factor effortlessly. Formally we set out our roadmap to address the following questions:

1. For (1) structured and fast tensored embedding (e.g., Tensor-SRHT as defined in Definition 1 below); and (2) Tensor-Subgaussian introduced in Definition 2, what is dictated from the embedding dimension m for the following guarantee to hold w.h.p

$$\left| \frac{1}{m} \sum_{i=1}^m \prod_{j=1}^d \langle v_i^j, u^j \rangle^2 - \|x\|^2 \right| \leq \max(\epsilon, \epsilon^2) \cdot \|x\|^2, \qquad (2)$$

   for all $x = u^1 \otimes \cdots \otimes u^d \in T^1 \times \cdots \times T^d$ (Cartesian product of d not necessarily convex cones), as a function of the geometric properties of the *individual* sets $T^1, \cdots, T^d$. This is a generalization of the Restricted Isometry Property (RIP) to (1) higher order tensored signals; (2) general cones beyond sparsity. Both sketches above are row-wise tensored and take the form $S_i = \text{vec}(v_i^1 \otimes \cdots \otimes v_i^d)$ for each row $i \in [m]$. We are interested in the regime $m \ll n^d$ and instantiate the embedding result for this sketch from Section 4 to bound the restricted singular value as required by a tensor signal recovery problem in Section 6.1.

2. To improve the dependence of m on the degree d (while maintaining computation efficiency), we consider a recursive embedding in Section 5 which repeatedly calls a degree-2 Tensor-SRHT $S^j \in \mathbb{R}^{m \times nm}$ as a subroutine as follows: $S(u^1 \otimes u^2 \otimes u^3 \cdots) := S^1(u^1 \otimes S^2(u^2 \otimes S^3(u^3 \otimes \cdots)))$. Similar uniform concentration is derived on the scaling of m with geometric properties of the individual sets for this alternative embedding, which is in turn called upon to speed up solving for optimization problem in Section 6.2.

3. Our technique is based on generic chaining - we include comparison with results one would get from more naive method in Section 3 and part with some discussions of lower bound on the embedding dimension in Section F and numerical results in Section 7.

We pause to emphasize it is the correlation in the tensor structure that introduces difficulty for tight concentration – result for general random tensor with i.i.d entries is less challenging to obtain, but at the same time less efficient to apply.

Definition 1 (Tensor-SRHT). *A random matrix constructed as $S = \frac{1}{\sqrt{m}} P_1 H_n D_1 \circ \cdots \circ P_d H_n D_d \in$*

$\mathbb{R}^{m \times n^d}$ *is called a Tensor-SRHT (Subsampled Randomized Hadamard Transform), if when acting on a rank-1 degree-d tensor, takes the form $S(u^1 \otimes \cdots \otimes u^d) = \frac{1}{\sqrt{m}} P' H_{n^d} D' vec(u^1 \otimes \cdots \otimes u^d) :=$*

2

$\frac{1}{\sqrt{m}} P_1 H_n D_1 u^1 \odot \cdots \odot P_d H_n D_d u^d$, where D' is a $n^d \times n^d$ diagonal matrix with entries $D_1 \otimes \cdots \otimes D_d$ (i.e., tensor product of independent Rademachers) and P' is a $m \times n^d$ subsampling matrix with a single 1 in each (independent) row and $H_{n^d} = H_n \otimes \cdots \otimes H_n$ where n is a power of 2 is the Hadamard matrix of size $n^d \times n^d$. Here $\odot$ denotes Hadamard product and $\circ$ denotes the transposed Khatri-Rao product. Moreover, such embedding can be carried out in time $\mathcal{O}(d(n \log n + m))$.

Definition 2 (Tensor-Subgaussian). *We call $S \in \mathbb{R}^{m \times n^d}$ a Tensor-Subgaussian embedding if every row $S_i = vec(v_i^1 \otimes \cdots \otimes v_i^d)$ is constructed where each factor is an independent σ-subgaussian isotropic random vector, i.e., (1) $\mathbb{E}[\langle v_i^j, u^j \rangle^2] = \|u^j\|_2^2$; (2) $\mathbb{E}[|\langle v_i^j, u^j \rangle|^p]^{1/p} \le \sqrt{\sigma p}\|u^j\|_2$ for all $p \ge 2$, $i \in [m], j \in [d]$ and any $u^j \in \mathbb{R}^n$.*

2 Related Work

In the case of vector-valued signal ($d = 1$), embedding analysis for infinite sets using structured matrices requires ingenuity and is significantly more involved in general. Notable extensions include [5, 10]. The work of [15] offered a unifying theme - the important message behind is that one can have a reduction from RIP based result to Gordon-type inequality by invoking it at different sparsity levels with various distortions à la Talagrand's multi-resolution generic chaining. An orthogonal thread for generalizing to heavier-tail distribution involves small-ball technique which gives an one-sided bound for nonnegative empirical process - such undertaking is present in e.g., [18].

Previous work on tensor concentration are mostly concerned with operator norm bounds for symmetric subgaussian and/or log-concave (potentially non-isotropic) factors [11, 23], where for symmetric forms $\|S\|_{op}$ is maximized by a single vector $u \in \mathbb{S}^{n-1}$ therefore for this we only need to content ourselves with a single index set and look at moment deviations of type: $\sup_{u \in \mathbb{S}^{n-1}} \left| \frac{1}{m} \sum_{i=1}^m \langle S_i, u \rangle^d - \mathbb{E}[\langle S, u \rangle^d] \right|$, an arguably simpler task. Indeed, a multi-resolution approach is not strictly beneficial here compared to more elementary arguments [11].

The case of non-symmetric factors warrant more care. Both [21, 4] studied pointwise tail bound of the form $\mathbb{P}(|\|Sx\|_2 - \|S\|_F| \ge t)$ for $S \in \mathbb{R}^{m \times n^d}$ a linear mapping, $x = u^1 \otimes \cdots \otimes u^d \in \mathbb{R}^{n^d}$, where u^k's are independent factors each with independent, mean 0, unit variance, subgaussian coordinates – this can in turn be used for deriving a high-probability lower bound on $\sigma_{\min}(X)$ for the $n^d \times m$ random matrix X where each column is formed by the aforementioned tensor x. Uniform deviation for general sets on tensors can be viewed as a special instance of 2nd-order chaos with mixed tails [16]. For example in the case of processes with subgaussian-subexponential increments (as is the case when $d = 2$ for Tensor-Subgaussian embedding in Definition 2), i.e., $\forall u > 0, s, t \in T$,

$$\mathbb{P}(\|X_t - X_s\| \ge \sqrt{u}d_2(t, s) + ud_1(t, s)) \le 2e^{-u},$$

the result of [9] gave a uniform deviation for $\sup_{t \in T} \|X_t\|$ as a combination of $\gamma_2(T, d_2)$ and $\gamma_1(T, d_1)$ but crucially these quantities are tied to the metric complexity of the *product index set* $T := T^1 \times T^2$ – something that is hard to compute by and large. Various works also study finite set embedding for Kronecker-structured sketches, some of which we will leverage for our results and will mention them in later contexts.

3 Discrete JL and a Single-scale Approach

At the heart of the following result is a generalized Khinchine inequality [2] which says if $\mathbb{E}[|\langle v^k, a \rangle|^p]^{1/p} \le C_p \|a\|_2$ for any vector $a \in \mathbb{R}^n$ and all independent $\{v^k\}_{k=1}^d$, then $\mathbb{E}[|\langle v^1 \otimes \cdots \otimes v^d, a \rangle|^p]^{1/p} \le C_p^d \|a\|_2$ for any (not necessarily rank-1) tensor $a \in \mathbb{R}^{n^d}$. This is closely related to an earlier result from [13] on the concentration of Gaussian chaos but generalized to broader class. Such moment control is only a hop away from tail bounds using standard arguments. We establish the finite-set embedding property for the row-wise-tensored embedding matrices below, building upon previous work. This serves as the stepping stone for the embedding of general sets.

Lemma 1 (Discrete-JL property for Tensor-SRHT and Tensor-Subgaussian). *For a set of cardinality p that the rank-1 tensor $x \in \mathbb{R}^{n^d}$ belongs, with probability at least $1 - e^{-\eta}$ for any $\eta > 0$ and $\epsilon > 0$, Tensor-SRHT as defined in Definition 1 satisfies $|\|Sx\|_2^2 - \|x\|_2^2| \le \max(\epsilon, \epsilon^2)\|x\|_2^2$ simultaneously for all p points provided $m = \mathcal{O}(C^d \epsilon^{-2}(\log^d(p) + (1 + \eta)^d))$. The same guarantee holds for*

3

Tensor-Subgaussian in Definition 2 with $m = \mathcal{O}(C^d \sigma^{2d} \epsilon^{-2}(\log^d(p) + (1 + \eta)^d))$ *for some universal constant C.*

Remark. Close inspection of the proof for Theorem 3 in [2] in fact uncovers that the discrete JL property above holds for more general class of SORS (Subsampled Orthogonal Random Sign) constructions for which $H^*H = n \cdot I_n$ and $\max_{i,j \in [n]} |H_{ij}| \leq c$. In the case $d = 1$, it also recovers the classical Johnson–Lindenstrauss lemma.

Without taking the multi-scale route, in the case $d = 1$, to guarantee ϵ-distortion over a continuous set, one needs to roughly speaking build a Δ-net for $x \in \mathbb{R}^n$ for $\Delta \lesssim \epsilon \cdot \sqrt{m/n}$ therefore the sample complexity one gets with a single-scale approach will scale as

$$m \gtrsim \frac{\log(|\mathcal{N}^\Delta|)}{\epsilon^2} \gtrsim \frac{nw^2(T)}{m\epsilon^4} \Rightarrow m \gtrsim \frac{\sqrt{n}w(T)}{\epsilon^2},$$

where we used Sudakov's minorization for bounding the size of the covering with Gaussian width of the set and the JL Lemma for SRHT/Subgaussian matrices for the first transition. This back-of-the-envelope calculation showcases that uniform covering is far from optimal, since in general it could be the case $w(T) \ll \sqrt{n}$ for $T \subset \mathbb{S}^{n-1}$ a subset of the unit sphere – and this insight is precisely the reason that motivated [15] to consider a multi-scale approximation that can establish the $m \asymp w^2(T)/\epsilon^2$ guarantee for wider classes of random ensembles beyond the Gaussian case in Theorem 1. To put things in perspective with later sections, we work out the sample complexity required from a naive uniform discretization below.

Lemma 2 (Δ-net Covering). *Using Tensor-SRHT, with a uniformly constructed Δ-net covering of the tensor, one requires $m = \mathcal{O}(\epsilon^{-2} \cdot n^{\frac{d^2}{1+d}} (\sum_{i=1}^d \gamma_2^2(T^i))^{\frac{d}{1+d}})$ for (2) to hold.*

Even in the prosaic case of Gaussian process indexed by ellipsoid and/or ℓ_1 ball, it is a well-known and disappointing fact that arguments based on union bound / Dudley integral don't give the optimal bound, whereas method based on generic chaining does [16], which we turn to next.

4 A Multi-scale Approach: Row-wise Tensored Embedding

One viable approach is to apply the result of [15] naively to $\text{vec}(u^1 \otimes \cdots \otimes u^d)$ without taking into consideration the Kronecker structure, but this is somewhat of a futile endeavor if one takes any interest in downstream applications of such bounds. In fact, this was also the impetus for Mendelson's work on product empirical processes [14] – it is generally hard to handle geometric properties of process indexed by product classes. We will instead derive results with an eye towards bounds involving *decoupled* geometric complexity measure for each factor that lends itself to explicit computations – this necessarily calls for a more intricate chaining argument. Another possibility is to use a contraction inequality à la Ledoux-Talagrand if the random factors $\{v_i^j\}_{j=1}^d$ come from bounded class but this will be crude in almost all cases.

Our agenda is to leverage the results on finite set embedding from the previous section, wrap them inside of a chaining argument by exploiting coverings at multiple scales with different distortions/probability tradeoff so each level of approximation demands roughly the same embedding dimension (as we will see, the final m depends on the maximum required across all resolutions).

4.1 Preliminaries

Throughout the paper, we use $\lesssim, \asymp, \gtrsim$ to hide absolute constants. To measure the size of the set $T^i \subset \mathbb{R}^n$, we use Gaussian width defined as for $g \sim \mathcal{N}(0, I_n)$,

$$w(T^i) = \mathbb{E}\left[\sup_{u \in T^i} g^\top u\right].$$

In our context, we define the γ_2^* functional as

$$\gamma_2^*(T^i) := \inf_{\{T_l^i\}} \sup_{u^i \in T^i} \sum_{l=0}^\infty 2^{l/2} \text{dist}(u^i, T_l^i)$$

4

where the infimum is taken over all sequences of nets $\{T_l^i\}_l$ with cardinality $|T_l^i| \leq 2^{2^l} =: N_l \; \forall i \in [d]$ and $|T_0^i| = 1 =: N_0$. For Gaussian process with canonical metric (i.e., Euclidean norm) on T^i, the expected supremum is completely characterized by $\gamma_2^*(T)$, i.e.,

$$\gamma_2^*(T^i) \asymp w(T^i)$$

where the upper bound is due to Fernique and the (much deeper, specific-to-gaussian-process) lower bound is due to Talagrand's majorizing measure theorem. A more general definition working with admissible sequences defines

$$\gamma_2(T^i) := \inf_{\{\mathcal{A}_l^i\}} \sup_{u^i \in T^i} \sum_{l=0}^{\infty} 2^{l/2} \mathrm{diam}(\mathcal{A}_l^i(u^i))$$

where the infimum is taken over all admissible sequences (i.e., increasing sequence of partitions of T^i with $|\mathcal{A}_l^i| \leq N_l$ for all $l \geq 0$) and $\mathcal{A}_l^i(u^i)$ denotes the (unique) element of $\mathcal{A}_l^i$ that contains u^i. It is not hard to see that by picking one point arbitrarily from each element of the partition, one can build a net which implies that we always have $\gamma_2^*(T^i) \leq \gamma_2(T^i)$. In fact, the work of [20] shows that these two quantities are always of the same order.

It is also an immediate consequence that for an optimal admissible sequence $\{\bar{\mathcal{A}}_l^i\}_l$, picking $\{\bar{T}_l^i\}_l$ as a sequence of nets with cardinally $|\bar{T}_l^i| \leq N_l$ constructed by choosing the center point in every element of the partition set $\{\bar{\mathcal{A}}_l^i\}_l$, we have for all $u^i \in T^i$, $i \in [d]$,

$$\sum_{l=0}^{\infty} 2^{l/2} \mathrm{dist}(u^i, \bar{T}_l^i) \leq \inf_{\{\mathcal{A}_l^i\}} \sup_{t \in T^i} \sum_{l=0}^{\infty} 2^{l/2} \mathrm{diam}(\mathcal{A}_l^i(t)) \,. \tag{3}$$

For our results, we will find it helpful to adopt the slightly more general γ_α-functional for $\alpha > 0$:

$$\sum_{l=0}^{\infty} 2^{l/\alpha} \mathrm{dist}(u^i, \bar{T}_l^i) \leq \gamma_\alpha(T^i) := \inf_{\{\mathcal{A}_l^i\}} \sup_{u^i \in T^i} \sum_{l=0}^{\infty} 2^{l/\alpha} \mathrm{diam}(\mathcal{A}_l^i(u^i))$$

and the infimum is taken over all admissible sequences in exactly the same way as (3). It is known that for a random variable with tail decay bounded as $e^{-|x|^\alpha}$, the supremum is upper bounded by the γ_α functional [9]. Moreover, we always have the following Dudley-style metric entropy integral estimate [16] where B_2^n denotes the unit-ℓ_2 ball in $\mathbb{R}^n$:

$$\gamma_\alpha(T^i) \lesssim C_\alpha \int_0^1 \left(\log N(T^i, sB_2^n) \right)^{1/\alpha} ds \,, \tag{4}$$

but the reverse is generally not true. Here the upper limit of the integral goes up to 1 because $N(T^i, sB_2^n) = 1$ for $s \geq 1$ by simply picking $\{0\}$ as cover. Covering number on the RHS of (4) can be bounded with estimates on Gaussian width. In particular, Sudakov minorization asserts

$$\sup_{s > 0} s\sqrt{\log N(T^i, sB_2^n)} \lesssim w(T^i) \,,$$

which uses covering number at a single scale. Various alternative options exist for upper bounding the covering number, including Volumetric estimates, Maurey's empirical method etc.

Estimate (4) above has the drawback of not being explicit in constants C_α, if one is keen on explicit dependence on α, the following lemma becomes timely.

Lemma 3 (Relationship between γ_α functionals). *For $\alpha \leq 1$, if set $T^i \subset \mathbb{S}^{n-1}$ has covering number $N(T^i, sB_2^n) \leq (\frac{a}{s})^b$ for some $b \geq 2$, $a \geq 2$, then*

$$\gamma_2(T^i) \leq \gamma_\alpha(T^i) \leq \left(1 + K \cdot \log_2(b/\alpha) \cdot b/\alpha \cdot \log_2(a)\right)^{\frac{2-\alpha}{2\alpha}} \gamma_2(T^i)$$

for some absolute constant K.

4.2 Multi-resolution embedding property

Instead of going through the multi-scale RIP (followed by column sign randomization) as done in [15] we will give ourselves more wiggle room by working with a multi-scale embedding property for finite sets. Definition 3 below will be featured prominently in subsequent sections and make the successive construction of approximations less mysterious than it may otherwise seem. We will invoke it for Tensor-SRHT and Tensor-Subgaussian in this section – both taking the form where each row $S_i = \mathrm{vec}(v_i^1 \otimes \cdots \otimes v_i^d)$.

5

Definition 3 (Multi-resolution Embedding Property). *A mapping $S : \mathbb{R}^{n^d} \mapsto \mathbb{R}^m$ fulfills the (ϵ, η, α)- Multi-resolution Embedding Property if for an increasing sequence of successive coverings $\{\bar{T}_l^i\}_l$ of $T^i \subset \mathbb{S}^{n-1}$ such that $|\bar{T}_l^i| \leq 2^{2^l}$ and $|\bar{T}_0^i| = 1 \,\forall i \in [d]$ defined in (3) for tensor $x := u^1 \otimes \cdots \otimes u^d$, the following holds simultaneously for all $1 \leq l \leq L \asymp \lceil \log_2(nd) \rceil$ with probability at least $1 - \exp(-\eta)$:*

- *For all $k \in [d]$ and $l \in [L]$,*

$$
\big| \|S(u_l^1 \otimes \cdots \otimes u_l^k \otimes \cdots \otimes u_{l-1}^d) - S(u_l^1 \otimes \cdots \otimes u_{l-1}^k \otimes \cdots \otimes u_{l-1}^d)\|_2^2
$$
$$
- \|u_l^1 \otimes \cdots \otimes (u_l^k - u_{l-1}^k) \otimes \cdots \otimes u_{l-1}^d\|_F^2 \big|
$$
$$
\leq \max(2^{l/\alpha}\epsilon, 2^{2l/\alpha}\epsilon^2) \cdot \|u_l^1\|_2^2 \cdots \|u_l^k - u_{l-1}^k\|_2^2 \cdots \|u_{l-1}^d\|_2^2
$$

- *For all $k \in [d]$ and $l \in [L]$,*

$$
\big| \|S(u_l^1 \otimes \cdots \otimes u_l^k \otimes \cdots \otimes u_{l-1}^d)\|_2^2 - \|u_l^1 \otimes \cdots \otimes u_l^k \otimes \cdots \otimes u_{l-1}^d\|_F^2 \big|
$$
$$
\leq \max(2^{l/\alpha}\epsilon, 2^{2l/\alpha}\epsilon^2) \cdot \|u_l^1\|_2^2 \cdots \|u_l^k\|_2^2 \cdots \|u_{l-1}^d\|_2^2
$$

- *For all $k \in [d]$ and $l \in [L]$,*

$$
\left| \left\| S\left(u_l^1 \otimes \cdots \otimes \left(\frac{u_l^k - u_{l-1}^k}{\|u_l^k - u_{l-1}^k\|_2} \right) \otimes \cdots \otimes u_{l-1}^d \right) \pm S(u_l^1 \otimes \cdots \otimes u_{l-1}^k \otimes \cdots \otimes u_{l-1}^d) \right\|_2^2 \right.
$$
$$
\left. - \left\| u_l^1 \otimes \cdots \otimes \left(\frac{u_l^k - u_{l-1}^k}{\|u_l^k - u_{l-1}^k\|_2} \pm u_{l-1}^k \right) \otimes \cdots \otimes u_{l-1}^d \right\|_F^2 \right|
$$
$$
\leq \max(2^{l/\alpha}\epsilon, 2^{2l/\alpha}\epsilon^2) \cdot \left\| \frac{u_l^k - u_{l-1}^k}{\|u_l^k - u_{l-1}^k\|_2} \pm u_{l-1}^k \right\|_2^2 \cdot \|u_l^1\|_2^2 \cdots \|u_l^{k-1}\|_2^2 \|u_{l-1}^{k+1}\|_2^2 \cdots \|u_{l-1}^d\|_2^2
$$

where tensor Frobenius norm $\|x\|_F := \prod_{k=1}^d \|u^k\|_2$ and u_l^k is the closest point to u^k in $\{\bar{T}_l^k\}$.

For the desired accuracy $\epsilon > 0$ in the final guarantee (2), in what follows we correspondingly define a sequence of distortion levels $\epsilon_0 = \epsilon, \epsilon_1 = 2^{1/\alpha}\epsilon, \cdots, \epsilon_L = 2^{L/\alpha}\epsilon$ for $L \asymp \lceil \log_2(nd) \rceil$ levels and let $\tilde{L} = \max(0, \lfloor \alpha \log_2(1/\epsilon) \rfloor)$ such that for $l \leq \tilde{L}$, $\epsilon_l \leq 1$ therefore $\max(\epsilon_l, \epsilon_l^2) = \epsilon_l$. Additionally, we define $x = u_{L+1}^1 \otimes \cdots \otimes u_{L+1}^d$ being the finest level of approximation. Give ϵ, n, d, we will pick $L = C\lceil \log_2(nd) \rceil$ for a constant C and work under the assumption that $\tilde{L} \leq L$ in the proofs presented in Section B – the case when $\tilde{L} > L$ allows us to draw the same conclusion and is deferred to Appendix D. Here the constant C is independent from all problem parameters.

Definition 3 takes center stage in the following lemma. The trade-off of η_l, ϵ_l and p_l specified in the proof of Lemma 4 below ensures that there's no occurrence of l in the final stated m. The $\{\epsilon_l\}$ plays the role of multi-level approximation close in spirit to what the γ-functional attempts to capture. The super-exponential factor of d^d also made an appearance in earlier work on embedding of finite set using Tensor-SRHT [3].

Lemma 4 (Multi-resolution embedding property of row-wise tensored sketches). *With $m = \mathcal{O}(C^d(d^d + (1 + \eta)^d)/\epsilon^2)$, Tensor-SRHT defined in Definition 1 satisfies Definition 3 for $\alpha = 2/d$. The same property also holds for Tensor Subgaussian defined in Definition 2 for $m = \mathcal{O}(C^d\sigma^{2d}(d^d + (1 + \eta)^d)/\epsilon^2)$ and $\alpha = 2/d$.*

4.3 Embedding of general sets with row-wise tensored sketches

Now we embark on our journey for the proof of our main result on row-wise Kronecker-structured sketches where Definition 3 and Lemma 4 will reveal their power.

Theorem 2 (Gordon-type Inequality for Tensor-SRHT and Tensor-Subgaussian). *Tensor-SRHT with $m = \mathcal{O}(C^d\epsilon^{-2}(\sum_{i=1}^d \gamma_{2/d}(T^i))^2 d^d)$ satisfies uniform concentration (2). The same guarantee carries over to Tensor-Subgaussian with $m = \mathcal{O}(C^d\sigma^{2d}\epsilon^{-2}(\sum_{i=1}^d \gamma_{2/d}(T^i))^2 d^d)$.*

6

This recovers the result of [15] for $d = 1$ (ignoring poly-logs). In light of the tail bound Theorem 2.1 in [4], it is also natural that the $\gamma_{2/d}$ functional shows up.

Remark. This concentration result can also be easily converted to be on $|\|Sx\|_2 - 1|$ using basic inequality $\frac{1}{3}\min\{|a^2 - 1|, \sqrt{|a^2 - 1|}\} \leq |a - 1| \leq \min\{|a^2 - 1|, \sqrt{|a^2 - 1|}\}$ for $a \geq 0$. For a short proof, see Appendix B.

It is worth noting that the above argument will generalize to other structured random ensembles, e.g., partial circulant matrix with random signs. To put things in context, we compare this bound with what we got from Lemma 2. Using Lemma 3,

$$\gamma_{2/d}(T^i) \leq (1 + K \cdot \log_2(b/\alpha) \cdot b/\alpha \cdot \log_2(a))^{\frac{d-1}{2}} \gamma_2(T^i),$$

which means substituting into Theorem 2, assuming for the sake of argument all the T^i are the same, focusing on the dependence on ϵ and γ_2, this approach gives

$$m = \mathcal{O}((\sum_{i=1}^{d} \gamma_{2/d}(T^i))^2 \epsilon^{-2}) = \mathcal{O}\left((b\log_2(a))^{d-1} \cdot \gamma_2(T^i)^2 \epsilon^{-2}\right). \tag{5}$$

if ignoring poly-logs. In contrast to Lemma 2 where we used a single-scale discretization $m = \mathcal{O}(\epsilon^{-2} \cdot n^{\frac{d^2}{1+d}}(\gamma_2^2(T^i))^{\frac{d}{1+d}})$, Sudakov informs us

$$\sqrt{b\log(a)} \leq \sup_{\epsilon \in (0,1]} \epsilon\sqrt{b\log(a/\epsilon)} \lesssim \gamma_2(T^i) \leq \sqrt{n}.$$

Therefore in the case of low complexity set ($\gamma_2(T^i) \ll \sqrt{n}$), the multi-resolution approach pays off.

5 Recursive Kronecker Embedding

The row-wise-tensored mapping from the previous section, despite its simplicity, gives exponential dependency on the degree d (and necessarily so, as a preview for Section F), suggesting it is ideal for embedding low-degree tensor. In this section, we analyze the "sketch and reduce" approach proposed by [2], which composes degree-2 sketches from the previous section in the following way: we define the operation S acting on rank-1 e.g., degree-3 tensor as

$$S(x \otimes y \otimes z) := S^1(x \otimes S^2(y \otimes S^3 z)). \tag{6}$$

The distinctive feature of the design is that at each layer, the Kronecker-structured sketch S^k only acts on degree-2, reduced-dimensional tensor – something it excels at. It is an easy exercise that the matrix $S \in \mathbb{R}^{m \times n^d}$, when acting on rank-1 degree-d tensor, can be deemed as $S = Q^0$ for

$$Q^d = 1 \text{ and } Q^{k-1} = S^k(Q^k \otimes I_n) \in \mathbb{R}^{m \times n^{d-k+1}} \text{ for } k = d, \cdots, 1,$$

where each $S^k \in \mathbb{R}^{m \times nm}$ for $k \in [d-1]$ and $S^d \in \mathbb{R}^{m \times n}$.

5.1 Building blocks for multi-resolution covering

The analysis follows the same template once we know how the JL moment property is preserved under matrix direct sum and multiplication, which was investigated in previous work. We have the following discrete JL property for the embedding matrix S introduced above.

Lemma 5 (Finite Set Embedding Property)**.** *The recursive embedding* (6) *satisfies* $|\|Sx\|_2^2 - 1| \leq \max(\epsilon, \epsilon^2)$ *for all unit-norm, rank-1 tensors* $x \in \mathbb{R}^{n^d}$ *belonging to a finite set of cardinality p with probability at least* $1 - e^{-\eta}$ *for any* $\eta > 0$ *with* $m = \mathcal{O}\left(\frac{d}{\epsilon^2}(\log^2(p) + \eta^2 \vee \eta)\right)$. *Moreover, such operation can be conducted in time* $\mathcal{O}(d(n\log n + m))$ *when each S^i is constructed from an degree-2 Tensor-SRHT sketch.*

The ensuing lemma makes it clear that we should be grateful for the result stated above.

Lemma 6 (Multi-resolution embedding property of Recursive Tensor-SRHT)**.** *With* $m = \mathcal{O}(d(d^2 + (1+\eta)^2)/\epsilon^2)$, *Recursive Tensor-SRHT satisfies the* (ϵ, η, α)-*Multi-resolution Embedding Property in Definition 3 with* $\alpha = 1$.

7

5.2 Embedding of general set using recursive sketch

We will employ a slightly different decomposition of the chain for this construction and dedicate the section to prove the following theorem. At a high level, the observation is that the sketch, albeit taking complicated hierarchical form, happens to be linear when acting on rank-1 tensor. Therefore the strategy is to have all the terms in the chain we need to control in the rank-1 form that only involves difference in one factor, after which the multi-resolution embedding property can be repeatedly instantiated as before.

Theorem 3 (Gordon-type Inequality for Recursive Kronecker Embedding). *The Recursive Tensor-SRHT with $m = \mathcal{O}(d\epsilon^{-2}(\sum_{i=1}^{d}\gamma_1(T^i))^2 \cdot (d^2 + (1+\eta)^2))$ satisfies $|\|Sx\|_2^2 - 1| \leq \max(\epsilon, \epsilon^2)$ for all $x = u^1 \otimes \cdots \otimes u^d \in T^1 \times \cdots \times T^d$ with probability at least $1 - \exp(-\eta)$ for $d \geq 2$.*

It is enlightening to compare with the previous embedding bound, assuming again the covering number admits $N(T^i, sB_2^n) \leq (\frac{a}{s})^b$ for all $i \in [d]$. With (4) we have

$$\gamma_1(T^i) \leq C_1 \int_0^1 \log N(T^i, sB_2^n) \, ds \leq C_1 \int_0^1 b \log(a/s) \, ds \leq C_1' \cdot b \log(a)$$

which means using Theorem 3 that $m = \mathcal{O}(d^5 b^2 \log^2(a)/\epsilon^2)$ for the desired embedding guarantee. This is favorable as the dependence on d has been reduced from exponential to polynomial. For example we can see that when each T^i consists of a set of p points on the unit sphere, $b = o(1)$ and $a = p$ we get $\log^2(p)/\epsilon^2$ as opposed to $\log^d(p)/\epsilon^2$ from the previous section (5) when focusing on the scaling with p.

6 Applications

In this section, we give applications of our result in two settings, deploying one type of random embedding for each, where we see how these bounds can take advantage of the underlying low complexity structure to move away from the (potentially much larger) ambient dimension. We note that these applications crucially exploit the fact that the object in $\mathbb{R}^{n^d}$ being acted upon has Kronecker structure – this departs from e.g., oblivious subspace embedding (OSE) result from [1] where the column span of *any* $n^d \times p$ matrix is preserved.

6.1 Signal Recovery

Inspired by compressed sensing, suppose we are given independent random (linear) 1-subgaussian measurements on Kronecker-structured rank-1 signal x of type

$$y_i = \langle S_i, x \rangle = \prod_{j=1}^{d} \langle v_i^j, u^j \rangle, \; i \in [m] \tag{7}$$

for $x = u^1 \otimes \cdots \otimes u^d$, $u^i \in T^i \subset \mathbb{S}^{n-1}$, and would like to know when does performing

$$\min_{\{z^j\}_{j=1}^d \in \mathbb{S}^{n-1}} \sum_{j=1}^{d} f_j(z^j) \quad \text{subject to } S(z^1 \otimes \cdots \otimes z^d) = y, \; f_j(z^j) \leq R_j \; \forall j \in [d] \tag{8}$$

uniquely reconstruct x, where f_j above is convex and $R_j := f_j(u^j)$ encodes the prior knowledge we have so that $\{u^j\}$ is feasible. In the case when such information is not available, the constraint can simply read as $\|z^j\|_2 \leq 1$, for example. Notice that the decision variable lives in a lower dimensional space (nd as opposed to n^d if we naively vectorize the signal) and one candidate could be alternating projected gradient descent over each factor. Computation aside on which algorithm to enlist for solving (8) (it involves solving a non-convex problem due to the multi-linear structure), the analysis below gives an information-theoretic lower bound on the sample complexity for successful recovery. The following quantities facilitate the analysis.

Definition 4 (Descent Cone and Restricted Singular Value). *We use $\mathcal{D}(f_j, u^j)$ to denote the descent cone of a convex function f_j at point $u^j \in \mathbb{R}^n$, that is, $\mathcal{D}(f_j, u^j) := \cup_{\tau>0}\{t \in \mathbb{R}^n : f_j(u^j + \tau t) \leq f_j(u^j)\}$. The correspondingly normalized descent cone is denoted as $\bar{\mathcal{D}}(f_j, u^j) := \mathcal{D}(f_j, u^j) \cap \mathbb{S}^{n-1}$. Let $\sigma_{\min}(S; \mathcal{C})$ be the minimum singular value of a matrix S restricted to set $\mathcal{C}$, i.e., $\sigma_{\min}(S; \mathcal{C}) := \min_{x \in \mathcal{C} \cap \mathbb{S}^{n-1}} \|Sx\|$. Furthermore, the descent cone of a proper convex function is always convex.*

8

We take hints from [8, 18] for the lemma below.

Lemma 7 (Recovery Guarantee). *If $\|Sw\| \geq (1-\epsilon)\|w\|$ for all $w = (u^1 + t^1) \cap \mathbb{S}^{n-1} \otimes \cdots \otimes (u^d + t^d) \cap \mathbb{S}^{n-1}$ for which $t^j \in \mathcal{D}(f_j, u^j)$ where $\epsilon < 1$, the optimizer $\{z_*^j\}_{j=1}^d$ returned by (8) satisfies $z_*^1 \otimes \cdots \otimes z_*^d = u^1 \otimes \cdots \otimes u^d$ for the measurement model (7).*

Invoking Theorem 2 with Tensor-Subgaussian, for $\epsilon \in (0,1)$, $\forall w \in \mathcal{W}^1 \times \cdots \times \mathcal{W}^d$ where $\mathcal{W}^j := (u^j + \mathcal{D}(f_j, u^j)) \cap \mathbb{S}^{n-1}$,

$$\left| \|Sw\| - 1 \right| \leq \min\{ | \|Sw\|_2^2 - 1 |, | \|Sw\|_2^2 - 1 |^{1/2} \} \leq \epsilon$$

if picking $m = \mathcal{O}(C^d (\sum_{i=1}^d \gamma_{2/d}(\mathcal{W}^i))^2 \cdot (d^d + (1+\eta)^d)/\epsilon^2)$, which means $\sigma_{\min}(S; \mathcal{W}^1 \times \cdots \times \mathcal{W}^d) \geq 1 - \epsilon > 0$ as needed by Lemma 7.

Using translation-invariance and subadditivity of the γ-functionals, one could show that this is order-wise the same as $m = \mathcal{O}(C^d (\sum_{i=1}^d \gamma_{2/d}(\bar{\mathcal{D}}(f_i, u^i)))^2 \cdot (d^d + (1+\eta)^d))$ if f_j's are convex – we refer to Appendix E for details. Now thanks to the decoupling, it reduces to d descent cone vector Gaussian width type calculation.

Example 1. *Supppose each of the d factors is k-sparse, i.e., $T^i = \{u^i \in \mathbb{R}^n : \|u^i\|_0 \leq k, \|u^i\|_2 = 1\}$, it is classical that the normalized descent cone for ℓ_1 norm at k-sparse vector is $\bar{\mathcal{D}}(f_i, u^i) = \{s : \|s\|_1 \leq 2\sqrt{k}\|s\|_2, \|s\|_2 = 1\}$. Since $conv(kB_0^n \cap B_2^n) \subset \sqrt{k}B_1^n \cap B_2^n \subset C \cdot conv(kB_0^n \cap B_2^n)$ for an absolute constant C, from known result one can deduce that the covering number and Gaussian width scale as*

$$w(\bar{\mathcal{D}}(\|\cdot\|_1, u^j)) \asymp \sqrt{k \log(en/k)}$$
$$\log(|\mathcal{N}^\Delta(\bar{\mathcal{D}}(\|\cdot\|_1, u^j))|) \asymp k \log(en/\Delta k),$$

consequently

$$\gamma_{2/d}^2(\mathcal{D}(\|\cdot\|_1, u^j)) \lesssim (kd \log(n/k) \log(kd))^{d-1} \cdot k \log(n/k).$$

This gives assuming $\log(n/k) \ll k$ (not worrying about the d^d factor, assuming d is small for this application) with $m = \mathcal{O}\left(k^d(1+\eta)^d\right)$, the recovery is successful with probability at least $1 - \exp(-\eta)$ when omitting poly-logs. It should be clarified that the minimizer of (8) may not be unique (as in the case with $f_j = \|\cdot\|_1$ up to sign ambiguity – which is the only *possible one for rank-1 tensor), but this sample complexity suffices for recovering any of the equivalent representations of the rank-1 signal under consideration.*

In general, the work of [8, 18] provide powerful recipe for bounding the Gaussian width of a descent cone based on duality and polar cones: for f_j a convex function, and $u^j \in \mathbb{R}^n$ a fixed point, $g \sim \mathcal{N}(0, I_n)$,

$$w^2(\mathcal{D}(f_j, u^j)) \leq \mathbb{E} \inf_{\tau \geq 0} \text{dist}^2(g, \tau \cdot \partial f_j(u^j)),$$

which cries out for more applications for such structured tensor recovery problems.

6.2 Optimization

Consider an optimization (tensor decomposition) problem, where for given signal $x = u^1 \otimes \cdots \otimes u^d \in T^1 \times \cdots \times T^d$ taking Kronecker structure, we wish to solve for

$$\min_{z^i \in T^i \ \forall i \in [d]} \|u^1 \otimes \cdots \otimes u^d - z^1 \otimes \cdots \otimes z^d\|_F^2. \tag{9}$$

In general, one could also consider the denoising version where there is noise in the observation $x + e$, but for simplicity we focus on the noiseless case below. With the hope of saving storage and speeding up, we apply sketching before solving a lower m-dimensional problem:

$$\min_{z^i \in T^i \ \forall i \in [d]} \|S(u^1 \otimes \cdots \otimes u^d) - S(z^1 \otimes \cdots \otimes z^d)\|_2^2 =: g(z^1, \cdots, z^d). \tag{10}$$

Let S be the recursive sketch from Section 5 and denote the optimizer of (10) as $\{z_*^i\}$. It is not hard to see that since $g(z_*^1, \cdots, z_*^d) \leq g(u^1, \cdots, u^d) = 0$, we must have $S(z_*^1 \otimes \cdots \otimes z_*^d) = S(u^1 \otimes \cdots \otimes u^d)$, which means that S restricted to set $T^1 \times \cdots \times T^d$ must have the smallest singular value bounded away from 0 for us to uniquely identify the rank-1 factors. Note again this doesn't resolve the inherent ambiguity between the factors such as sign flips but the resulting sample complexity is sufficient to recover any such signal consistent with the measurement (i.e., the returned rank-1 solution obeys $z_*^1 \otimes \cdots \otimes z_*^d = u^1 \otimes \cdots \otimes u^d$ hence in x space it is unique). We give an example in Section E.

9

7 Experiments

In this section, we numerically investigate (1) embedding dimension scaling with d for the two types of random embeddings in Section 4 and 5; (2) signal recovery from random Gaussian measurements as elaborated in Section 6.1 where the signal is rank-1 belonging to a product of cones. This is in some sense a generalization of the non-convex Wirtinger flow formulation for the phase retrieval problem [7], where both the random measurement and the signal are non-symmetric, in addition to the availability of potential prior knowledge on the factors.

For the first experiment, we let $n = 10, d = 5$ and pick each factor $\{u^j\}$ to be 20% sparse. The figure below reports the average distortion of the embedding $|\|Sx\|_2 - 1|$ over 25 runs for both the row-wise tensored and recursive sketch with Gaussian random factors. Variance of the distortion across the trials is also shown as we vary $m = 0.8 \times n \times d^\alpha$ for $\alpha \in \{1, \cdots, 5\}$. For the second experiment, we perform projected gradient descent on the following objective:

$$\min_{\|z^1\|_1 \leq R_1, \cdots, \|z^d\|_1 \leq R_d} \frac{1}{2m} \sum_{i=1}^m \left(y_i - \prod_{j=1}^d \langle v_i^j, z^j \rangle \right)^2 =: \mathcal{L}(z^1, \cdots, z^d). \quad (11)$$

We use the spectral factorization of $\frac{1}{m} \sum_{i=1}^m \langle v_i^1, u^1 \rangle v_i^1 \otimes \cdots \otimes \langle v_i^d, u^d \rangle v_i^d = \frac{1}{m} \sum_{i=1}^m y_i v_i^1 \otimes \cdots \otimes v_i^d$ as initialization, as in expectation this is the signal $x = u^1 \otimes \cdots \otimes u^d$ since we assumed v_i^j are independent across the j's. We use the `tucker-als` function from the Matlab Tensor Toolbox[1] for computing the best rank-$(1, 1, 1)$ tensor approximation, after which gradient update is made on each factor followed by ℓ_1 projection. We set each factor $\{u^j\}$ to be 20% sparse and let $d = 3, n = 10$, $m = 2 \times 0.8 \times n \times d^\alpha$ for $\alpha \in \{1, \cdots, 3\}$ and record the successful recovery out of 25 trials. Stepsize is picked to be 0.1 and success is defined as $\mathcal{L}(z^1, \cdots, z^d) \leq 0.1$ after 500 gradient steps.
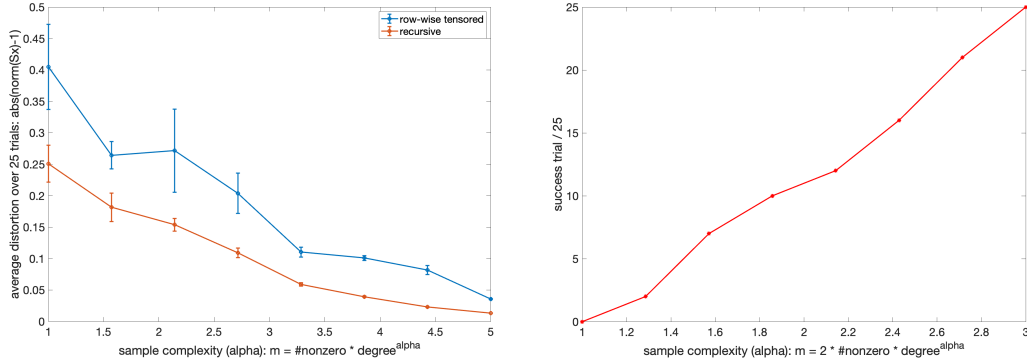
Figure 1: Left: Embedding scaling with degree. Right: Tensor recovery from Gaussian measurement.

8 Discussion

In this paper, we generalized Gordon's inequality to two families of tensor-structured random embeddings, which admit efficient implementation when acting on rank-1 Kronecker-structured signals. As future work, we deem rigorously establishing a lower bound on the embedding dimension for general sets both an interesting and challenging direction to pursue. On the other hand, tools developed here should be helpful for analyzing performance of the gradient-based recovery algorithms that we numerically tested in Section 7.

Acknowledgments and Disclosure of Funding

Our sincere appreciation goes to Rachel Ward for several discussions related to the topic. This work was performed at UT Austin supported under NSF IFML 2019844 and NSF 1934932.

---

[1]http://www.tensortoolbox.org

10

References

[1] Thomas D Ahle, Michael Kapralov, Jakob BT Knudsen, Rasmus Pagh, Ameya Velingker, David P Woodruff, and Amir Zandieh. Oblivious sketching of high-degree polynomial kernels. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 141–160. SIAM, 2020.

[2] Thomas D Ahle and Jakob BT Knudsen. Almost optimal tensor sketch. *arXiv preprint arXiv:1909.01821*, 2019.

[3] Stefan Bamberger, Felix Krahmer, and Rachel Ward. Johnson-Lindenstrauss Embeddings with Kronecker Structure. *arXiv preprint arXiv:2106.13349*, 2021.

[4] Stefan Bamberger, Felix Krahmer, and Rachel Ward. The Hanson-Wright Inequality for Random Tensors. *arXiv preprint arXiv:2106.13345*, 2021.

[5] Jean Bourgain, Sjoerd Dirksen, and Jelani Nelson. Toward a unified theory of sparse dimensionality reduction in Euclidean space. *Geometric and Functional Analysis*, 25(4):1009–1088, 2015.

[6] Jian-Feng Cai and Weiyu Xu. Guarantees of total variation minimization for signal recovery. *Information and Inference: A Journal of the IMA*, 4(4):328–353, 2015.

[7] Emmanuel J Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[8] Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational mathematics*, 12(6):805–849, 2012.

[9] Sjoerd Dirksen. Tail bounds via generic chaining. *Electronic Journal of Probability*, 20:1–29, 2015.

[10] Sjoerd Dirksen. Dimensionality reduction with subgaussian matrices: a unified theory. *Foundations of Computational Mathematics*, 16(5):1367–1396, 2016.

[11] Mathieu Even and Laurent Massoulie. Concentration of Non-Isotropic Random Tensors with Applications to Learning and Empirical Risk Minimization. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 1847–1886. PMLR, 15–19 Aug 2021.

[12] Yehoram Gordon. On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$. In *Geometric aspects of functional analysis*, pages 84–106. Springer, 1988.

[13] Rafał Latała. Estimates of moments and tails of Gaussian chaoses. *The Annals of Probability*, 34(6):2315–2331, 2006.

[14] Shahar Mendelson. Upper bounds on product and multiplier empirical processes. *Stochastic Processes and their Applications*, 126(12):3652–3680, 2016.

[15] Samet Oymak, Benjamin Recht, and Mahdi Soltanolkotabi. Isometric sketching of any set via the restricted isometry property. *Information and Inference: A Journal of the IMA*, 7(4):707–726, 2018.

[16] Michel Talagrand. *Upper and lower bounds for stochastic processes: modern methods and classical problems*, volume 60. Springer Science & Business Media, 2014.

[17] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. Regularized Linear Regression: A Precise Analysis of the Estimation Error. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1683–1709, Paris, France, 03–06 Jul 2015. PMLR.

[18] Joel A Tropp. Convex recovery of a structured signal from independent random linear measurements. In *Sampling Theory, a Renaissance*, pages 67–101. Springer, 2015.

11

[19] Aad W Van Der Vaart and Jon Wellner. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 1996.

[20] Ramon van Handel. Chaining, interpolation and convexity II: The contraction principle. *The Annals of Probability*, 46(3):1764–1805, 2018.

[21] Roman Vershynin. Concentration inequalities for random tensors. *Bernoulli*, 26(4):3139–3162, 2020.

[22] Vincent Q Vu and Jing Lei. Squared-norm empirical process in Banach space. *arXiv preprint arXiv:1312.1005*, 2013.

[23] Nikita Zhivotovskiy. Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle. *arXiv preprint arXiv:2108.08198*, 2021.

Checklist

1. For all authors...

    (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]

    (b) Did you describe the limitations of your work? [Yes]

    (c) Did you discuss any potential negative societal impacts of your work? [N/A]

    (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

    (a) Did you state the full set of assumptions of all theoretical results? [Yes]

    (b) Did you include complete proofs of all theoretical results? [Yes]

3. If you ran experiments...

    (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]

    (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]

    (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]

    (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

    (a) If your work uses existing assets, did you cite the creators? [Yes]

    (b) Did you mention the license of the assets? [N/A]

    (c) Did you include any new assets either in the supplemental material or as a URL?[N/A]

    (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

    (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

    (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

    (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

    (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

12

A   Proofs for Section 3

This section deals with uniform covering, i.e., a single-scale approach.

Proof of Lemma 1. Tracing the footsteps for the proof of Theorem 3 in [2], one could check that we can smuggle in the term $\max(\epsilon, \epsilon^2)$ replacing $\epsilon \in (0,1)$ for the following guarantee: with

$$m = \mathcal{O}\left(C^d \frac{1}{\epsilon^2} \log(\frac{1}{\delta}) \log^d(\frac{1}{\delta})\right),$$

the resulting Tensor SRHT matrix $S \in \mathbb{R}^{m \times n^d}$ constructed as in Definition 1 satisfies (1) $\mathbb{E}[\|Sx\|_2^2] = 1$ for all $\|x\| = 1$; (2) $\mathbb{E}[\|\|Sx\|_2^2 - 1|^{\log(1/\delta)}] \leq (\frac{1}{e} \max(\epsilon, \epsilon^2))^{\log(1/\delta)}$. This implies via Markov's inequality,

$$\mathbb{P}(\|\|Sx\|_2^2 - 1| \geq \max(\epsilon, \epsilon^2)) \leq \delta$$

for any unit norm x and $\delta \in (0,1)$. Therefore for a set of cardinality p, we take a union bound to conclude with probability at least $1 - e^{-\eta}$ for any $\eta > 0$,

$$\|\|Sx\|_2^2 - 1| \leq \max(\epsilon, \epsilon^2)$$

simultaneously for all p points on the unit sphere in the set provided

$$m = \mathcal{O}\left(C^d \frac{1}{\epsilon^2}(\log^d(p) + \eta^d \vee \eta)\right)$$

for some universal constant C, which renders the advertised bound by recognizing $\eta^d \vee \eta \leq (1+\eta)^d$. The same argument applies to Tensor-Subgaussian embedding by working with Theorem 2 in [2] instead. $\qquad\square$

Proof of Lemma 2. Let $\Delta < 1/2$, to cast an Δ-net (in Frobenius norm) for the rank-1 tensor, suppose for each of the d factors $u^i \in T^i \subset \mathbb{S}^{n-1}$ we find $v^i \in T^i \cap \mathcal{N}^i$ such that $\|v^i - u^i\|_2 \leq \Delta/d$, then

$$\|v^1 \otimes \cdots \otimes v^d - u^1 \otimes \cdots \otimes u^d\|_F \leq \sum_{i=1}^d \|v^1\| \cdots \|v^{i-1}\|\|v^i - u^i\|\|u^{i+1}\| \cdots \|u^d\|$$

$$\leq d \times \Delta/d = \Delta.$$

To extend (2) to hold for all $x \in T$, write $u^i = v^i + l^i$ for $\|l^i\|_2 \leq \Delta/d$, and recall $\|u^1 \otimes \cdots \otimes u^d\| = \|v^1 \otimes \cdots \otimes v^d\| = 1$,

$$|\text{vec}(u^1 \otimes \cdots \otimes u^d)^\top(S^\top S - I)\text{vec}(u^1 \otimes \cdots \otimes u^d)|$$

$$\leq |\text{vec}(v^1 \otimes \cdots \otimes v^d)^\top(S^\top S - I)\text{vec}(v^1 \otimes \cdots \otimes v^d)|$$

$$+ 2|(\text{vec}(u^1 \otimes \cdots \otimes u^d) - \text{vec}(v^1 \otimes \cdots \otimes v^d))^\top(S^\top S - I)\text{vec}(v^1 \otimes \cdots \otimes v^d)|$$

$$\leq \max_{v^i \in T^i \cap \mathcal{N}^i} |\text{vec}(v^1 \otimes \cdots \otimes v^d)^\top(S^\top S - I)\text{vec}(v^1 \otimes \cdots \otimes v^d)| + 2\Delta\|S^\top S - I\|_{op}$$

so taking sup over $u^i \in T^i$, we have

$$\|S^\top S - I\|_{op,T} \leq \max_{v^i \in T^i \cap \mathcal{N}^i} |\text{vec}(v^1 \otimes \cdots \otimes v^d)^\top(S^\top S - I)\text{vec}(v^1 \otimes \cdots \otimes v^d)| + 2\Delta\|S^\top S - I\|_{op}.$$

Therefore to have the distortion below order $\max(\epsilon, \epsilon^2)$ for all x, it suffices to cover each factor to accuracy Δ/d for $\Delta \lesssim \max\{\epsilon, \epsilon^2\}/\|S^\top S - I\|_{op}$ and union bound over this finite set to guarantee $\max(\epsilon, \epsilon^2)$ distortion on it. Now since on the unit sphere $T^i = \mathbb{S}^{n-1}$ for any $\Delta = o(1)$,

$$\|S^\top S - I\|_{op} \leq \frac{1}{1 - 2\Delta} \cdot \max_{\tilde{x} \in \mathbb{S} \cap \mathcal{N}^\Delta} |\tilde{x}^\top(S^\top S - I)\tilde{x}|,$$

this suggests $\|S^\top S - I\|_{op} \lesssim n^d/m$.

Using Sudakov's minorization, the cardinality of the finite set $p \lesssim \prod_{i=1}^d \exp\left(\frac{d^2}{\Delta^2} w^2(T^i)\right) \lesssim \prod_{i=1}^d \exp\left(\frac{d^2 n^d}{\epsilon^2 m} w^2(T^i)\right)$. Owing to the existence of Lemma 1 on Discrete-JL for Tensor-SRHT, it yields the dependence on $w(T^i)$ (and therefore $\gamma_2(T^i)$) should scale as

$$m = \mathcal{O}\left(\frac{1}{\epsilon^2} \cdot n^{\frac{d^2}{1+d}}\left(\sum_{i=1}^d \gamma_2^2(T^i)\right)^{\frac{d}{1+d}}\right)$$

for such a uniform concentration to hold, ignoring d^d dependence. $\qquad\square$

13

B   Proofs for Section 4

We begin with a few helper lemmas for setting up the chaining and multi-resolution covering argument. Below is an easy observation on the sequence of successive coverings $\{\bar{T}_l^i\}_l$ as defined in (3).

Lemma 8 (Choice of L). *For all $u^i \in T^i$, $dist(u^i, \bar{T}_L^i) \lesssim \frac{1}{d}$ for $L \gtrsim \lceil \log_2(nd) \rceil$.*

Proof of Lemma 8. Since $\gamma_2(T^i) \asymp \gamma_2^*(T^i)$ and (3) shows that

$$\sup_{u^i \in T^i} \sum_{l=0}^{\infty} 2^{l/2} \text{dist}(u^i, \bar{T}_l^i) \leq \gamma_2(T^i),$$

it is necessarily the case that (using the $\inf$ in the definition of γ_2^*)

$$\sup_{u^i \in T^i} \sum_{l=0}^{\infty} 2^{l/2} \text{dist}(u^i, \bar{T}_l^i) \asymp \gamma_2^*(T^i),$$

so $\{\bar{T}_l^i\}_l$ is almost optimal, which means that for the classical greedily constructed ϵ_l-net for the unit Euclidean ball with cardinality $(1 + 2/\epsilon_l)^n \leq N_l$ for $\epsilon_l = n/2^{l-2}$, and for any $u^i \in T^i$,

$$\sum_{l=0}^{\infty} 2^{l/2} \text{dist}(u^i, \bar{T}_l^i) \lesssim \sum_{l=0}^{\infty} 2^{l/2} \cdot \epsilon_l = n \cdot \text{const}$$

therefore $\text{dist}(u^i, \bar{T}_l^i) \lesssim \frac{n}{(l+1) \cdot 2^{l/2}}$ with a proof of contradiction, where in the above we pushed $\sup$ inside. This in turn indicates that sending $L \gtrsim \lceil \log_2(nd) \rceil$ the distortion for approximating any $u^i \in T^i \subset \mathbb{S}^{n-1}$ with the net $\{\bar{T}_L^i\}$ is below order $1/d$. $\square$

Proof of Lemma 3. By definition, the γ_α functional is monotonically non-increasing in α. The other side of the inequality involves a careful look into the admissible sequence. Pick a cutoff level l_c to be specified later, for the optimal admissible sequence $\{\bar{\mathcal{A}}_l^i\}_l$ for the γ_2 functional we construct another admissible sequence $\{\mathcal{B}_l^i\}_l$ that coincides with $\{\bar{\mathcal{A}}_l^i\}_l$ for $l \leq l_c$, and observe that

$$\sup_{u^i \in T^i} \sum_{l \leq l_c} 2^{l/\alpha} \text{diam}(\bar{\mathcal{A}}_l^i(u^i)) \leq \sup_{u^i \in T^i} 2^{\frac{(2-\alpha)l_c}{2\alpha}} \sum_{l \leq l_c} 2^{l/2} \text{diam}(\bar{\mathcal{A}}_l^i(u^i)). \tag{12}$$

For the scales $l > l_c$, we aim to pick l_c large enough so that $\sum_{l > l_c} 2^{l/\alpha} \text{diam}(\mathcal{B}_l^i(u^i)) \leq 1$ for $|\mathcal{B}_l^i| \leq 2^{2^l}$ being the tightest covering of elements of $\bar{\mathcal{A}}_l^i$ for each $l > l_c$. Since $N(T^i, sB_2^n) \leq (\frac{a}{s})^b$, we have $s_l = a2^{-2^l/b}$ distortion, which means for $l = l_c + e$ where $e > 0$, and any $u^i \in T^i$,

$$2^{l/\alpha} \text{diam}(\mathcal{B}_l^i(u^i)) \leq a2^{l/\alpha - 2^l/b} = 2^{\log_2(a) + l/\alpha - 2^l/b}.$$

Therefore put $l_c = \log_2\left(K \cdot \log_2(b/\alpha) \cdot b/\alpha \cdot \log_2(a)\right)$ for a sufficiently large constant K (essentially we need l_c large enough such that $2^{l_c} - b/\alpha \cdot l_c \geq (\log_2(a) + 1)b$),

$$\sum_{l > l_c} 2^{l/\alpha} \text{diam}(\mathcal{B}_l^i(u^i)) \leq \sum_{e \geq 0} \frac{1}{2^{e+1}} \leq 1.$$

Plugging l_c back into (12), altogether this gives (since γ_α takes $\inf$ over all admissible sequences, of which $\{\mathcal{B}_l^i\}_l$ is one)

$$\gamma_\alpha(T^i) \leq 1 + \left(K \cdot \log_2(b/\alpha) \cdot b/\alpha \cdot \log_2(a)\right)^{\frac{2-\alpha}{2\alpha}} \cdot \gamma_2(T^i)$$

$$\leq \left(1 + K \cdot \log_2(b/\alpha) \cdot b/\alpha \cdot \log_2(a)\right)^{\frac{2-\alpha}{2\alpha}} \cdot \gamma_2(T^i)$$

where we used $\gamma_2(T^i) \geq \text{diam}(T^i)/2 = 1$. $\square$

Remark. The polynomial covering number assumption for Lemma 3 is a natural one: For VC class with VC dimension v, we have the covering number bound $N(T^i) \leq Kv(4e)^v(\frac{1}{s})^{2(v-1)}$ for some universal constant K. (cf. Theorem 2.6.4 of [19])

14

The corollary below illustrates that with the specified choice of L, together with the embedding property Definition 3, we have control on the distortion of S acting on all rank-1 tensors of interest.

Corollary 1 (Approximation at the L-th level)**.** *For any* $x \in T^1 \times \cdots \times T^d$, *with* $L \asymp \lceil \log_2(nd) \rceil$ *and the net* $\{\bar{T}_l^i\}_l$ *constructed from an optimal admissible sequence as in* (3) *for each set* $T^i \subset \mathbb{S}^{n-1}$,

$$\|S(x - \tilde{x}_L)\|_2 \lesssim (1 + 2^{L/\alpha} \epsilon).$$

Proof. Since $L \asymp \lceil \log_2(nd) \rceil$, it ensures $\mathrm{dist}(u^i, \bar{T}_L^i) \lesssim \frac{1}{d}$ using Lemma 8 for all $u^i \in T^i$. Now for $\tilde{x}_L = u_L^1 \otimes \cdots \otimes u_L^d \in \bar{T}_L^1 \times \cdots \times \bar{T}_L^d$,

$$\|S(x - \tilde{x}_L)\|_2 \leq \sum_{k=1}^{d} \|S(u^1 \otimes \cdots \otimes u^{k-1} \otimes (u^k - u_L^k) \otimes u_L^{k+1} \otimes \cdots \otimes u_L^d)\|_2$$

$$\lesssim \sum_{k=1}^{d} (1 + 2^{L/\alpha} \epsilon) \frac{1}{d} \lesssim (1 + 2^{L/\alpha} \epsilon)$$

where in the above we used part 1 of the multi-resolution embedding property Definition 3 which implies

$$\|S(u^1 \otimes \cdots \otimes u^{k-1} \otimes (u^k - u_L^k) \otimes u_L^{k+1} \otimes \cdots \otimes u_L^d)\|_2^2 \lesssim (\max(2^{L/\alpha} \epsilon, 2^{2L/\alpha} \epsilon^2) + 1) \cdot \frac{1}{d^2}$$

$$\lesssim (1 + 2^{L/\alpha} \epsilon)^2 \cdot \frac{1}{d^2}$$

for any $k \in [d]$. $\qquad\square$

The following two concern the row-wise tensored embeddings and are the main results of this section.

Proof of Lemma 4. The requirement entails that the cardinality of the set $p_l \leq 5d \cdot (2^{2^l})^{d-1} \cdot (2^{2^l})^2$ for each level of distortion $1 \leq l \leq L$ with $\epsilon_l = 2^{ld/2}\epsilon$ and $\eta_l = l(\eta + 1) \geq 1$ (i.e., we only look at points belonging to neighboring scales). Union bounding over $L \asymp \lceil \log_2(nd) \rceil$ levels, using Lemma 1, we get with

$$m = \mathcal{O}\Big(C^d \frac{1}{\epsilon_l^2} (\log^d(p_l) + \eta_l^d \vee \eta_l) \Big)$$

which is $\mathcal{O}(C^d(d^d + (1 + \eta)^d)/\epsilon^2)$ hiding poly-logs that all events as required in Definition 3 holds with probability at least

$$1 - \sum_{l=1}^{L} \exp(-\eta_l) \geq 1 - \sum_{l=1}^{\infty} \exp(-l(\eta + 1)) \geq 1 - \exp(-\eta),$$

as claimed. For the Tensor-Subgaussian sketch, this becomes $\mathcal{O}(C^d \sigma^{2d}(d^d + (1 + \eta)^d)/\epsilon^2)$ using again Lemma 1. $\qquad\square$

Proof of Theorem 2. Throughout the section, we work with the net $\{\bar{T}_l^k\}_{l \in [L]}$ constructed from (3) for each $k \in [d]$. Using triangle inequality, forming a telescoping sum and let $\tilde{x}_l = u_l^1 \otimes \cdots \otimes u_l^d$ for each $l \in [\tilde{L}]$ where $\max(\epsilon_l, \epsilon_l^2) = \epsilon_l$, and on the event $\tilde{L} < L$,

$$\big| \|Sx\|_2^2 - \|x\|_2^2 \big|$$

$$\leq \left| \sum_{l=1}^{\tilde{L}} \sum_{i=1}^{m} \left(\prod_{k=1}^{d} \langle v_i^k, u_l^k \rangle^2 - \prod_{k=1}^{d} \langle v_i^k, u_{l-1}^k \rangle^2 \right) - \left(\prod_{k=1}^{d} \|u_l^k\|_2^2 - \prod_{k=1}^{d} \|u_{l-1}^k\|_2^2 \right) \right|$$

$$+ \big| \|Sx\|_2^2 - \|S\tilde{x}_{\tilde{L}}\|_2^2 \big| + \big| \|x\|_2^2 - \|\tilde{x}_{\tilde{L}}\|_2^2 \big| + \big| \|S\tilde{x}_0\|_2^2 - \|\tilde{x}_0\|_2^2 \big|$$

$$\leq \sum_{l=1}^{\tilde{L}} \sum_{k=1}^{d} \left| \sum_{i=1}^{m} \left(\langle v_i^k, u_l^k \rangle^2 - \langle v_i^k, u_{l-1}^k \rangle^2 \right) \times \prod_{s=1}^{k-1} \langle v_i^s, u_l^s \rangle^2 \times \prod_{s=k+1}^{d} \langle v_i^s, u_{l-1}^s \rangle^2 \right|$$

15

$$- \left(\|u_l^k\|_2^2 - \|u_{l-1}^k\|_2^2 \right) \times \prod_{s=1}^{k-1} \|u_l^s\|_2^2 \times \prod_{s=k+1}^{d} \|u_{l-1}^s\|_2^2 \bigg|$$

$$+ \left| \|Sx\|_2^2 - \|S\tilde{x}_{\tilde{L}}\|_2^2 \right| + \left| \|x\|_2^2 - \|\tilde{x}_{\tilde{L}}\|_2^2 \right| + \left| \|S\tilde{x}_0\|_2^2 - \|\tilde{x}_0\|_2^2 \right|$$

$$\leq \sum_{l=1}^{\tilde{L}} \sum_{k=1}^{d} \sum_{i=1}^{m} \bigg| \sum \left(\langle v_i^k, u_l^k - u_{l-1}^k \rangle^2 + 2 \langle v_i^k, u_l^k - u_{l-1}^k \rangle \langle v_i^k, u_l^k - u_{l-1}^k \rangle \right) \times \prod_{s=1}^{k-1} \langle v_i^s, u_l^s \rangle^2 \times \prod_{s=k+1}^{d} \langle v_i^s, u_{l-1}^s \rangle^2$$

$$- \left(\|u_l^k - u_{l-1}^k\|_2^2 + 2 \langle u_{l-1}^k, u_l^k - u_{l-1}^k \rangle \right) \times \prod_{s=1}^{k-1} \|u_l^s\|_2^2 \times \prod_{s=k+1}^{d} \|u_{l-1}^s\|_2^2 \bigg|$$

$$+ \left| \|Sx\|_2^2 - \|S\tilde{x}_{\tilde{L}}\|_2^2 \right| + \left| \|x\|_2^2 - \|\tilde{x}_{\tilde{L}}\|_2^2 \right| + \left| \|S\tilde{x}_0\|_2^2 - \|\tilde{x}_0\|_2^2 \right|$$

$$\leq \sum_{k=1}^{d} \sum_{l=1}^{\tilde{L}} \sum_{i=1}^{m} \bigg| \sum \langle v_i^k, u_l^k - u_{l-1}^k \rangle^2 \prod_{s=1}^{k-1} \langle v_i^s, u_l^s \rangle^2 \prod_{s=k+1}^{d} \langle v_i^s, u_{l-1}^s \rangle^2 - \|u_l^k - u_{l-1}^k\|_2^2 \prod_{s=1}^{k-1} \|u_l^s\|_2^2 \prod_{s=k+1}^{d} \|u_{l-1}^s\|_2^2 \bigg|$$

$$+ 2 \sum_{k=1}^{d} \sum_{l=1}^{\tilde{L}} \sum_{i=1}^{m} \bigg| \sum \langle v_i^k, u_l^k - u_{l-1}^k \rangle \langle v_i^k, u_l^k - u_{l-1}^k \rangle \prod_{s=1}^{k-1} \langle v_i^s, u_l^s \rangle^2 \prod_{s=k+1}^{d} \langle v_i^s, u_{l-1}^s \rangle^2$$

$$- \langle u_{l-1}^k, u_l^k - u_{l-1}^k \rangle \prod_{s=1}^{k-1} \|u_l^s\|_2^2 \prod_{s=k+1}^{d} \|u_{l-1}^s\|_2^2 \bigg|$$

$$+ \left| \|Sx\|_2^2 - \|S\tilde{x}_{\tilde{L}}\|_2^2 \right| + \left| \|x\|_2^2 - \|\tilde{x}_{\tilde{L}}\|_2^2 \right| + \left| \|S\tilde{x}_0\|_2^2 - \|\tilde{x}_0\|_2^2 \right|$$

and we attend to each of the 5 terms above in turn.

Term 1: Fixing a $k \in [d]$ and $l \in [\tilde{L}]$, invoking Definition 3, the first term can be written as for $S \in \mathbb{R}^{m \times n^d}$ (recall u_l^k is the closet point to $u^k \in T^k \subset \mathbb{S}^{n-1}$ in the l-th level covering of T^k therefore $\|u_l^k\|_2 = 1$)

$$\left| \|S \cdot \mathrm{vec}(u_l^1 \otimes \cdots \otimes (u_l^k - u_{l-1}^k) \otimes \cdots \otimes u_{l-1}^d)\|_2^2 - \|\mathrm{vec}(u_l^1 \otimes \cdots \otimes (u_l^k - u_{l-1}^k) \otimes \cdots \otimes u_{l-1}^d)\|_2^2 \right|$$
$$\leq \max(2^{ld/2}\epsilon, 2^{ld}\epsilon^2) \cdot \|u_l^1\|_2^2 \cdots \|u_l^k - u_{l-1}^k\|_2^2 \cdots \|u_{l-1}^d\|_2^2$$
$$\leq \max(2^{ld/2}\epsilon, 2^{ld}\epsilon^2) \cdot (\|u_l^k - u^k\|_2 + \|u^k - u_{l-1}^k\|_2)^2$$
$$\leq 2^{ld/2}\epsilon \cdot 4 \cdot \mathrm{dist}^2(u^k, \bar{T}_{l-1}^k) \leq 2^{ld/2}\epsilon \cdot 8 \cdot \mathrm{dist}(u^k, \bar{T}_{l-1}^k)$$

Summing over l and d gives Term 1 is upper bouned by

$$8\sqrt{2}\epsilon \sum_{k=1}^{d} \sum_{l=0}^{\tilde{L}-1} 2^{ld/2} \mathrm{dist}(u^k, \bar{T}_l^k) \leq 8\sqrt{2}\epsilon \cdot \sum_{k=1}^{d} \gamma_{2/d}(T^k) \, .$$

Term 2: The second term is

$$2 \left| \mathrm{vec}(u_l^1 \otimes \cdots \otimes u_{l-1}^k \otimes \cdots \otimes u_{l-1}^d) \cdot (S^\top S - I) \cdot \mathrm{vec}(u_l^1 \otimes \cdots \otimes (u_l^k - u_{l-1}^k) \otimes \cdots \otimes u_{l-1}^d) \right|$$
$$\leq \max(2^{ld/2}\epsilon, 2^{ld}\epsilon^2) \cdot 2 \cdot \|u_l^k - u_{l-1}^k\|_2 \leq 2^{ld/2}\epsilon \cdot 4 \cdot \mathrm{dist}(u^k, \bar{T}_{l-1}^k) \, ,$$

where we used that part 3 of Definition 3 implies that since S is linear on rank-1 tensors, therefore

$$\left| \mathrm{vec}\!\left(u_l^1 \otimes \cdots \otimes \frac{u_l^k - u_{l-1}^k}{\|u_l^k - u_{l-1}^k\|_2} \otimes \cdots \otimes u_{l-1}^d \right) \cdot (S^\top S - I) \cdot \mathrm{vec}(u_l^1 \otimes \cdots \otimes u_{l-1}^k \otimes \cdots \otimes u_{l-1}^d) \right|$$

$$\leq \frac{1}{4} \bigg\{ \bigg| \mathrm{vec}\!\left(u_l^1 \otimes \cdots \otimes \left(\frac{u_l^k - u_{l-1}^k}{\|u_l^k - u_{l-1}^k\|_2} + u_{l-1}^k \right) \otimes \cdots \otimes u_{l-1}^d \right) \cdot (S^\top S - I)$$

$$\cdot \mathrm{vec}\!\left(u_l^1 \otimes \cdots \otimes \left(\frac{u_l^k - u_{l-1}^k}{\|u_l^k - u_{l-1}^k\|_2} + u_{l-1}^k \right) \otimes \cdots \otimes u_{l-1}^d \right) \bigg|$$

$$+ \bigg| \mathrm{vec}\!\left(u_l^1 \otimes \cdots \otimes \left(\frac{u_l^k - u_{l-1}^k}{\|u_l^k - u_{l-1}^k\|_2} - u_{l-1}^k \right) \otimes \cdots \otimes u_{l-1}^d \right) \cdot (S^\top S - I)$$

16

$$\cdot \operatorname{vec}\left(u_l^1 \otimes \cdots \otimes \left(\frac{u_l^k - u_{l-1}^k}{\|u_l^k - u_{l-1}^k\|_2} - u_{l-1}^k\right) \otimes \cdots \otimes u_{l-1}^d\right)\bigg|\bigg\}$$

$$\leq \max(2^{ld/2}\epsilon, 2^{ld}\epsilon^2) \cdot \frac{1}{4} \cdot \left(2 + 2\left(\frac{u_l^k - u_{l-1}^k}{\|u_l^k - u_{l-1}^k\|_2}\right)^\top u_{l-1}^k + 2 - 2\left(\frac{u_l^k - u_{l-1}^k}{\|u_l^k - u_{l-1}^k\|_2}\right)^\top u_{l-1}^k\right)$$

$$= \max(2^{ld/2}\epsilon, 2^{ld}\epsilon^2).$$

Summing over l and d, the second term is upper bounded by

$$4\sqrt{2}\epsilon \sum_{k=1}^{d} \sum_{l=0}^{\tilde{L}-1} 2^{ld/2}\operatorname{dist}(u^k, \bar{T}_l^k) \leq 4\sqrt{2}\epsilon \cdot \sum_{k=1}^{d} \gamma_{2/d}(T^k).$$

Term 3: For the third term, we begin by noting that

$$\left|\|Sx\|_2^2 - \|S\tilde{x}_{\tilde{L}}\|_2^2\right| \leq \left|\|Sx\|_2 - \|S\tilde{x}_{\tilde{L}}\|_2\right| \cdot \left[\|Sx\|_2 + \|S\tilde{x}_{\tilde{L}}\|_2\right]$$
$$\leq \left|\|Sx\|_2 - \|S\tilde{x}_{\tilde{L}}\|_2\right|^2 + 2 \cdot \left|\|Sx\|_2 - \|S\tilde{x}_{\tilde{L}}\|_2\right| \cdot \|S\tilde{x}_{\tilde{L}}\|_2,$$

therefore we are left to wrestle with

$$\left|\|Sx\|_2 - \|S\tilde{x}_{\tilde{L}}\|_2\right|$$
$$\leq \|S(x - \tilde{x}_L)\| + \|S(\tilde{x}_L - \tilde{x}_{\tilde{L}})\|$$

$$\leq \|S(x - \tilde{x}_L)\|_2 + \sum_{l=\tilde{L}+1}^{L} \|S(\tilde{x}_l - \tilde{x}_{l-1})\|_2$$

$$\lesssim (1 + 2^{Ld/2}\epsilon) + \sum_{l=\tilde{L}+1}^{L} \sum_{k=1}^{d} \|S \cdot \operatorname{vec}(u_l^1 \otimes \cdots \otimes u_l^{k-1} \otimes (u_l^k - u_{l-1}^k) \otimes u_{l-1}^{k+1} \otimes \cdots \otimes u_{l-1}^d)\|_2$$

$$\lesssim (1 + 2^{Ld/2}\epsilon) + \sum_{l=\tilde{L}+1}^{L} \sum_{k=1}^{d} (1 + 2^{ld/2}\epsilon) \cdot \|u_l^k - u_{l-1}^k\|_2$$

$$\lesssim \sum_{k=1}^{d} \left\{2 \cdot 2^{Ld/2}\frac{\epsilon}{d} + \sum_{l=\tilde{L}+1}^{L} 2 \cdot 2^{ld/2}\epsilon \cdot 2 \cdot \operatorname{dist}(u^k, \bar{T}_{l-1}^k)\right\}$$

$$\lesssim \sum_{k=1}^{d} \left\{2 \cdot 2^{Ld/2}\frac{\epsilon}{d} + 4 \cdot 2^{d/2} \cdot \sum_{l=\tilde{L}+1}^{L} 2^{(l-1)d/2}\epsilon \cdot \operatorname{dist}(u^k, \bar{T}_{l-1}^k)\right\}$$

$$\lesssim 4 \cdot 2^{d/2}\epsilon \cdot \sum_{k=1}^{d} \gamma_{2/d}(T^k)$$

where we used Definition 3 and that $\epsilon_l \geq 1$ for $l \geq \tilde{L}$ and Corollary 1. It remains to bound $\|S\tilde{x}_{\tilde{L}}\|_2$, for this,

$$\|S\tilde{x}_{\tilde{L}}\|_2 \leq 2^{\tilde{L}d/2}\epsilon + 1 \leq 2$$

where we again used $1 + \max(\epsilon_{\tilde{L}}, \epsilon_{\tilde{L}}^2) \leq (1 + \epsilon_{\tilde{L}})^2$ and $\epsilon_{\tilde{L}} \leq 1$. Altogether this yields

$$\left|\|Sx\|_2^2 - \|S\tilde{x}_{\tilde{L}}\|_2^2\right| \lesssim 16 \cdot 2^d\epsilon^2 \cdot \left(\sum_{k=1}^{d} \gamma_{2/d}(T^k)\right)^2 + 16 \cdot 2^{d/2}\epsilon \cdot \sum_{k=1}^{d} \gamma_{2/d}(T^k).$$

Term 4: Analogous to the previous part, we have

$$\left|\|x\|_2^2 - \|\tilde{x}_{\tilde{L}}\|_2^2\right| \leq \left|\|x\|_2 - \|\tilde{x}_{\tilde{L}}\|_2\right| \cdot \left[\|x\|_2 + \|\tilde{x}_{\tilde{L}}\|_2\right]$$
$$\leq \left|\|x\|_2 - \|\tilde{x}_{\tilde{L}}\|_2\right|^2 + 2 \cdot \left|\|x\|_2 - \|\tilde{x}_{\tilde{L}}\|_2\right| \cdot \|\tilde{x}_{\tilde{L}}\|_2,$$

where

$$\left|\|x\|_2 - \|\tilde{x}_{\tilde{L}}\|_2\right| \leq \sum_{l=\tilde{L}}^{L} \|\tilde{x}_{l+1} - \tilde{x}_l\|_2 \leq \sum_{l=\tilde{L}}^{L} \sum_{k=1}^{d} \|u_{l+1}^1\| \cdots \|u_{l+1}^{k-1}\|\|u_{l+1}^k - u_l^k\|\|u_l^{k+1}\| \cdots \|u_l^d\|$$

17

$$\leq \sum_{l=\tilde{L}}^{L}\sum_{k=1}^{d} 2 \cdot \mathrm{dist}(u^k, \bar{T}_l^k) \leq \sum_{l=\tilde{L}}^{L}\sum_{k=1}^{d} 2 \cdot 2^{ld/2}\epsilon \cdot \mathrm{dist}(u^k, \bar{T}_l^k)$$

$$\leq 2\epsilon \sum_{k=1}^{d} \gamma_{2/d}(T^k)$$

using $\epsilon_l \geq 1$ for $l \geq \tilde{L}$. Therefore since $\|\tilde{x}_{\tilde{L}}\|_2 = 1$,

$$\left|\|x\|_2^2 - \|\tilde{x}_{\tilde{L}}\|_2^2\right| \leq 4\epsilon^2 \left(\sum_{k=1}^{d} \gamma_{2/d}(T^k)\right)^2 + 4\epsilon \sum_{k=1}^{d} \gamma_{2/d}(T^k).$$

Term 5: The last missing piece directly follows from part 2 of Definition 3:
$$\left|\|S\tilde{x}_0\|_2^2 - \|\tilde{x}_0\|_2^2\right| \leq \max(\epsilon, \epsilon^2).$$
Now to finish the train of thought, we collect the results and use the definition of the γ_α-functional

$$2^{d/2}\sum_{k=1}^{d} \gamma_{2/d}(T^k) \geq 2^{d/2}\sum_{k=1}^{d} \gamma_2(T^k) \geq 2^{d/2}\sum_{k=1}^{d} \mathrm{diam}(T^k)/2 = 2^{d/2}d > 1$$

since $T^k \subset \mathbb{S}^{n-1}$, to reach

$$\left|\|Sx\|_2^2 - \|x\|_2^2\right| \lesssim 2^d\epsilon^2 \cdot \left(\sum_{k=1}^{d} \gamma_{2/d}(T^k)\right)^2 + 2^{d/2}\epsilon \cdot \sum_{k=1}^{d} \gamma_{2/d}(T^k) + \max(\epsilon, \epsilon^2)$$

$$\lesssim \max\left\{2^{d/2}\epsilon \cdot \sum_{k=1}^{d} \gamma_{2/d}(T^k), \, 2^d\epsilon^2 \cdot \left(\sum_{k=1}^{d} \gamma_{2/d}(T^k)\right)^2\right\} + \max(\epsilon, \epsilon^2)$$

$$\lesssim \max\left\{2^{d/2}\epsilon \cdot \sum_{k=1}^{d} \gamma_{2/d}(T^k), \, 2^d\epsilon^2 \cdot \left(\sum_{k=1}^{d} \gamma_{2/d}(T^k)\right)^2\right\}$$

hence a re-scaling $\epsilon \mapsto \frac{\epsilon}{2^{d/2}\sum_{k=1}^{d}\gamma_{2/d}(T^k)}$ will deliver the desired embedding property. Putting together with Lemma 4 we have the sample complexity

$$m = \mathcal{O}\left(C^d(\sum_{i=1}^{d}\gamma_{2/d}(T^i))^2 \cdot (d^d + (1+\eta)^d)/\epsilon^2\right)$$

for Tensor-SRHT. The claim for Tensor-Subgaussian also follows modulo notation adjustments. $\quad\square$

Remark. It is likely that one would be able to handle more general concentration of Lipschitz nonlinearities using modification of the argument, but the expectation will quite possibly become hard to compute. By reckoning that $\mathbb{E}_{s\sim\mathcal{N}(0,I)}[(s^\top x)^2(s^\top y)^2] = \|x\|_2^2 \cdot \|y\|_2^2 + 2(x^\top y)^2$, for $d = 2$ and orthogonal factors, one could also save randomness by potentially using a symmetric degree-2 sketch and appeal to [14]'s result on product processes for near-isometric embedding of arbitrary sets, with dependence on geometric properties of individual sets T^1 and T^2 (they studied quantities of type $\sup_{f\in\mathcal{F},h\in\mathcal{H}}|\frac{1}{m}\sum_{i=1}^{m} f(X_i)h(X_i) - \mathbb{E}[fh]|)$.

The lemma below is relatively standard – we include it here for completeness.

Lemma 9. *The following numerical inequality holds:* $\frac{1}{3}\min\{|a^2-1|, \sqrt{|a^2-1|}\} \leq |a-1| \leq \min\{|a^2-1|, \sqrt{|a^2-1|}\}$ *for* $a \geq 0$.

Proof. For the right inequality, $|a-1| = |a^2-1|/|a+1| \leq |a^2-1|$ always since $a \geq 0$. In the case $\sqrt{|a^2-1|} \leq |a^2-1|$, in which case $|a^2-1| \geq 1 \Rightarrow a^2 \geq 2$, therefore $|a-1| \leq \sqrt{|(a+1)(a-1)|}$ since it's obviously true for $a \geq \sqrt{2}$ (recall by assumption $a \geq 0$).

For the left inequality, in the case $a^2 \leq 2$, $|a-1| = \frac{|a^2-1|}{|a+1|} \geq \frac{|a^2-1|}{3}$ since $1 \leq a+1 \leq 3$; otherwise if $a^2 \geq 2 \Rightarrow a \geq \sqrt{2}$, we have $\sqrt{|(a-1)(a+1)|} \leq 3|a-1|$ as it holds that $\sqrt{a+1} \leq 3\sqrt{a-1}$. This last transition can be seen from the fact that $a \mapsto \sqrt{\frac{a+1}{a-1}}$ is monotonically decreasing on $a \in [\sqrt{2}, \infty)$, which completes the proof. $\quad\square$

18

C Proofs for Section 5

We specialize the discussion to recursive embeddings in this section, again leveraging a multi-resolution covering, with the first two lemmas being the building blocks for the ensuing main result.

Proof of Lemma 5. Invoking the JL moment condition for degree-2 Kronecker embedding $S^t \in \mathbb{R}^{m \times nm}$, for Tensor-SRHT matrix constructed from $S_i^t = v_i^{(1)} \otimes v_i^{(2)}$ at each level $t \in [d]$, with

$$m = \mathcal{O}\Big(\frac{1}{\epsilon^2} \log(\frac{1}{\delta}) \log^2(\frac{1}{\epsilon\delta})\Big)$$

the resulting compositional matrix $S \in \mathbb{R}^{m \times n^d}$ exhibits (1) $\mathbb{E}[\|Sx\|_2^2] = 1$ for all $\|x\| = 1$; (2) $\mathbb{E}[(\|Sx\|_2^2 - 1)^{\log(1/\delta)}] \leq (\frac{1}{e} \max(\epsilon/\sqrt{d}, \epsilon^2/d))^{\log(1/\delta)}$ using Theorem 1 of [2]. This implies via Markov's inequality, $\delta \in (0, 1)$,

$$\mathbb{P}(|\|Sx\|_2^2 - 1| \geq \max(\epsilon/\sqrt{d}, \epsilon^2/d)) \leq \delta$$

for any unit norm x. So for a set of cardinality p, taking a union bound we reach with probability at least $1 - e^{-\eta}$ for any $\eta > 0$, $|\|Sx\|_2^2 - 1| \leq \max(\epsilon, \epsilon^2)$ simultaneously for all p points on the unit sphere in the set provided (tilde hides poly-logs in $1/\epsilon$ and d)

$$m = \tilde{\mathcal{O}}\Big(\frac{d}{\epsilon^2} (\log^2(p) + \eta^2 \vee \eta)\Big).$$

This reduces the dependency on d from exponential to linear. $\square$

Proof of Lemma 6. The requirement entails that $p_l \leq 5d \cdot (2^{2^l})^{d-1} \cdot (2^{2^l})^2$ for each level of distortion $1 \leq l \leq L$ with $\epsilon_l = 2^l \epsilon$ and $\eta_l = l(\eta + 1) \geq 1$ (i.e., we only look at points engaging in neighboring scales). Union bounding over $L \asymp \lceil \log_2(nd) \rceil$ levels, using Lemma 5, we get with

$$m = \tilde{\mathcal{O}}\Big(\frac{d}{\epsilon_l^2} (\log^2(p_l) + \eta_l^2 \vee \eta_l)\Big)$$

which is $\tilde{\mathcal{O}}(d(d^2 + (1 + \eta)^2)/\epsilon^2)$ that all events required in Definition 3 hold with probability at least $1 - \sum_{l=1}^{L} \exp(-\eta_l) \geq 1 - \sum_{l=1}^{\infty} \exp(-l(\eta + 1)) \geq 1 - \exp(-\eta)$, as promised. $\square$

The calculation below taxes one's patience with algebra but is otherwise relatively mechanical.

Proof of Theorem 3. Let $\odot$ denote elementwise product and $\bar{x}_l^t := S^t(u_l^t \otimes \cdots S^d u_l^d) \in \mathbb{R}^m$ for each $l \in [L]$ and $t \in [d]$. Denote S_1^t and S_2^t the two $m \times n$ and $m \times m$ independent sketches at each level $t \in [d-1]$ and $S^d \in \mathbb{R}^{m \times n}$. Note that $\bar{x}_l^t$ is a recursive sketch of degree $d - t + 1$ tensor with all factors i belonging to l-th level approximation in $\{\bar{T}_l^i\}$. Furthermore, let $\tilde{L} = \max(0, \lfloor \log_2(1/\epsilon) \rfloor)$ such that for $l \leq \tilde{L}$, $\epsilon_l \leq 1$. Forming a telescoping sum and keeping in mind we are working under $\tilde{L} < L$ (the other case is discussed in Section D),

$|\|Sx\|_2^2 - \|x\|_2^2|$

$$\leq \sum_{l=1}^{\tilde{L}} \sum_{i=1}^{m} \langle S_{1,i}^1, u_l^1 \rangle^2 \langle S_{2,i}^1, \bar{x}_l^2 \rangle^2 - \langle S_{1,i}^1, u_{l-1}^1 \rangle^2 \langle S_{2,i}^1, \bar{x}_{l-1}^2 \rangle^2 - \sum_{k=1}^{d} \big(\|u_l^k\|_2^2 - \|u_{l-1}^k\|_2^2\big) \prod_{s=1}^{k-1} \|u_l^s\|_2^2 \prod_{s=k+1}^{d} \|u_{l-1}^s\|_2^2$$

$$+ \big|\|Sx\|_2^2 - \|S\tilde{x}_{\tilde{L}}\|_2^2\big| + \big|\|x\|_2^2 - \|\tilde{x}_{\tilde{L}}\|_2^2\big| + \big|\|S\tilde{x}_0\|_2^2 - \|\tilde{x}_0\|_2^2\big|$$

$$\leq \sum_{l=1}^{\tilde{L}} \Big(\sum_{i=1}^{m} (\langle S_{1,i}^1, u_l^1 \rangle^2 - \langle S_{1,i}^1, u_{l-1}^1 \rangle^2) \cdot \langle S_{2,i}^1, \bar{x}_{l-1}^2 \rangle^2 - \big(\|u_l^1\|_2^2 - \|u_{l-1}^1\|_2^2\big) \times \prod_{s=2}^{d} \|u_{l-1}^s\|_2^2 \Big) \qquad (*)$$

$$+ \sum_{l=1}^{\tilde{L}} \Big(\sum_{i=1}^{m} (\langle S_{2,i}^1, \bar{x}_l^2 \rangle^2 - \langle S_{2,i}^1, \bar{x}_{l-1}^2 \rangle^2) \langle S_{1,i}^1, u_l^1 \rangle^2 - \sum_{k=2}^{d} \big(\|u_l^k\|_2^2 - \|u_{l-1}^k\|_2^2\big) \times \prod_{s=1}^{k-1} \|u_l^s\|_2^2 \times \prod_{s=k+1}^{d} \|u_{l-1}^s\|_2^2 \Big)$$

$$+ \big|\|Sx\|_2^2 - \|S\tilde{x}_{\tilde{L}}\|_2^2\big| + \big|\|x\|_2^2 - \|\tilde{x}_{\tilde{L}}\|_2^2\big| + \big|\|S\tilde{x}_0\|_2^2 - \|\tilde{x}_0\|_2^2\big|$$

19

Term 1 + 2: Expand the second line above using $\bar{x}_l^2 = \langle S_1^2, u_l^2 \rangle \odot \langle S_2^2, \bar{x}_l^3 \rangle$ as

$$\leq \sum_{l=1}^{\tilde{L}} \Big(\sum_{i=1}^{m} \Big(\big\langle S_{2,i}^1, \langle S_1^2, u_l^2 \rangle \odot \langle S_2^2, \bar{x}_l^3 \rangle \big\rangle^2 - \big\langle S_{2,i}^1, \langle S_1^2, u_{l-1}^2 \rangle \odot \langle S_2^2, \bar{x}_{l-1}^3 \rangle \big\rangle^2 \Big) \langle S_{1,i}^1, u_l^1 \rangle^2 \quad (13)$$

$$- \sum_{k=2}^{d} \big(\|u_l^k\|_2^2 - \|u_{l-1}^k\|_2^2 \big) \times \prod_{s=1}^{k-1} \|u_l^s\|_2^2 \times \prod_{s=k+1}^{d} \|u_{l-1}^s\|_2^2 \Big) \quad (14)$$

$$\leq \sum_{l=1}^{\tilde{L}} \sum_{i=1}^{m} \Big(\big\langle S_{2,i}^1, \langle S_2^2, \bar{x}_{l-1}^3 \rangle \odot \langle S_1^2, u_l^2 \rangle \big\rangle^2 - \big\langle S_{2,i}^1, \langle S_2^2, \bar{x}_{l-1}^3 \rangle \odot \langle S_1^2, u_{l-1}^2 \rangle \big\rangle^2 \Big) \langle S_{1,i}^1, u_l^1 \rangle^2 \quad (15)$$

$$- \big(\|u_l^2\|_2^2 - \|u_{l-1}^2\|_2^2 \big) \times \|u_l^1\|_2^2 \times \prod_{s=3}^{d} \|u_{l-1}^s\|_2^2 \quad (16)$$

$$+ \sum_{l=1}^{\tilde{L}} \sum_{i=1}^{m} \Big(\big\langle \langle S_1^2, u_l^2 \rangle \odot \langle S_2^2, \bar{x}_l^3 \rangle, S_{2,i}^1 \big\rangle^2 - \big\langle \langle S_1^2, u_l^2 \rangle \odot \langle S_2^2, \bar{x}_{l-1}^3 \rangle, S_{2,i}^1 \big\rangle^2 \Big) \langle S_{1,i}^1, u_l^1 \rangle^2 \quad (17)$$

$$- \sum_{k=3}^{d} \big(\|u_l^k\|_2^2 - \|u_{l-1}^k\|_2^2 \big) \times \prod_{s=1}^{k-1} \|u_l^s\|_2^2 \times \prod_{s=k+1}^{d} \|u_{l-1}^s\|_2^2 \quad (18)$$

where we used the fact that for vectors a, b, c of same length, $\langle a, b \odot c \rangle = \langle a \odot c, b \rangle$. Notice that for a fixed l, (15) above is nothing but a recursive sketch on tensor

$$\|S(u_l^1 \otimes u_l^2 \otimes u_{l-1}^3 \otimes \cdots \otimes u_{l-1}^d)\|_2^2 - \|S(u_l^1 \otimes u_{l-1}^2 \otimes u_{l-1}^3 \otimes \cdots \otimes u_{l-1}^d)\|_2^2$$

therefore both $(*)$ and (15)-(16) above involve bounding distortion of the form below for which we can invoke Definition 3 and follow similar steps as the previous section to reach

$$\big| \|S(u_l^1 \otimes \cdots \otimes u_l^k \otimes \cdots \otimes u_{l-1}^d)\|_2^2 - \|S(u_l^1 \otimes \cdots \otimes u_{l-1}^k \otimes \cdots \otimes u_{l-1}^d)\|_2^2$$

$$- \|\mathrm{vec}(u_l^1 \otimes \cdots \otimes u_l^k \otimes \cdots \otimes u_{l-1}^d)\|_2^2 + \|\mathrm{vec}(u_l^1 \otimes \cdots \otimes u_{l-1}^k \otimes \cdots \otimes u_{l-1}^d)\|_2^2 \big|$$

$$\leq \big| \|S(u_l^1 \otimes \cdots \otimes u_l^k - u_{l-1}^k \otimes \cdots \otimes u_{l-1}^d)\|_2^2 - \|u_l^1 \otimes \cdots \otimes u_l^k - u_{l-1}^k \otimes \cdots \otimes u_{l-1}^d\|_2^2 \big|$$

$$+ 2 \big| (u_l^1 \otimes \cdots \otimes u_l^k - u_{l-1}^k \otimes \cdots \otimes u_{l-1}^d)^\top (S^\top S - I)(u_l^1 \otimes \cdots \otimes u_l^k \otimes \cdots \otimes u_{l-1}^d) \big|$$

$$\leq 2^l \epsilon \cdot 8 \cdot \mathrm{dist}(u^k, \bar{T}_{l-1}^k) + 2^l \epsilon \cdot 4 \cdot \mathrm{dist}(u^k, \bar{T}_{l-1}^k).$$

Proceeding by unfolding (17)-(18) above in a similar fashion, and summing over $l \in [\tilde{L}]$ gives

$$12\sqrt{2}\epsilon \sum_{k=1}^{d} \sum_{l=0}^{\tilde{L}-1} 2^l \mathrm{dist}(u^k, \bar{T}_l^k) \leq 12\sqrt{2}\epsilon \cdot \sum_{k=1}^{d} \gamma_1(T^k).$$

The derivation for the rest terms mirrors that from the previous section so we will be terse.

Term 3: For the third term, we note that $\|S(x - \tilde{x}_L)\|_2 \lesssim (1 + 2^{L/\alpha}\epsilon)$ for any rank-1 tensor x under consideration. Now recall $\tilde{x}_{\tilde{L}} = \bar{x}_{\tilde{L}}^1$ in our notation,

$$\big| \|Sx\|_2^2 - \|S\tilde{x}_{\tilde{L}}\|_2^2 \big| \leq \big| \|Sx\|_2 - \|S\tilde{x}_{\tilde{L}}\|_2 \big|^2 + 2 \cdot \big| \|Sx\|_2 - \|S\tilde{x}_{\tilde{L}}\|_2 \big| \cdot \|S\tilde{x}_{\tilde{L}}\|_2$$

$$\lesssim 64\epsilon^2 \cdot \left(\sum_{k=1}^{d} \gamma_1(T^k) \right)^2 + 32\epsilon \cdot \sum_{k=1}^{d} \gamma_1(T^k).$$

Term 4: Analogous to the previous part, we have

$$\big| \|x\|_2^2 - \|\tilde{x}_{\tilde{L}}\|_2^2 \big| \leq \big| \|x\|_2 - \|\tilde{x}_{\tilde{L}}\|_2 \big|^2 + 2 \cdot \big| \|x\|_2 - \|\tilde{x}_{\tilde{L}}\|_2 \big| \cdot \|\tilde{x}_{\tilde{L}}\|_2$$

$$\lesssim 4\epsilon^2 \left(\sum_{k=1}^{d} \gamma_1(T^k) \right)^2 + 4\epsilon \sum_{k=1}^{d} \gamma_1(T^k).$$

Term 5: Directly invoking part 2 of Definition 3 gives

$$\big| \|S\tilde{x}_0\|_2^2 - \|\tilde{x}_0\|_2^2 \big| \leq \max(\epsilon, \epsilon^2).$$

20

The finishing touch is done by noting $\sum_{k=1}^{d} \gamma_1(T^k) \geq d \geq 1$, assembling the pieces,

$$\left|\|Sx\|_2^2 - \|x\|_2^2\right| \lesssim \epsilon^2 \cdot \left(\sum_{k=1}^{d} \gamma_1(T^k)\right)^2 + \epsilon \cdot \sum_{k=1}^{d} \gamma_1(T^k) + \max(\epsilon, \epsilon^2)$$

$$\lesssim \max\left\{\epsilon \cdot \sum_{k=1}^{d} \gamma_1(T^k),\ \epsilon^2 \cdot \left(\sum_{k=1}^{d} \gamma_1(T^k)\right)^2\right\} + \max(\epsilon, \epsilon^2)$$

hence a change of variable $\epsilon \mapsto \frac{\epsilon}{\sum_{k=1}^{d} \gamma_1(T^k)}$ will make the stars align. Invoking Lemma 6 we end up with the sample complexity

$$m = \tilde{\mathcal{O}}\left(d(\sum_{k=1}^{d} \gamma_1(T^k))^2 \cdot (d^2 + (1+\eta)^2)/\epsilon^2\right)$$

for tensors with degree $d \geq 2$. $\qquad\square$

D Embedding Distortion when $\tilde{L} > L$

We will only use the Multi-resolution Embedding Property Definition 3 in the argument below, making it valid extension for both Theorem 2 and Theorem 3. We decompose:

$$\left|\|Sx\|_2^2 - \|x\|_2^2\right| \leq \sum_{l=1}^{L}\left|\|S\tilde{x}_l\|_2^2 - \|\tilde{x}_l\|_2^2\right| - \left|\|S\tilde{x}_{l-1}\|_2^2 - \|\tilde{x}_{l-1}\|_2^2\right|$$
$$+ \left|\left|\|Sx\|_2^2 - \|x\|_2^2\right| - \left|\|S\tilde{x}_L\|_2^2 - \|\tilde{x}_L\|_2^2\right|\right| + \left|\|S\tilde{x}_0\|_2^2 - \|\tilde{x}_0\|_2^2\right|\ .$$

The proofs in the previous parts have already taught us that the first term is $\lesssim \epsilon \cdot \sum_{k=1}^{d} \gamma_\alpha(T^k)$ by simply recalling that Definition 3 holds up until level L and since $L < \tilde{L}$, all the distortion take the first term ϵ_l instead of ϵ_l^2 for $l \in [L]$. The last term is also easily bounded by $\max(\epsilon, \epsilon^2)$ as before. This leaves us with the middle two terms. For this, note that both x and $\tilde{x}_L$ are rank-1 tensors, the inner product $x^\top \tilde{x}_L$ denotes $\prod_{k=1}^{d} \langle u^k, u_L^k \rangle$ here,

$$\left|\left|\|Sx\|_2^2 - \|x\|_2^2\right| - \left|\|S\tilde{x}_L\|_2^2 - \|\tilde{x}_L\|_2^2\right|\right| \leq \left|(\|Sx\|_2^2 - \|x\|_2^2) - (\|S\tilde{x}_L\|_2^2 - \|\tilde{x}_L\|_2^2)\right|$$
$$\leq \left|\|Sx - S\tilde{x}_L\|_2^2 - \|x - \tilde{x}_L\|_2^2\right| + 2\left|(Sx - S\tilde{x}_L)^\top S\tilde{x}_L - (x^\top \tilde{x}_L - \tilde{x}_L^\top \tilde{x}_L)\right|.\quad (*)$$

We only sketch the calculation and omit the details below as it is largely similar to earlier ones. Using polarization identity for the second term and telescoping over the degree, it amounts to looking at the distortion of S acting on $\cdots \otimes u^k - u_L^k \otimes \cdots,\ \cdots \otimes \frac{u^k - u_L^k}{\|u^k - u_L^k\|_2} + u_L^k \otimes \cdots$ and $\cdots \otimes \frac{u^k - u_L^k}{\|u^k - u_L^k\|_2} - u_L^k \otimes \cdots$, for which we can simply delegate Corollary 1 at level L with the first term in the $\max$ for the job (since $L < \tilde{L}$). After some simplification, one would reach that $(*) \lesssim \epsilon \cdot \sum_{k=1}^{d} 2^{L/\alpha}\mathrm{dist}(u^k, \bar{T}_L^k) \lesssim \epsilon \cdot \sum_{k=1}^{d} \gamma_\alpha(T^k)$. Merging with other parts,

$$\left|\|Sx\|_2^2 - \|x\|_2^2\right| \lesssim \epsilon \cdot \sum_{k=1}^{d} \gamma_\alpha(T^k) + \max(\epsilon, \epsilon^2),$$

from which it should become evident that the same distortion conclusion holds in this case as well.

E Additional Details for Section 6

E.1 Signal Recovery

Proof of Lemma 7. Let the optimizer of (8) be $z_*^j = t_*^j + u^j$ for all $j \in [d]$, where t_*^j denotes the error from the true unknown u^j. The constraint entails that t_*^j verifies the following condition:

$$t_*^j \in \mathcal{D}(f_j, u^j)\, \forall j \in [d] \text{ and } S(u^1 + t_*^1 \otimes \cdots \otimes u^d + t_*^d) = y\,,\ \|u^j + t_*^j\|_2 = 1\,.$$

21

Introduce the shorthand $\mathcal{W}^j := (u^j + \mathcal{D}(f_j, u^j)) \cap \mathbb{S}^{n-1}$. Using the definition of the restricted singular value and the stated assumption,

$$\sigma_{\min}(S; \mathcal{W}^1 \times \cdots \times \mathcal{W}^d) \geq 1 - \epsilon > 0.$$

Identifying $z_*^1 \otimes \cdots \otimes z_*^d$ as belonging to the set $\mathcal{W}^1 \times \cdots \times \mathcal{W}^d$, the restricted strong convexity condition allows us to conclude the uniqueness claim. $\qquad\square$

Lemma 10. *We have $\gamma_{2/d}((u^i + \mathcal{D}(f_i, u^i)) \cap \mathbb{S}^{n-1}) \lesssim \gamma_{2/d}(\mathcal{D}(f_i, u^i) \cap \mathbb{S}^{n-1})$ for all $i, d \geq 1$, and f_i convex. In the above, $\lesssim$ hide d factor.*

Proof. For any $v \in (u^i + \mathcal{D}(f_i, u^i)) \cap \mathbb{S}^{n-1}$, where the sum is the Minkowski sum, we can write $v = \frac{u^i + v'}{\|u^i + v'\|}$ for some $v' \in \mathcal{D}(f_i, u^i)$. The set on the RHS of the theorem statement consists of $\{g/\|g\| : g \in \mathcal{D}(f_i, u^i)\}$. Let $v' = \alpha \cdot u^i + \beta \cdot p$ for $\beta \neq 0$ and some $p \perp u^i$ where $\|p\| = \|u^i\| = 1$, in which case

$$\frac{u^i + v'}{\|u^i + v'\|} = \frac{1}{\|u^i + v'\|} u^i + \frac{1}{\|u^i + v'\|} v'$$

$$= \frac{1+\alpha}{\|(1+\alpha)u^i + \beta \cdot p\|} u^i + \frac{\beta}{\|(1+\alpha)u^i + \beta \cdot p\|} p$$

$$= \frac{1+\alpha}{\sqrt{(1+\alpha)^2 + \beta^2}} u^i + \frac{\beta}{\sqrt{(1+\alpha)^2 + \beta^2}} p$$

therefore we can see that v is a linear combination of u^i and p. We distinguish between cases: If $\alpha = 0$, then $p \in \{g/\|g\|\}_{g \in \mathcal{D}(f_i, u^i)}$ so $v \in \mathrm{conv}(u^i \cup \{g/\|g\|\}_{g \in \mathcal{D}(f_i, u^i)})$. If $\alpha \neq 0$, then $u^i \in \{g/\|g\|\}_{g \in \mathcal{D}(f_i, u^i)}$ as well as $p \in \{g/\|g\|\}_{g \in \mathcal{D}(f_i, u^i)}$, (since f_i is convex, the descent cone $\mathcal{D}(f_i, u^i)$ is a convex cone). Therefore the resulting vector v should also be a member of the RHS $\{g/\|g\|\}_{g \in \mathcal{D}(f_i, u^i)}$. It is evident that if $\beta = 0, \alpha \neq 0$, i.e., v' parallel u^i, the same inclusion holds. Therefore up to constants, the set on the LHS can be expressed as belonging to a union of itself with its shifted version.

Now using Lemma 2.1 of [22] – the result is stated with γ_α where $\alpha \geq 1$, but one could check it holds for any $\alpha > 0$ up to factor depending on d. Let $T_1 = \mathcal{D}(f_i, u^i) \cap \mathbb{S}^{n-1}, T_2 = u^i + (\mathcal{D}(f_i, u^i) \cap \mathbb{S}^{n-1})$,

$$\gamma_\alpha(T_1 \cup T_2) \lesssim \mathrm{diam}(T_1 \cup T_2) + \gamma_\alpha(T_1) + \gamma_\alpha(T_2) \lesssim \gamma_\alpha(T_1)$$

since $\|u^i\| = 1$ is bounded by assumption, γ-functionals are translation-invariant, and diam is upper bounded by γ_α. $\qquad\square$

Example 2. *Another example comes from signals taking quantized values (e.g, binary vectors). In this case, we may choose the regularizer $f_j = \|\cdot\|_\infty$ when $u^j \in \{\pm 1\}^n / \sqrt{n}$. Since $\bar{\mathcal{D}}(\|\cdot\|_\infty, u^j) = \{s \colon s_i \cdot u_i^j \leq 0 \; \forall i \in [n], \|s\|_2 = 1\}$ for a binary vector u^j, and the cone can be confirmed to be self-dual, the calculation in [8] suggests*

$$w^2(\bar{\mathcal{D}}(\|\cdot\|_\infty, u^j)) \leq n/2.$$

Moreover the covering number is that of $1/2^n$ of the unit sphere, using $N(\mathbb{S}^{n-1}, sB_2^n) \leq (3/s)^n$,

$$\gamma_{2/d}(\bar{\mathcal{D}}(\|\cdot\|_\infty, u^j)) \lesssim \int_0^1 (\log N(\bar{\mathcal{D}}(\|\cdot\|_\infty, u^j), sB_2^n))^{d/2} \, ds$$

$$\lesssim C_{2/d} \int_0^1 (n \log(1/s))^{d/2} \, ds \leq C'_{2/d} \cdot n^{d/2}.$$

Therefore with sample complexity (not concerned about the d^d factor) $m = \mathcal{O}\left(n^d (1+\eta)^d\right)$, one would be able to recover with high probability.

If one is interested in optimizing (8) using gradient information, the prescribed concentration in Theorem 2 will come in handy as well, as one needs to analyze quantities of type:

$$\sup_{\{z^j \colon f_j(z^j) \leq R_j\}, h \in \mathbb{S}^{n-1}} \left| \frac{1}{m} \sum_{i=1}^m \prod_{k \neq j \in [d]} \langle v_i^j, z^j \rangle \langle v_i^k, h \rangle - \mathbb{E}\left[\prod_{k \neq j \in [d]} \langle v_i^j, z^j \rangle \langle v_i^k, h \rangle \right] \right|.$$

The sampling matrix S, of course, can come from more structured / fast-multiply-equipped random ensembles apart from subgaussian factors.

22

E.2 Optimization

Example 3. *For a concrete example, suppose we have a-priori knowledge that the factors are smooth ($\|Du^i\|_0 \leq k, \|u^i\|_2 = 1$), in which case picking $T^i = \{s^i \colon \|Ds^i\|_1 \leq 4\sqrt{k}, \|s^i\|_2 \leq 1\}$ for*

$$D = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ 0 & 0 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \\ 0 & 0 & 0 & \cdots & 0 & -1 \end{bmatrix} \in \mathbb{R}^{n \times n}$$

the 1D total variation regularization becomes a natural choice. To carry out the program, it remains to bound the covering number of this set. For this, the work of [6] showed that

$$w_2(T^i) \asymp \gamma_2(T^i) \asymp (nk)^{1/4}\sqrt{\log(n)}\,.$$

A short calculation together with Theorem 3 allow us to reach that $m = \tilde{\mathcal{O}}(d^5 nk)$ ensures the solution to (10) will identify the correct factors.

Other examples could be signals taking block sparse structure where a $\|\cdot\|_{\ell_1/\ell_2}$ may be appropriate.

F Tightness of the Embedding Dimension

We contemplate on lower bounds for the embedding dimension and provide evidence for the row-wise-tensored sketch considered in Section 4 in this section.

In the case of finite set, [2] used the key ingredient of tight moment bound which states for mean-0 i.i.d random variables where $\|X_i\|_p \asymp p^\alpha$, it holds that $\|\sum_{i=1}^n X_i\|_p \asymp \max\{2^\alpha\sqrt{pn}, (n/p)^{1/p}p^\alpha\}$ for all $2 \leq p \leq 2n$. With minor massaging, one can extract from their result that for Tensor-Rademacher, the embedding dimension has to scale as $m \gtrsim (\log p)^d$ for simultaneously preserving the norms of p^d points. Close examination of their proof of Theorem 4 in fact reveals that the only critical assumptions responsible for such scaling are (1) each factor is independent; and (2) has $\|\langle v_i^k, h\rangle\|_p \asymp \sqrt{p}$ for any $\|h\|_2 = 1$, therefore similar conclusion holds for e.g., independent Gaussian factors. In the case of Tensor-SRHT, [3] showed that one needs at least $m \gtrsim (\log p)^d$ as well. Compared with our Theorem 2, this is tight, since we will have each factor belonging to a set of cardinality p, therefore $\gamma_2(T^i)^2 = \log(p), a = p, b = o(1)$, manifesting the inevitability of exponential dependence on d for this sketch.

On the occasion of unit sphere $T^i = \mathbb{S}^{n-1}$, consider the case when each $\{v_i^k\}_{i\in[m],k\in[d]}$ is independent random vector uniform on the sphere of radius $\sqrt{n}$. This closely resembles an n-dimensional standard Gaussian in high dimension ($\sigma = 1$). Now let every $u^k = v_1^k/\sqrt{n} \in \mathbb{S}^{n-1}$ for $k \in [d]$. Since random vectors on the unit sphere are almost orthogonal to each other (i.e., $\langle v_i^k, u^k\rangle = o(1)$ for $i \neq 1$),

$$\frac{1}{m}\sum_{i=1}^m \prod_{k=1}^d \langle v_i^k, u^k\rangle^2 \approx \frac{1}{m}\left(\frac{n^2}{n}\right)^d\,.$$

Therefore for this particular example, for the quantity to be $o(1)$, we need $m \gtrsim n^d$. Putting side-by-side with our Theorem 2, the sample complexity $m = \mathcal{O}(n^d)$ is sharp as $\gamma_2(\mathbb{S}^{n-1})^2 = n, b = n, a = o(1)$ in this case.

23