New Bounds for Matrix Multiplication: from Alpha to Omega

Virginia Vassilevska Williams* Yinzhan Xu† Zixuan Xu‡ Renfei Zhou§

Abstract

The main contribution of this paper is a new improved variant of the laser method for designing matrix multiplication algorithms. Building upon the recent techniques of [Duan, Wu, Zhou, FOCS 2023], the new method introduces several new ingredients that not only yield an improved bound on the matrix multiplication exponent ω, but also improve the known bounds on rectangular matrix multiplication by [Le Gall and Urrutia, SODA 2018].

In particular, the new bound on ω is

$$\omega \leq 2.371552 \text{ (improved from } \omega \leq 2.371866).$$

For the dual matrix multiplication exponent α defined as the largest α for which $\omega(1, \alpha, 1) = 2$, we obtain the improvement

$$\alpha \geq 0.321334 \text{ (improved from } \alpha \geq 0.31389).$$

Similar improvements are obtained for various other exponents for multiplying rectangular matrices.

1 Introduction

Matrix multiplication is arguably the most basic linear algebraic operation, with plentiful applications throughout computer science and beyond. Its algorithmic complexity has been studied for many decades. In 1969 a breakthrough result by Strassen [31] showed that $n \times n$ matrices can be multiplied faster than the naive cubic time algorithm. Since then there has been an explosion of results obtaining lower and lower bounds on the exponent ω defined as the smallest constant such that for all $\varepsilon > 0$, $n \times n$ matrices can be multiplied using $O(n^{\omega+\varepsilon})$ arithmetic operations (additions, subtractions, multiplications and divisions; this is called the arithmetic circuit model of computation). In recent years, the bound $\omega < 2.373$ has been obtained [33, 15, 23, 4]. A new paper by Duan, Wu and Zhou [17] shows that $\omega < 2.3719$.

The dream bound would be $\omega = 2$, implying a near-linear time algorithm for multiplying matrices. Unfortunately, a series of works [6, 2, 10, 1, 3, 8, 5, 9] has shown that the known techniques for multiplying matrices cannot achieve $\omega = 2$.

All work on matrix multiplication since 1986 [32, 33, 15, 23, 4, 17] has used various variants of the so-called *laser method*. The strongest limitation result known for the laser method and its variants [6] is that such techniques cannot show that $\omega < 2.3078$.

The limitation results could mean that radically new methods need to be produced to make big strides. Yet, even if one stays within the laser method framework, it is still an intriguing question: *how close can we get to the* 2.3078 *barrier bound?*

In many applications of matrix multiplication, one needs to multiply *rectangular* matrices: $n^a \times n^b$ by $n^b \times n^c$, where a, b, c are different. Here one defines $\omega(a, b, c)$ to be the exponent for which matrix products of such dimensions can be multiplied in $O(n^{\omega(a,b,c)+\varepsilon})$ time for all $\varepsilon > 0$, in the arithmetic circuit model of computation.

For instance, in the study of All-Pairs Shortest Paths (APSP) in unweighted directed graphs [34], the complexity of APSP depends on the value μ which is defined as the real number satisfying the equation $\omega(1, \mu, 1) = 1 + 2\mu$. The same value is needed for the best known algorithms for computing minimum witnesses of Boolean Matrix Multiplication [14], for All-Pairs Bottleneck Paths in node-weighted graphs [30] and other problems.

*Massachusetts Institute of Technology. virgi@mit.edu. Supported by NSF Grants CCF-2129139 and CCF-2330048 and BSF Grant 2020356.

†Massachusetts Institute of Technology. xyzhan@mit.edu. Partially supported by NSF Grants CCF-2129139 and CCF-2330048 and BSF Grant 2020356.

‡Massachusetts Institute of Technology. zixuanxu@mit.edu.

§Institute for Interdisciplinary Information Sciences, Tsinghua University. zhourf20@mails.tsinghua.edu.cn.

In the work on k-clique detection, the value of $\omega(1,2,1)$ is important, as it is known [18] that 4-cliques in n-node graphs can be detected in $O(n^{\omega(1,2,1)+\varepsilon})$ time for any $\varepsilon > 0$. Moreover, if $\omega(1,2,1) < 3.16$, this would improve the known algorithms for k-clique detection for all $k \geq 8$ [27].

A final value of interest is α, the largest constant so that $\omega(1,\alpha,1) = 2$, first studied by Coppersmith [11, 12]. If $\omega = 2$, then $\alpha = 1$. So one can view the goal of increasing α as another way to attempt to prove that $\omega = 2$.

The best bounds on rectangular matrix multiplication to date are given by Le Gall and Urrutia [25], which improved upon [11, 12, 20, 21, 22]. For the values listed above, the bounds obtained by [25] are as follows: $\mu < 0.5286$, $\omega(1,2,1) < 3.25164$ and $\alpha < 0.31389$.

The goal of this paper is to obtain better bounds on ω, α, μ and rectangular matrix multiplication in general.

1.1 Our results.

The main result of this paper is a new improved variant of the laser method for designing matrix multiplication algorithms. Applying the new method, we obtain improved bounds for both square and rectangular matrix multiplication.

In particular, we show that $\alpha > 0.321334$ (improving upon the previous bound 0.31389), $\mu < 0.527661$ (improving upon the previous bound 0.5286) and $\omega(1,2,1) < 3.250385$ (improving upon 3.25164).

As a consequence, Zwick's algorithm for APSP in directed unweighted graphs (and several other algorithms, e.g., minimum witnesses for Boolean Matrix Multiplication [14] and All-Pairs Bottleneck Paths in node-weighted graphs [30]) runs in $O(n^{2.527661})$ time and 4-cliques can be found in $O(n^{3.250385})$ time.

For many other bounds on rectangular matrix multiplication, see Table 1.

Table 1: Our bounds on $\omega(1,\kappa,1)$ by analyzing the fourth power of the CW tensor compared to the best previous bounds. The previous bound for $\kappa = 1$ comes from [17]'s eighth-power analysis, while all other entries come from [25].

κ	upper bound on $\omega(1,\kappa,1)$	previous bound on $\omega(1,\kappa,1)$
0.321334	2	N/A
0.33	2.000100	2.000448
0.34	2.000600	2.001118
0.35	2.001363	2.001957
0.40	2.009541	2.010314
0.45	2.023788	2.024801
0.50	2.042994	2.044183
0.527661	2.055322	N/A
0.55	2.066134	2.067488
0.60	2.092631	2.093981
0.65	2.121734	2.123097
0.70	2.153048	2.154399

κ	upper bound on $\omega(1,\kappa,1)$	previous bound on $\omega(1,\kappa,1)$
0.75	2.186210	2.187543
0.80	2.220929	2.222256
0.85	2.256984	2.258317
0.90	2.294209	2.295544
0.95	2.332440	2.333789
1.00	**2.371552**	2.371866
1.10	2.452056	2.453481
1.20	2.535063	2.536550
1.50	2.794941	2.796537
2.00	3.250385	3.251640
2.50	3.720468	3.721503
3.00	4.198809	4.199712

Independent Work. Independently, Le Gall [24] also obtained bounds on rectangular matrix multiplication, improving over [25]. His method generalizes the approach of [17] to rectangular matrices. For technical reasons, the bound on ω produced by his method does not match the bound in [17]. In comparison, our method is not only a generalization of [17] to rectangular matrices, but also an improvement. As a result, our bounds are better than the bounds in [24].

2 Technical Overview

2.1 Overview of previous work.

For positive integers a, b, c, the $a \times b \times c$ matrix multiplication tensor $\langle a, b, c\rangle$ is a tensor over the variable sets $\{x_{ij}\}_{i\in[a],j\in[b]}, \{y_{jk}\}_{j\in[b],k\in[c]}, \{z_{ki}\}_{k\in[c],i\in[a]}$ defined as the tensor computing the $a \times c$ product matrix $\{z_{ki}\}_{k\in[c],i\in[a]}$ of an $a \times b$ matrix $\{x_{ij}\}_{i\in[a],j\in[b]}$ and a $b \times c$ matrix $\{y_{jk}\}_{j\in[b],k\in[c]}$.[1] Specifically,

[1] For integer $n \geq 0$, the notation $[n]$ denotes $\{1, \ldots, n\}$.

$\langle a, b, c\rangle$ can be written as the following trilinear form

$$\langle a, b, c\rangle = \sum_{i\in[a]} \sum_{j\in[b]} \sum_{k\in[c]} x_{ij} y_{jk} z_{ki}.$$

It is not hard to check that $\langle a, b, c\rangle \otimes \langle d, e, f\rangle = \langle ad, be, cf\rangle$. For a tensor T, let $R(T)$ denote the tensor rank of T and the matrix multiplication exponent ω is defined as

$$\omega \coloneqq \inf_{q\in\mathbb{N},\, q\geq 2} \log_q R(\langle q, q, q\rangle).$$

It is hard to directly bound the tensor rank of $\langle q, q, q\rangle$ in general, so current approaches to bounding ω utilize Schönhage's asymptotic sum inequality [29], which states that if one can bound the asymptotic rank of a direct sum of matrix multiplication tensors, where the asymptotic rank $\widetilde{R}(T)$ of a tensor T is defined as

$$\widetilde{R}(T) \coloneqq \lim_{n\to\infty} R(T^{\otimes n})^{1/n},$$

then one can get a bound on ω. More specifically, we recall the asymptotic sum inequality as follows.

THEOREM 2.1 (Asymptotic sum inequality [29]). *For positive integers $r > m$ and a_i, b_i, c_i for $i \in [m]$, if*

$$\widetilde{R}\left(\bigoplus_{i=1}^{m} \langle a_i, b_i, c_i\rangle\right) \leq r,$$

then $\omega \leq 3\tau$ where $\tau \in [2/3, 1]$ is the solution to the equation

$$\sum_{i=1}^{m} (a_i \cdot b_i \cdot c_i)^{\tau} = r.$$

Schönhage's asymptotic sum inequality gave rise to the following approach to bounding ω: start with a tensor T whose asymptotic rank $\widetilde{R}(T)$ is easy to bound. Consider $T^{\otimes n}$ for some n sufficiently large and we want to transform $T^{\otimes n}$ into a direct sum of matrix multiplication tensors whose asymptotic rank is upper bounded by the asymptotic rank of $\widetilde{R}(T^{\otimes n}) = \widetilde{R}(T)^n$. The common ways of doing such transformation is via zeroing-out, i.e., setting some variables in $T^{\otimes n}$ to zero, or the more general degeneration, whose definition is deferred to Section 3. Then we can apply the asymptotic sum inequality to get a bound on ω. Observe that if $T^{\otimes n}$ can be degenerated into $\bigoplus_{i=1}^{m} \langle a_i, b_i, c_i\rangle$, then for a fixed τ, we want to maximize the value of $\sum_{i=1}^{m} (a_i \cdot b_i \cdot c_i)^{\tau}$. This gives a notion of the "matrix multiplication value" of a tensor T that we want to maximize. Then notice that a lower bound on the value of $T^{\otimes n}$ would directly imply an upper bound on ω via the asymptotic sum inequality. It still remains unknown how to get the best possible bound on ω via a tensor power $T^{\otimes n}$, but the laser method provides one way to give a nontrivial bound.

Laser method. Let T be a tensor over three sets of variables X, Y, Z. For positive integers s_X, s_Y, s_Z, let $X = \bigsqcup_{i\in[s_X]} X_i, Y = \bigsqcup_{j\in[s_Y]} Y_j$ and $Z = \bigsqcup_{k\in[s_Z]} Z_k$ be partitions of the X-, Y-, Z-variable sets into s_X, s_Y, s_Z parts respectively. Then T can be written as a sum of subtensors $\sum_{i,j,k} T_{i,j,k}$, where $T_{i,j,k}$ denotes the subtensor of T restricted to variables X_i, Y_j, Z_k.

Suppose for now that each subtensor $T_{i,j,k}$ is a matrix multiplication tensor. If T is a direct sum of matrix multiplication tensors, then we can apply Schönhage's asymptotic sum inequality [29] to obtain a bound on ω. However, T is a sum of $T_{i,j,k}$, not necessarily a direct sum.

The laser method [32] is devised to overcome this issue. First, we take the n-th tensor power of T for some large n, which is a tensor over variables X^n, Y^n, Z^n:

$$T^{\otimes n} = \sum_{I\in[s_X]^n} \sum_{J\in[s_Y]^n} \sum_{K\in[s_Z]^n} T_{I,J,K},$$

where

$$T_{I,J,K} = T_{I_1,J_1,K_1} \otimes T_{I_2,J_2,K_2} \otimes \cdots \otimes T_{I_n,J_n,K_n}.$$

We will refer to these three sets of variables as X-variables, Y-variables and Z-variables respectively. Because each $T_{i,j,k}$ is a matrix multiplication tensor and the tensor products of several $T_{i,j,k}$'s will still be matrix multiplication tensors, $T_{I,J,K}$ is a matrix multiplication tensor for any $I \in [s_X]^n, J \in [s_Y]^n, K \in [s_Z]^n$. For any $I \in [s_X]^n$, let X_I denote $X_{I_1} \times X_{I_2} \times \cdots \times X_{I_n}$, which is a subset of X^n. Similarly we define Y_J and Z_K. It is not difficult to see that $T_{I,J,K}$ is exactly the subtensor of $T^{\otimes n}$ when restricted to X_I, Y_J, Z_K. We call such subsets X_I, Y_J, Z_K *variable blocks*.

The goal of the laser method is to select some of the variable blocks X_I, Y_J or Z_K and zero out all of the variables in these blocks, i.e. "zero out the blocks", so that the remaining tensor is a direct sum of $T_{I,J,K}$'s.

The laser method specifies a distribution α over triples (i,j,k) where $i \in [s_X], j \in [s_Y], k \in [s_Z]$, so that for each $T_{I,J,K}$ that we want to keep in the direct sum, we require that

$$\left|\{t \in [n] \mid (I_t, J_t, K_t) = (i,j,k)\}\right| = \alpha(i,j,k) \cdot n. \tag{2.1}$$

If a subtensor $T_{I,J,K}$ satisfies (2.1), we say that it is *consistent* with the distribution α.

The distribution α induces the marginal distributions $\alpha_X, \alpha_Y, \alpha_Z$ on the X-, Y-, Z-variables over the indices $[s_X], [s_Y], [s_Z]$ respectively as follows. Let α_X, α_Y, α_Z be the marginal distributions of α on the three dimensions respectively, i.e.,

$$\begin{aligned}
\alpha_X(i) &= \sum_{j \in [s_Y], k \in [s_Z]} \alpha(i,j,k) \quad \forall i \in [s_X], \\
\alpha_Y(j) &= \sum_{i \in [s_X], k \in [s_Z]} \alpha(i,j,k) \quad \forall j \in [s_Y], \\
\alpha_Z(k) &= \sum_{i \in [s_X], j \in [s_Y]} \alpha(i,j,k) \quad \forall k \in [s_Z].
\end{aligned}$$

In the laser method, we zero out all X-variable blocks X_I that are not consistent with α_X (X_I is consistent with α_X if $|\{t \in [n] : I_t = i\}| = \alpha_X(i) \cdot n$ for every $i \in [s_X]$). We similarly zero out all Y-variable blocks Y_J that are not consistent with α_Y and Z-variable blocks Z_K that are not consistent with α_Z.

At this stage, a subtensor $T_{I,J,K}$ remains if X_I, Y_J and Z_K all remain. Thus, all remaining $T_{I,J,K}$'s are consistent with some distribution α' that induces the same marginal distributions $\alpha_X, \alpha_Y, \alpha_Z$, though α' might be different from α. The final stages of the laser method aim to keep a collection of independent subtensors $T_{I,J,K}$ and zero out the subtensors $T_{I,J,K}$ that are consistent with a distribution $\alpha' \neq \alpha$, using techniques such as hashing and greedy procedures. Eventually, the laser method obtains multiple independent copies of the tensor isomorphic to:

$$\mathcal{T} \coloneqq \bigotimes_{i,j,k} T_{i,j,k}^{\otimes \alpha(i,j,k) \cdot n}.$$

The Coppersmith-Winograd tensor CW_q. Prior works [13, 15, 33, 23, 4, 17] that obtained the best bounds on ω used the Coppersmith-Winograd tensor CW_q and its powers as the starting tensor T in the laser method. The Coppersmith-Winograd tensor CW_q for a nonnegative integer q is defined as

$$\mathrm{CW}_q \coloneqq x_0 y_0 z_{q+1} + x_0 y_{q+1} z_0 + x_{q+1} y_0 z_0 + \sum_{i=1}^{q} (x_0 y_i z_i + x_i y_0 z_i + x_i y_i z_0).$$

Observe that

$$\sum_{i=1}^{q} (x_0 y_i z_i + x_i y_0 z_i + x_i y_i z_0) \equiv \langle 1,1,q \rangle + \langle q,1,1 \rangle + \langle 1,q,1 \rangle,$$

so CW_q is the sum of six matrix multiplication tensors where the other three are copies of $\langle 1,1,1 \rangle$. Next, we describe the leveled partitions of CW_q and $\mathrm{CW}_q^{\otimes 2^\ell}$ that are crucial to our analysis. For simplicity, we denote $T^{(\ell)} \coloneqq \mathrm{CW}_q^{\otimes 2^{\ell-1}}$.

For $T^{(1)} = \mathrm{CW}_q$, its variable sets are partitioned into three parts

$$X^{(1)} = X_0^{(1)} \sqcup X_1^{(1)} \sqcup X_2^{(1)} = \{x_0\} \sqcup \{x_1, \dots, x_q\} \sqcup \{x_{q+1}\},$$
$$Y^{(1)} = Y_0^{(1)} \sqcup Y_1^{(1)} \sqcup Y_2^{(1)} = \{y_0\} \sqcup \{y_1, \dots, y_q\} \sqcup \{y_{q+1}\},$$
$$Z^{(1)} = Z_0^{(1)} \sqcup Z_1^{(1)} \sqcup Z_2^{(1)} = \{z_0\} \sqcup \{z_1, \dots, z_q\} \sqcup \{z_{q+1}\}.$$

Notice that under this partition, a constituent tensor $T^{(1)}_{i,j,k}$ is nonzero if and only if $i + j + k = 2$.

For $T^{(\ell)} = \mathrm{CW}_q^{\otimes 2^{\ell-1}}$ with variable sets $X^{(\ell)}, Y^{(\ell)}, Z^{(\ell)}$, the above partition on $T^{(1)}$ directly induces a partition on the variable sets $X^{(\ell)}, Y^{(\ell)}, Z^{(\ell)}$ where each part of the partition is indexed by a $\{0,1,2\}$-sequence of length $2^{\ell-1}$. Specifically, this gives the partition

$$X^{(\ell)} = \bigsqcup_{(\hat{i}_1, \hat{i}_2, \dots, \hat{i}_{2^{\ell-1}}) \in \{0,1,2\}^{2^{\ell-1}}} X^{(1)}_{\hat{i}_1} \otimes X^{(1)}_{\hat{i}_2} \otimes \cdots \otimes X^{(1)}_{\hat{i}_{2^{\ell-1}}}$$

for X-variables and analogous partitions for Y- and Z-variables.

One can use the laser method on these partitions. However, this would not yield an improved bound on ω from what one would get just by analyzing $T^{(1)}$. The reason behind the improvement obtained by analyzing higher powers of CW_q comes from the fact that we can consider the following coarsening of the above partition where the parts corresponding to sequences with the same sum are "merged" into a single part. More specifically, we have

$$X^{(\ell)} = \bigsqcup_{i=0}^{2^\ell} X_i^{(\ell)}, \qquad \text{where} \quad X_i^{(\ell)} \coloneqq \bigsqcup_{\substack{(\hat{i}_1, \hat{i}_2, \dots, \hat{i}_{2^{\ell-1}}) \in \{0,1,2\}^{2^{\ell-1}} \\ \sum_t \hat{i}_t = i}} X^{(1)}_{\hat{i}_1} \otimes X^{(1)}_{\hat{i}_2} \otimes \cdots \otimes X^{(1)}_{\hat{i}_{2^{\ell-1}}}.$$

We refer to this specific partition of $T^{(\ell)}$ as the *level-ℓ partition*. Note that we can also view this partition as obtained from coarsening the level-$(\ell-1)$ partition, i.e.,

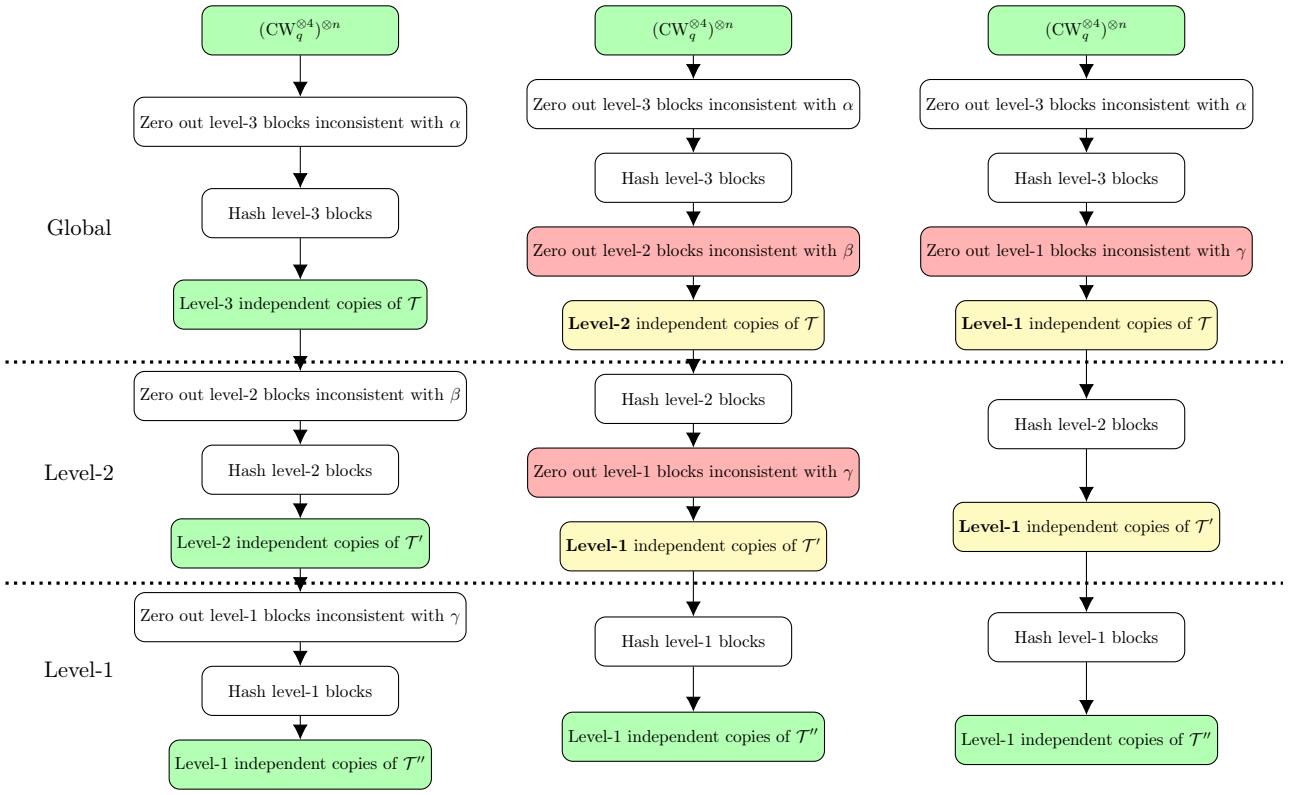
$$X_i^{(\ell)} = \bigsqcup_{\substack{0 \le i' \le i \\ 0 \le i', i - i' \le 2^\ell}} X_{i'}^{(\ell-1)} \otimes X_{i-i'}^{(\ell-1)}.$$

We can partition the variable sets $Y^{(\ell)}$ and $Z^{(\ell)}$ similarly. Then we use $T^{(\ell)}_{i,j,k}$ to denote the subtensor of $T^{(\ell)}$ restricted to the variable subsets $X_i^{(\ell)}, Y_j^{(\ell)}, Z_k^{(\ell)}$ and note that $T^{(\ell)}_{i,j,k}$ is nonzero if and only if $i + j + k = 2^\ell$. We call $T^{(\ell)}_{i,j,k}$ a level-ℓ *constituent tensor*, $X_i^{(\ell)}, Y_j^{(\ell)}, Z_k^{(\ell)}$ level-ℓ variable blocks, and we omit the superscript (ℓ) when ℓ is clear from context.

For $\ell > 1$, some level-ℓ constituent tensors $T^{(\ell)}_{i,j,k}$ are no longer matrix multiplication tensors, so each independent copy of $\mathcal{T} = \bigotimes_{i,j,k} \big(T^{(\ell)}_{i,j,k}\big)^{\otimes \alpha(i,j,k) \cdot n}$ may also no longer be a matrix multiplication tensor. To resolve this issue, prior works [13, 15, 33, 23, 4] use the laser method recursively to analyze $T_{i,j,k}$'s that are not matrix multiplication tensors.

The work of [17]. Consider the analysis on the tensor $T^{(\ell)}$ of the laser method. In previous approaches prior to the work of Duan, Wu and Zhou [17], one would first apply the laser method on $T^{(\ell)}$ to obtain multiple copies of $\mathcal{T} = \bigotimes_{i,j,k} \big(T^{(\ell)}_{i,j,k}\big)^{\otimes \alpha(i,j,k) \cdot n}$ which consists of level-ℓ constituent tensors $T^{(\ell)}_{i,j,k}$ and do not share level-ℓ variable blocks. Then for each $T^{(\ell)}_{i,j,k}$ that is not a matrix multiplication tensor, one would recursively apply the laser method to obtain multiple copies of some other tensors that are independent over level-$(\ell-1)$ variable blocks.

Recall that for a level-ℓ constituent tensor $T^{(\ell)}_{i,j,k}$, we can partition its variable set $X_i^{(\ell)}, Y_j^{(\ell)}, Z_k^{(\ell)}$ recursively into $\bigsqcup_{i'} X_{i'}^{(\ell-1)} \otimes X_{i-i'}^{(\ell-1)}$, $\bigsqcup_{j'} Y_{j'}^{(\ell-1)} \otimes Y_{j-j'}^{(\ell-1)}$ and $\bigsqcup_{k'} Z_{k'}^{(\ell-1)} \otimes Z_{k-k'}^{(\ell-1)}$ respectively. In the first recursive step, when applying the laser method on $T^{(\ell)}_{i,j,k}$, we take the n'-th tensor power $\big(T^{(\ell)}_{i,j,k}\big)^{\otimes n'}$ of $T^{(\ell)}_{i,j,k}$ for some large n' and specify a distribution β over triples $((i', i-i'), (j', j-j'), (k', k-k'))$ where $0 \le i' \le i$, $0 \le j' \le j$, $0 \le k' \le k$,



(a) [33, 15, 23, 4] (b) [17] (c) This work

Figure 1: High-level comparison between this work and prior works on $(\mathrm{CW}_q^{\otimes 4})^{\otimes n}$. Here, α is a distribution over level-3 constituent tensors, β is a collection of distributions over level-2 constituent tensors, and γ is a collection of distributions over level-1 constituent tensors.

and zero out all variables blocks that are not consistent with the marginal distributions induced by β. Therefore, in $T_{i,j,k}^{\otimes n'}$, only a subset of the level-$(\ell-1)$ variable blocks survive the above zeroing-out.

Now suppose we can move the above zeroing-out step earlier, say before we have independent copies of $\mathcal{T}$ when we first apply the laser method on $T^{(\ell)}$, then instead of keeping independent copies of $\mathcal{T}$, we only need to keep a subtensor $\mathcal{T}'$ of it, where $\mathcal{T}'$ is $\mathcal{T}$ after applying the above zeroing-out step. This leads to one of the key observations in [17]: we do not need to have copies of $\mathcal{T}'$ that are fully independent over the level-ℓ variable blocks. Instead, any two copies can share the same level-ℓ variable block as long as they do not share the same level-$(\ell-1)$ variable blocks that would survive the first zeroing-out in the recursive application of the laser method on the level-ℓ constituent tensors. As a result, we can potentially keep more independent copies of $\mathcal{T}'$, because of the relaxed constraints, and each copy $\mathcal{T}'$ would still be essentially as good as $\mathcal{T}$ for the purpose of the analysis because we are merely moving a later zeroing-out earlier. Because we are keeping more copies, by the asymptotic sum inequality, we will achieve a better bound for ω.

As illustrated in Fig. 1, consider $(\mathrm{CW}_q^{\otimes 4})^{\otimes n}$ and suppose α, β, γ are (collections of) distributions over level-3, level-2, level-1 constituent tensors respectively. In subfigure (a), works prior to [17] including [33, 4, 23] zero out level-3 blocks according to α and obtain level-3-independent[2] copies of $\mathcal{T}'$ before zeroing out level-2 blocks. As shown in subfigure (b), Duan et al. [17] moved the step of zeroing out level-2 blocks according to β earlier and only obtained level-2-independence as opposed to level-3-independence.

[2]We say several subtensors of $\mathrm{CW}_q^{\otimes N}$ are *level-ℓ-independent* if they do not share any level-ℓ variable block, and thus they are also *independent*.

It is not obvious how one can accomplish the above modification. Duan et al. [17] considered the notion of *split distributions*, which roughly measures how a level-ℓ block "splits" into level-$(\ell-1)$ blocks with respect to the recursive leveled partition. By observing the split distribution of a level-ℓ block, one gains some partial information about the level-$(\ell-1)$ blocks that allows the modification of zeroing out level-$(\ell-1)$ blocks inconsistent with β earlier. Ideally, one would hope to achieve this modification symmetrically over the X-, Y-, and Z-variables, i.e., allow the sharing of level-ℓ variable blocks in all three dimensions, but the method in [17] did not achieve that. Instead, their technique works when the multiple copies of $\mathcal{T}$ only share the same level-ℓ Z-variable block while each X- and Y-variable block needs to be contained in a unique level-ℓ subtensor. (More generally, their technique works when level-ℓ variable blocks are shared in exactly one of X-, Y-, Z-variables). In order to set up the tensor satisfying the required constraints, they need to zero out the Z-variable blocks asymmetrically with respect to the X- and Y-variables. It still remains an open question whether the techniques in [17] can be symmetrized over the three dimensions.

Another technical detail is that the obtained independent copies of tensors in [17] are not all necessarily full copies of $\mathcal{T}'$. That is, some variables of the independent tensors are zeroed out. This creates independent copies of $\mathcal{T}'$ but with some "holes". Because of the asymmetry of their method, such holes can only appear in Z-variables. In order to overcome this issue, they showed that, as long as the fraction of holes is small, and all holes are in Z-variables, one can degenerate a small number of independent copies of $\mathcal{T}'$ with holes to a full copy of $\mathcal{T}$. Prior to their work, Schönhage [29] also studied this problem of degenerating multiple independent copies of a tensor with holes to a full copy of the tensor. Schönhage's method applied to the case when two of the three dimensions can have holes, but it focuses only on matrix multiplication tensors.

2.2 Our improvements.

Complete split distribution. We take the observation of [17] one step further. The high-level idea is the following: instead of keeping copies of $\mathcal{T}$ that are independent over level-$(\ell-1)$ variable blocks, we keep copies of it that are independent over level-1 variable blocks. For $\ell > 1$, this should give more degrees of freedom and enable us to keep more copies of $\mathcal{T}$. As illustrated in Fig. 1 (c), we directly move the step of zeroing out level-1 blocks according to γ earlier and obtain level-1 independence as opposed in level-2 independence in [17].

To implement the above idea, we utilize the notion of *complete split distributions*, which can be viewed as an extension of the notion of split distributions used in [17]. Recall that in [17], a level-ℓ split distribution measures how a level-ℓ variable block splits into level-$(\ell-1)$ blocks. A level-ℓ complete split distribution measures how a level-ℓ block splits into level-1 variable blocks. Specifically, a level-1 block sequence of length $2^{\ell-1}\cdot n$ in $T^{(\ell)}$ can be viewed as n consecutive chunks of $\{0,1,2\}$-sequences each of length $2^{\ell-1}$, and we consider the proportion of each of these $3^{2^{\ell-1}}$ possible types of chunks in the n chunks. A level-ℓ complete split distribution is a distribution on these $3^{2^{\ell-1}}$ types of chunks, and a level-1 block sequence (and its corresponding level-1 variable block) is said to be consistent with a level-ℓ complete split distribution if the proportion of each type of chunks matches the corresponding probability specified in the complete split distribution.

Let $\beta_X, \beta_Y, \beta_Z$ be three level-ℓ complete split distributions, and let $T_{i,j,k}$ be a level-ℓ constituent tensor. We will consider the tensor $T^{\otimes n}_{i,j,k}[\beta_X, \beta_Y, \beta_Z]$, which is obtained from $T^{\otimes n}_{i,j,k}$ by zeroing out all level-1 X-, Y-, Z-variable blocks that are not consistent with $\beta_X, \beta_Y, \beta_Z$ respectively. We call this "enforcing the complete split distributions". In our recursive steps, we will analyze $T^{\otimes n}_{i,j,k}[\beta_X, \beta_Y, \beta_Z]$ instead of $T^{\otimes n}_{i,j,k}$.

Enforcing split distributions in all three dimensions. Dual et al. [17] only enforce their split distribution in one of the dimensions (the Z variables). In our method, we need to enforce complete split distributions in all three dimensions. Here we explain why.

First of all, when analyzing a level-ℓ constituent tensor $T^{\otimes n}_{i,j,k}$, [17] only consider split distributions, instead of complete split distributions. Every level-$(\ell-1)$ block sequence in $T^{\otimes n}_{i,j,k}$ can be viewed as a length-$(2n)$ sequence on $\{0,1,\ldots,2^{\ell-1}\}$. If we split the sequence to chunks of length 2, we obtain a length-n sequence of pairs in $\{0,1,\ldots,2^{\ell-1}\}^2$. The split distribution used in [17] essentially specifies the proportion of each type of pairs, and they zero out all level-$(\ell-1)$ variable blocks that are not consistent with the specified proportions.

Similar to what we discussed earlier, when enforcing the split distribution on the tensor $T^{\otimes n}_{i',j',k'}$ (or $T^{\otimes n}_{i-i',j-j',k-k'}$), the constraint becomes a constraint that enforces the proportion of each level-$(\ell-1)$ variable block in the level-$(\ell-1)$ variable blocks in $T^{\otimes n}_{i',j',k'}$. Since there is only one level-$(\ell-1)$ block in $T^{\otimes n}_{i',j',k'}$, either the whole tensor $T^{\otimes n}_{i',j',k'}$ satisfies the constraints, or it does not. Thus, the constraints of the split distribution do

not carry over to further recursion levels.

When analyzing each constituent tensor $T_{i,j,k}^{\otimes n}$, Duan et al. [17] aim to obtain some "symmetrized value" of $T_{i,j,k}$, similar to previous works [13, 15, 33, 23, 4]. As a result, when analyzing $T_{i,j,k}^{\otimes n}$, they apply their method multiple times to enforce a split distribution on each of the three possible dimensions, i.e., they can choose to share X-, Y-, or Z-variables depending on which application of their method it is. Still, the constraints of the split distribution do not carry over to the next recursion level as discussed in the previous paragraph. Thus, in their analysis, holes only appear in one of the dimensions.

However, when enforcing a complete split distribution, the constraints carry over to further recursion levels: say in the analysis for $T_{i,j,k}$ in some application of the method in the current level, we choose to enforce a complete split distribution on Z-variables. This constraint still has an effect on the next level. However, in the analysis at the next level, we can choose to enforce a complete split distribution on Y-variables instead. This creates constraints on the complete split distribution in two dimensions. In general, these constraints can appear in all three dimensions, and therefore, we need to handle holes in all three dimensions.

A technical issue. A technical issue arises if we enforce complete split distributions in three dimensions. We consider a simplified scenario where the support of the distribution β has size 1 to explain the issue. In other words, we aim to zero out $T_{i,j,k}^{\otimes n}$ into independent copies of $(T_{i',j',k'} \otimes T_{i-i',j-j',k-k'})^{\otimes n}$ for some i', j', k'. In this simplified scenario, if we do not enforce complete split distributions, we could rewrite $(T_{i',j',k'} \otimes T_{i-i',j-j',k-k'})^{\otimes n}$ equivalently as $T_{i',j',k'}^{\otimes n} \otimes T_{i-i',j-j',k-k'}^{\otimes n}$ by simply permuting the indices, and then recursively analyze $T_{i',j',k'}^{\otimes n}$ and $T_{i-i',j-j',k-k'}^{\otimes n}$ separately. Now with complete split distribution, this step becomes problematic. Suppose we are able to obtain independent copies of

$$\mathcal{T}_1 \coloneqq (T_{i',j',k'} \otimes T_{i-i',j-j',k-k'})^{\otimes n}[\beta_X, \beta_Y, \beta_Z],$$

for some $\beta_X, \beta_Y, \beta_Z$. Then in order to recursively analyze $\mathcal{T}_1$, we instead need a tensor

$$\mathcal{T}_2 \coloneqq \left(T_{i',j',k'}^{\otimes n}\left[\beta_X^{(L)}, \beta_Y^{(L)}, \beta_Z^{(L)}\right]\right) \otimes \left(T_{i-i',j-j',k-k'}^{\otimes n}\left[\beta_X^{(R)}, \beta_Y^{(R)}, \beta_Z^{(R)}\right]\right),$$

for some level-$(\ell-1)$ complete split distributions $\beta_X^{(L)}, \beta_Y^{(L)}, \beta_Z^{(L)}, \beta_X^{(R)}, \beta_Y^{(R)}, \beta_Z^{(R)}$.

Let us discuss how the above level-$(\ell-1)$ complete split distributions are related to $\beta_X, \beta_Y, \beta_Z$. To give some intuition, in each length-2^ℓ chunk of a level-1 block sequence in $\mathcal{T}_1$, the first half-chunk belongs to some $T_{i',j',k'}$, and the second half-chunk belongs to some $T_{i-i',j-j',k-k'}$. In $\mathcal{T}_2$, we permute the indices so that all the first half-chunks belonging to some $T_{i',j',k'}$ are put together in the first half of the resulting sequence, and all the second half-chunks belonging to some $T_{i-i',j-j',k-k'}$ are put together in the second half of the resulting sequence. If we enforce a level-ℓ complete split distribution β_X on a level-1 block sequence $\hat{I} \in \{0,1,2\}^{2^{\ell-1}}$ in $\mathcal{T}_1$, what would the permuted sequence look like? Let $\sigma_1, \sigma_2 \in \{0,1,2\}^{2^{\ell-2}}$ denote two length-$2^{\ell-2}$ chunks and let $\sigma_1 \circ \sigma_2$ denote their concatenation. Since $\hat{I}$ is consistent with β_X, $\hat{I}$ contains $\beta_X(\sigma_1 \circ \sigma_2) \cdot n$ chunks $\sigma_1 \circ \sigma_2$ for every σ_1, σ_2. For each of these chunks, σ_1 gets permuted to the first half of the permuted level-1 block sequence in $\mathcal{T}_2$, and σ_2 gets permuted to the second half of the permuted level-1 block sequence in $\mathcal{T}_2$. Summing over all σ_1, σ_2, it is not difficult to verify that

$$\beta_X^{(L)}(\sigma_1) = \sum_{\sigma_2} \beta_X(\sigma_1 \circ \sigma_2), \qquad \beta_X^{(R)}(\sigma_2) = \sum_{\sigma_1} \beta_X(\sigma_1 \circ \sigma_2).$$

In this sense, $\beta_X^{(L)}$ and $\beta_X^{(R)}$ can be viewed as two marginal distributions of β_X. This similarly holds for Y and Z.

One set of constraints we can add to make $\beta_X^{(L)}$ and $\beta_X^{(R)}$ always the two marginal distributions of β_X is $\beta_X = \beta_X^{(L)} \times \beta_X^{(R)}$, namely we enforce β_X to be the joint distribution of (independently distributed) $\beta_X^{(L)}$ and $\beta_X^{(R)}$. Similarly we can add the constraints $\beta_Y = \beta_Y^{(L)} \times \beta_Y^{(R)}$ and $\beta_Z = \beta_Z^{(L)} \times \beta_Z^{(R)}$.

However, even with these constraints, $\mathcal{T}_1$ might not necessarily be equivalent to $\mathcal{T}_2$. By the above reasoning, every level-1 block sequence in $\mathcal{T}_1$ is permuted into a level-1 block sequence in $\mathcal{T}_2$, but not all block sequences in $\mathcal{T}_2$ can be obtained this way. Intuitively, this is because joint distributions can determine marginal distributions, which means that, for instance, $\beta_X^{(L)} \times \beta_X^{(R)}$ can determine both $\beta_X^{(L)}$ and $\beta_X^{(R)}$. The other way is not true, and there could be multiple joint distributions whose marginals satisfy $\beta_X^{(L)}$ and $\beta_X^{(R)}$.

By a careful calculation, one can still show that the proportion of X-, Y-, Z-variables in $\mathcal{T}_2$ that are not in $\mathcal{T}_1$ is at most a $1-2^{-o(N)}$ fraction of those in $\mathcal{T}_2$. These variables become holes. Unfortunately, the methods in previous works [29, 17] do not apply, as they are unable to fix holes that are present in all three dimensions (X-, Y-, Z-variables).

Next, we discuss how we fix the technical issue.

Intuition of the fix. The first step towards resolving this issue is to decrease the fraction of holes in all three dimensions, from $1-2^{-o(N)}$ all the way down to $2^{-\Omega(N)}$. Then we describe a generic method adapted from [16] for fixing holes in all three dimensions as long as the fractions of holes are small.

For the first step, we slightly relax the condition for zeroing out variables in $\mathcal{T}_1$ and $\mathcal{T}_2$. Let $\varepsilon > 0$ be an arbitrary constant. For any $T_{i,j,k}$, we use $T^{\otimes n}_{i,j,k}[\beta_X, \beta_Y, \beta_Z, \varepsilon]$ to denote $T^{\otimes n}_{i,j,k}$ but we zero out all level-1 X-, Y-, Z-variables, where the proportion of each chunk in $\{0,1,2\}^{2^{\ell-1}}$ in their level-1 block sequence differs at most ε from the corresponding probability in $\beta_X, \beta_Y, \beta_Z$ respectively. That is, we allow some small flexibility when zeroing out variables. Then, let

$$\mathcal{T}_1' \coloneqq \left(T_{i',j',k'} \otimes T_{i-i',j-j',k-k'}\right)^{\otimes n}\left[\beta_X^{(L)} \times \beta_X^{(R)},\ \beta_Y^{(L)} \times \beta_Y^{(R)},\ \beta_Z^{(L)} \times \beta_Z^{(R)},\ \varepsilon\right],$$

and recall

$$\mathcal{T}_2 = \left(T^{\otimes n}_{i',j',k'}\left[\beta_X^{(L)}, \beta_Y^{(L)}, \beta_Z^{(L)}\right]\right) \otimes \left(T^{\otimes n}_{i-i',j-j',k-k'}\left[\beta_X^{(R)}, \beta_Y^{(R)}, \beta_Z^{(R)}\right]\right).$$

Intuitively, we allow more flexibility in $\mathcal{T}_1$ than that in $\mathcal{T}_2$, so that more variables remain in $\mathcal{T}_1$ compared to $\mathcal{T}_2$, and the fraction of holes should become smaller. The idea for proving this is to use concentration bounds: if we pick a uniformly random level-1 X-variable block from $T^{\otimes n}_{i',j',k'}\left[\beta_X^{(L)}, \beta_Y^{(L)}, \beta_Z^{(L)}\right]$ and another uniformly random level-1 X-variable block from $T^{\otimes n}_{i-i',j-j',k-k'}\left[\beta_X^{(R)}, \beta_Y^{(R)}, \beta_Z^{(R)}\right]$, then with very high probability ($1-2^{-\Omega(n)}$), the combination (interleaving the length $2^{\ell-2}$ chunks between their level-1 block sequences) of them satisfies $\beta_X^{(L)} \times \beta_X^{(R)}$, up to ε additive error. Then the fraction of holes is $2^{-\Omega(n)}$. Similar reasons also apply to Y- and Z-variables.

Fixing the holes in all three dimensions. Suppose we have many "broken" copies of some tensor T, in each of which a small fraction of variables (holes) are missing. The goal of this step is to degenerate these broken tensors into one without holes. Schönhage [29] solved this problem for matrix multiplication tensors with holes in only X- and Y-variables, but not Z, via an elegant linear transformation. Duan et al. [17] introduced another method for so-called standard form tensors, which are quite general and are able to capture tensor products of constituent tensors, but can only deal with holes in a single dimension. Duan [16] developed a method utilizing an elegant recursive approach for fixing holes in all three dimensions, but only for matrix multiplication tensors.

We generalize the method of [16] so that it can fix holes in all three dimensions simultaneously, while supporting a broad class of tensors similar to [17]. The only additional requirement compared to [17] is that the fraction of holes is below $O(1/\log N)$, where N is the number of variables in the tensor T. This requirement is satisfied via the previous step of the fix.

Next, we provide some intuition of the recursive hole-fixing approach. Assume T is supported on variable sets X, Y, Z, and the fraction of holes in every copy of T does not exceed $c \ll 1$. We first take one broken copy of T, which we call T_{hole}, and let $X^{(0)}, Y^{(0)}, Z^{(0)}$ denote the set of holes in T_{hole}; let $X^{(1)} \coloneqq X \setminus X^{(0)}$, $Y^{(1)} \coloneqq Y \setminus Y^{(0)}$, $Z^{(1)} \coloneqq Z \setminus Z^{(0)}$ represent the set of non-hole variables. We can further divide T into the sum of eight subtensors:

$$T = \sum_{a,b,c \in \{0,1\}} T\big|_{X^{(a)},Y^{(b)},Z^{(c)}} = T_{\text{hole}} + \sum_{\substack{a,b,c \in \{0,1\} \\ 1 \in \{a,b,c\}}} T\big|_{X^{(a)},Y^{(b)},Z^{(c)}},$$

where $T|_{X',Y',Z'}$ denotes the subtensor of T over subsets of variables $X' \subseteq X$, $Y' \subseteq Y$, and $Z' \subseteq Z$. We directly use the broken copy T_{hole} for the first term, and recurse into seven subproblems to produce the other terms. In each subproblem, at least one of the variable sets is X_1, Y_1 or Z_1, which is c times the size of X, Y, or Z. As long as c is very small, the number of broken copies of T used in this recursive algorithm is affordable.

Rectangular matrix multiplication. In the analysis for square matrix multiplication, we could lower bound the "symmetrized value" of every constituent tensor $T_{i,j,k}$, which captures the asymptotic ability of $T^{\otimes n}_{i,j,k} \otimes T^{\otimes n}_{j,k,i} \otimes T^{\otimes n}_{k,i,j}$ to degenerate into matrix multiplication tensors. The reason why we could symmetrize

the constituent tensors is that we want to obtain square matrix multiplication tensors $\langle a, a, a\rangle$ for some a, which is symmetric about all three dimensions. The situation is different when we consider rectangular matrix multiplications, where we produce matrix multiplication tensors of the form $\langle a, a^\kappa, a\rangle$ to bound $\omega(1, \kappa, 1)$. Thus, we no longer treat the analysis of each constituent tensor $T_{i,j,k}$ as an individual subproblem, because the proportion of $T_{i,j,k}$, $T_{j,k,i}$, and $T_{k,i,j}$ could be different. Hence, it is natural to adopt the framework introduced by Le Gall [22] (and further used in [25]) for rectangular matrix multiplication: we directly apply the laser method on a tensor consisting of multiple constituent tensors, e.g., on $\mathcal{T} = \bigotimes_{i,j,k} T_{i,j,k}^{\otimes \alpha(i,j,k)\cdot n}$, rather than doing this for every term $T_{i,j,k}^{\otimes \alpha(i,j,k)\cdot n}$ separately.

Difficulty of applying the refined laser method. Another natural attempt would be to combine our techniques with the refined laser method introduced in [4], which aims to reduce the "penalty term" that arises when we deal with the block triples inconsistent with the selected distribution α but consistent with the marginals of α. Alman and Vassilevska W. [4] pick a collection of disjoint level-ℓ block triples $X_I Y_J Z_K$ consistent with the chosen distribution α, which we call the "wanted" triples. Then, they zero out a wanted triple with probability $1-p$ and keep it with probability p. Any "unwanted" triple $X_{I'} Y_{J'} Z_{K'}$ only remains with probability p^3, since three involved variable blocks come from three different wanted triples and are zeroed out independently; in contrast, every wanted triple has probability p to remain. The gap between p and p^3 makes it a nontrivial improvement beyond the older method (increasing the modulus of hashing, see, e.g., [15, 33, 23]), which produces a gap between p and p^2.

However, a difficulty arises when the refined laser method is combined with the asymmetric hashing technique in [17] and this paper. Since we allow, e.g., level-ℓ Z-variable blocks to be shared, we can no longer zero out all three blocks X_I, Y_J, Z_K when we decide to give up on this triple, as Z_K might be utilized by other wanted triples. If we only zero out X_I and Y_J simultaneously, the probability of remaining becomes p (for a wanted triple) versus p^2 (for an unwanted triple), which results in the same bound as the older approach.

3 Preliminaries

3.1 Tensors and tensor operations.

Tensors. A tensor T over variable sets $X = \{x_1, \dots, x_{|X|}\}$, $Y = \{y_1, \dots, y_{|Y|}\}$, $Z = \{z_1, \dots, z_{|Z|}\}$ and field $\mathbb{F}$ is a trilinear form

$$T = \sum_{i=1}^{|X|} \sum_{j=1}^{|Y|} \sum_{k=1}^{|Z|} a_{i,j,k} \cdot x_i y_j z_k,$$

where all $a_{i,j,k}$ are from $\mathbb{F}$. X, Y, Z are also called the *support* of the tensor. If all $a_{i,j,k} \in \{0, 1\}$, the tensor T can be considered as over any field $\mathbb{F}$, which is the case for all tensors involved in this paper.

In the following, assume T is a tensor over $X = \{x_1, \dots, x_{|X|}\}$, $Y = \{y_1, \dots, y_{|Y|}\}$, $Z = \{z_1, \dots, z_{|Z|}\}$ and T' is a tensor over $X' = \{x'_1, \dots, x'_{|X'|}\}$, $Y' = \{y'_1, \dots, y'_{|Y'|}\}$, $Z' = \{z'_1, \dots, z'_{|Z'|}\}$, written as

$$T = \sum_{i=1}^{|X|} \sum_{j=1}^{|Y|} \sum_{k=1}^{|Z|} a_{i,j,k} \cdot x_i y_j z_k, \qquad T' = \sum_{i=1}^{|X'|} \sum_{j=1}^{|Y'|} \sum_{k=1}^{|Z'|} b_{i,j,k} \cdot x'_i y'_j z'_k,$$

Tensor operations. Recall the following tensor operations between two tensors T and T':

- The *sum* $T + T'$ is only defined when both tensors are supported on the same sets $(X, Y, Z) = (X', Y', Z')$, given by
$$T + T' = \sum_{i=1}^{|X|} \sum_{j=1}^{|Y|} \sum_{k=1}^{|Z|} (a_{i,j,k} + b_{i,j,k}) \cdot x_i y_j z_k.$$
- The *direct sum* $T \oplus T'$ equals the sum $T + T'$ over disjoint unions $X \sqcup X'$, $Y \sqcup Y'$, and $Z \sqcup Z'$, i.e., we first relabel the variables so that T and T' have disjoint supports, and then take their sum. If T and T' are supported on disjoint variable sets, their sum is the same as their direct sum, in which case we say T and T' are *independent*. We write $T^{\oplus n} \coloneqq \underbrace{T \oplus T \oplus \cdots \oplus T}_{n \text{ copies}}$ to denote the sum of n independent copies of T.

- The *tensor product*, a.k.a. the *Kronecker product*, is defined as the tensor

$$T \otimes T' = \sum_{i=1}^{|X|} \sum_{j=1}^{|Y|} \sum_{k=1}^{|Z|} \sum_{i'=1}^{|X'|} \sum_{j'=1}^{|Y'|} \sum_{k'=1}^{|Z'|} a_{i,j,k} \cdot b_{i',j',k'} \cdot (x_i, x'_{i'}) \cdot (y_j, y'_{j'}) \cdot (z_k, z'_{k'})$$

over variable sets $X \times X'$, $Y \times Y'$, and $Z \times Z'$. We write $T^{\otimes n} \coloneqq \underbrace{T \otimes T \otimes \cdots \otimes T}_{n \text{ times}}$ to denote the *n-th tensor power* of T.

- We say T and T' are *isomorphic*, denoted by $T \equiv T'$, if $|X| = |X'|$, $|Y| = |Y'|$, $|Z| = |Z'|$, and there are permutations π_X, π_Y, π_Z over $[|X|], [|Y|], [|Z|]$ respectively, such that $a_{i,j,k} = b_{\pi_X(i),\pi_Y(j),\pi_Z(k)}$ for all i, j, k. In other words, both tensors are equivalent up to a relabeling of the variables.

3.2 Tensor rank. Given a tensor T over X, Y, Z, the tensor rank $R(T)$ is defined to be the minimum integer $r \geq 0$ such that T can be written as

$$T = \sum_{t=1}^{r} \left(\sum_{i=1}^{|X|} a_{t,i} \cdot x_i \right) \left(\sum_{j=1}^{|Y|} b_{t,j} \cdot y_j \right) \left(\sum_{k=1}^{|Z|} c_{t,k} \cdot z_k \right),$$

where the above sum is called the *rank decomposition* of T.

Given two tensors T, T', the tensor rank satisfies the following property with respect to tensor operations.

- $R(T + T') \leq R(T) + R(T')$.
- $R(T \oplus T') \leq R(T) + R(T')$.
- $R(T \otimes T') \leq R(T) \cdot R(T')$.

The *asymptotic rank* $\widetilde{R}(T)$ of T is defined as

$$\widetilde{R}(T) \coloneqq \lim_{n \to \infty} \left(R(T^{\otimes n}) \right)^{1/n}.$$

Due to the third item above and Fekete's lemma, the asymptotic rank is well-defined and upper bounded by $R(T^{\otimes m})^{1/m}$ for any fixed integer $m > 0$.

3.3 Degenerations, restrictions, zero-outs. Let T be a tensor over X, Y, Z and T' be a tensor over X', Y', Z'. Both T and T' are tensors over a field $\mathbb{F}$.

Degeneration. Let $\mathbb{F}[\lambda]$ be the ring of polynomials of the formal variable λ. We say that T' is a degeneration of T, written as $T \trianglerighteq T'$, if there exists $\mathbb{F}[\lambda]$-linear maps

$$\phi_X : \mathrm{span}_{\mathbb{F}[\lambda]}(X) \to \mathrm{span}_{\mathbb{F}[\lambda]}(X'),$$
$$\phi_Y : \mathrm{span}_{\mathbb{F}[\lambda]}(Y) \to \mathrm{span}_{\mathbb{F}[\lambda]}(Y'),$$
$$\phi_Z : \mathrm{span}_{\mathbb{F}[\lambda]}(Z) \to \mathrm{span}_{\mathbb{F}[\lambda]}(Z'),$$

and $d \in \mathbb{N}$ such that

$$T' = \lambda^{-d} \left(\sum_{i=1}^{|X|} \sum_{j=1}^{|Y|} \sum_{k=1}^{|Z|} a_{i,j,k} \cdot \phi_X(x_i) \cdot \phi_Y(y_j) \cdot \phi_Z(z_k) \right) + O(\lambda).$$

It is not hard to check that if $T' \trianglerighteq T$, then $\widetilde{R}(T') \leq \widetilde{R}(T)$.

Restriction. Restriction is a special type of degeneration that considers the case where the maps ϕ_X, ϕ_Y, ϕ_Z are $\mathbb{F}$-linear maps. More specifically, T' is a restriction of T if there exist $\mathbb{F}$-linear maps

$$\phi_X : \mathrm{span}_{\mathbb{F}}(X) \to \mathrm{span}_{\mathbb{F}}(X'),$$
$$\phi_Y : \mathrm{span}_{\mathbb{F}}(Y) \to \mathrm{span}_{\mathbb{F}}(Y'),$$
$$\phi_Z : \mathrm{span}_{\mathbb{F}}(Z) \to \mathrm{span}_{\mathbb{F}}(Z'),$$

such that

$$T' = \sum_{i=1}^{|X|}\sum_{j=1}^{|Y|}\sum_{k=1}^{|Z|} a_{i,j,k} \cdot \phi_X(x_i) \cdot \phi_Y(y_j) \cdot \phi_Z(z_k).$$

It is not hard to see that since the maps ϕ_X, ϕ_Y, ϕ_Z are linear transformations, we have $R(T') \leq R(T)$ and consequently $\widetilde{R}(T') \leq \widetilde{R}(T)$.

Zero-out. In the laser method, we only consider a limited type of restriction called zero-outs, namely the maps ϕ_X, ϕ_Y, ϕ_Z set some variables to zero. More specifically, we choose subsets $X' \subseteq X$, $Y' \subseteq Y$, $Z' \subseteq Z$ and define the maps as

$$\phi_X(x_i) = \begin{cases} x_i & \text{If } x_i \in X', \\ 0 & \text{otherwise,} \end{cases}$$

and similarly for ϕ_Y, ϕ_Z. The resulting tensor

$$T' = \sum_{i=1}^{|X|}\sum_{j=1}^{|Y|}\sum_{k=1}^{|Z|} a_{i,j,k} \cdot \phi_X(x_i) \cdot \phi_Y(y_j) \cdot \phi_Z(z_k) = \sum_{x_i \in X'}\sum_{y_j \in Y'}\sum_{z_k \in Z'} a_{i,j,k} \cdot x_i y_j z_k$$

is called a zero-out of T. Throughout this paper, we use the notation $T' = T|_{X',Y',Z'}$ to denote such a tensor T' obtained as a zero-out of T and we say that the variables in $X \setminus X'$, $Y \setminus Y'$, $Z \setminus Z'$ are zeroed out. In this case, we also call T' the *subtensor* of T over X', Y', Z'.

3.4 Matrix multiplication tensors. For positive integers a, b, c, the $a \times b \times c$ matrix multiplication tensor $\langle a, b, c\rangle$ is a tensor over the variable sets $\{x_{ij}\}_{i\in[a],j\in[b]}, \{y_{jk}\}_{j\in[b],k\in[c]}, \{z_{ki}\}_{i\in[a],k\in[c]}$ defined as the tensor computing the $a \times c$ product matrix $\{z_{ki}\}_{i\in[a],k\in[c]}$ of an $a \times b$ matrix $\{x_{ij}\}_{i\in[a],j\in[b]}$ and $b \times c$ matrix $\{y_{jk}\}_{j\in[b],k\in[c]}$. Specifically, $\langle a, b, c\rangle$ can be written as the trilinear form

$$\langle a, b, c\rangle = \sum_{i\in[a]}\sum_{j\in[b]}\sum_{k\in[c]} x_{ij} y_{jk} z_{ki}.$$

It is not hard to check that $\langle a, b, c\rangle \otimes \langle d, e, f\rangle \equiv \langle ad, be, cf\rangle$.

Following from the recursive approach introduced by Strassen in [31], for any integer $q \geq 2$, if $R(\langle q, q, q\rangle) \leq r$, then one can use the rank decomposition of $\langle q, q, q\rangle$ to design an arithmetic circuit of size $O(n^{\log_q(r)})$ to multiply two $n \times n$ matrices. This motivates the definition of the *matrix multiplication exponent* ω as follows:

$$\omega := \inf_{q\in\mathbb{N},\, q\geq 2} \log_q(R(\langle q, q, q\rangle)).$$

Namely, for every $\varepsilon > 0$, there exists an arithmetic circuit of size $O(n^{\omega+\varepsilon})$ that computes the multiplication of two $n \times n$ matrices. Since $\langle q, q, q\rangle^{\otimes n} \equiv \langle q^n, q^n, q^n\rangle$, equivalently ω can be written in terms of the asymptotic rank of $\langle q, q, q\rangle$ as

$$\omega = \log_q(\widetilde{R}(\langle q, q, q\rangle)).$$

In this paper, we also consider the arithmetic complexity of multiplying rectangular matrices of sizes $n^a \times n^b$ and $n^b \times n^c$ where $a, b, c \in \mathbb{R}_{\geq 0}$. We define the quantity $\omega(a, b, c)$ similar to ω as

$$\omega(a, b, c) = \log_q\Big(\widetilde{R}(\langle q^a, q^b, q^c\rangle)\Big)$$

where $q \geq 2$ is a positive integer. This means that for any $\varepsilon > 0$, there exists an arithmetic circuit of size $O(n^{\omega(a,b,c)+\varepsilon})$ that computes the multiplication of an $n^a \times n^b$ matrix with an $n^b \times n^c$ matrix. In this paper, we focus on bounds for the values of the form $\omega(1, \kappa, 1)$ for $\kappa > 0$. We remark that it is known that $\omega(1, 1, \kappa) = \omega(1, \kappa, 1) = \omega(\kappa, 1, 1)$.

3.5 Schönhage's asymptotic sum inequality. By the above definition of ω, it is clear that if one can bound the asymptotic rank of matrix multiplication tensors, then one would get an upper bound on ω. In fact, Schönhage showed in [29] that one can obtain an upper bound on ω if one can bound the asymptotic rank of a direct sum of matrix multiplication tensors. Specifically, we recall Shönhage's asymptotic sum inequality as follows.

THEOREM 3.1 (Asymptotic sum inequality [29]). *For positive integers $r > m$ and a_i, b_i, c_i for $i \in [m]$, if*

$$\widetilde{R}\left(\bigoplus_{i=1}^{m}\langle a_i, b_i, c_i\rangle\right) \leq r,$$

then $\omega \leq 3\tau$ where $\tau \in [2/3, 1]$ is the solution to the equation

$$\sum_{i=1}^{m}(a_i \cdot b_i \cdot c_i)^{\tau} = r.$$

Analogously, the asymptotic sum inequality can also be used to obtain bounds on the rectangular matrix multiplication as follows.

THEOREM 3.2 (Asymptotic sum inequality for $\omega(a, b, c)$ [29]). *Let $t, q > 0$ be positive integers and $a, b, c \geq 0$, then*

$$t \cdot q^{\omega(a,b,c)} \leq \widetilde{R}\left(\bigoplus_{i=1}^{t}\langle q^a, q^b, q^c\rangle\right).$$

3.6 The Coppersmith-Winograd tensor. For a nonnegative integer $q \geq 0$, the Coppersmith-Winograd tensor CW_q over the variables $X = \{x_0, \ldots, x_{q+1}\}$, $Y = \{y_0, \ldots, y_{q+1}\}$, $Z = \{z_0, \ldots, z_{q+1}\}$ is defined as

$$\mathrm{CW}_q := x_0y_0z_{q+1} + x_0y_{q+1}z_0 + x_{q+1}y_0z_0 + \sum_{i=1}^{q}(x_0y_iz_i + x_iy_0z_i + x_iy_iz_0).$$

Observe that

$$\sum_{i=1}^{q} x_0y_iz_i + \sum_{i=1}^{q} x_iy_0z_i + \sum_{i=1}^{q} x_iy_iz_0 \equiv \langle 1, 1, q\rangle + \langle q, 1, 1\rangle + \langle 1, q, 1\rangle,$$

so CW_q is the sum of six matrix multiplication tensors where the other three are copies of $\langle 1, 1, 1\rangle$. It is known from Coppersmith and Winograd [13] that $\widetilde{R}(\mathrm{CW}_q) \leq q + 2$.

3.7 Base leveled partition of CW_q. We will consider the $2^{\ell-1}$-th tensor power of CW_q for $\ell \geq 1$. For convenience, we use the notation $T^{(\ell)} := \mathrm{CW}_q^{\otimes 2^{\ell-1}}$. There is a natural partitioning of the variables of CW_q introduced in [13] and consequently used in all following works including [33, 4, 23, 17]. We now describe the leveled partition of $T^{(\ell)}$.

Level-1 partition. For $T^{(1)} = \mathrm{CW}_q$, its variable sets $X^{(1)}, Y^{(1)}, Z^{(1)}$ are partitioned into three parts

$$\begin{aligned}
X^{(1)} &= X_0^{(1)} \sqcup X_1^{(1)} \sqcup X_2^{(1)} = \{x_0\} \sqcup \{x_1, \ldots, x_q\} \sqcup \{x_{q+1}\},\\
Y^{(1)} &= Y_0^{(1)} \sqcup Y_1^{(1)} \sqcup Y_2^{(1)} = \{y_0\} \sqcup \{y_1, \ldots, y_q\} \sqcup \{y_{q+1}\},\\
Z^{(1)} &= Z_0^{(1)} \sqcup Z_1^{(1)} \sqcup Z_2^{(1)} = \{z_0\} \sqcup \{z_1, \ldots, z_q\} \sqcup \{z_{q+1}\}.
\end{aligned}$$

We use $T^{(1)}_{i,j,k}$ to denote the subtensor $T^{(1)}|_{X_i,Y_j,Z_k}$ and we call $T^{(1)}_{i,j,k}$ a *level-1 constituent tensor*. Then notice that under the above partition, the constituent tensor $T^{(1)}_{i,j,k}$ is nonzero if and only if $i + j + k = 2$. In particular, we can write CW_q as a sum of constituent tensors as follows

$$T^{(1)} = \mathrm{CW}_q = \sum_{\substack{i,j,k\geq 0\\ i+j+k=2}} T^{(1)}_{i,j,k}.$$

Level-ℓ partition. For $T^{(\ell)} = \mathrm{CW}_q^{\otimes 2^{\ell-1}}$ with variable sets $X^{(\ell)}, Y^{(\ell)}, Z^{(\ell)}$, the above level-1 partition on $T^{(1)}$ directly induces a partition on the variable sets $X^{(\ell)}, Y^{(\ell)}, Z^{(\ell)}$ where each part of the partition is indexed by a $\{0,1,2\}$-sequence of length $2^{\ell-1}$. Specifically, this gives the partition

$$X^{(\ell)} = \bigsqcup_{(\hat{i}_1,\hat{i}_2,\ldots,\hat{i}_{2^{\ell-1}})\in\{0,1,2\}^{2^{\ell-1}}} X^{(1)}_{\hat{i}_1} \otimes X^{(1)}_{\hat{i}_2} \otimes \cdots \otimes X^{(1)}_{\hat{i}_{2^{\ell-1}}}$$

for X-variables and analogous partitions for Y- and Z-variables.

In order to obtain an improvement by analyzing higher tensor powers of CW_q, we need to consider the following coarsening of the induced partition where the parts corresponding to sequences with the same sum are "merged" into a single part. More specifically, we have

$$X^{(\ell)} = \bigsqcup_{i=0}^{2^\ell} X^{(\ell)}_i, \qquad \text{where} \quad X^{(\ell)}_i := \bigsqcup_{\substack{(\hat{i}_1,\hat{i}_2,\ldots,\hat{i}_{2^{\ell-1}})\in\{0,1,2\}^{2^{\ell-1}} \\ \sum_t \hat{i}_t = i}} X^{(1)}_{\hat{i}_1} \otimes X^{(1)}_{\hat{i}_2} \otimes \cdots \otimes X^{(1)}_{\hat{i}_{2^{\ell-1}}}.$$

We refer to this above coarsened partition of $T^{(\ell)}$ as the *level-ℓ partition*. Note that we can also view this partition as obtained from coarsening the level-$(\ell-1)$ partition, i.e.,

$$X^{(\ell)}_i = \bigsqcup_{\substack{0\le i'\le i \\ 0\le i', i-i'\le 2^\ell}} X^{(\ell-1)}_{i'} \otimes X^{(\ell-1)}_{i-i'}.$$

We can partition the variable sets $Y^{(\ell)}$ and $Z^{(\ell)}$ similarly.

Under the level-ℓ partition, we use $T^{(\ell)}_{i,j,k}$ to denote the subtensor $T^{(\ell)}|_{X^{(\ell)}_i, Y^{(\ell)}_j, Z^{(\ell)}_k}$ and note that $T^{(\ell)}_{i,j,k}$ is nonzero if and only if $i+j+k = 2^\ell$. So we have

$$T^{(\ell)} = \mathrm{CW}_q^{\otimes 2^{\ell-1}} = \sum_{\substack{i,j,k\ge 0 \\ i+j+k=2^\ell}} T^{(\ell)}_{i,j,k}.$$

We call each $T^{(\ell)}_{i,j,k}$ a *level-ℓ constituent tensor*, $X^{(\ell)}_i, Y^{(\ell)}_j, Z^{(\ell)}_k$ level-ℓ variable blocks, and we omit the superscript (ℓ) when ℓ is clear from context.

3.8 Leveled partition for large tensor powers of CW_q.

In the laser method, we consider a large tensor power of CW_q in the form $(T^{(\ell)})^{\otimes n} = (\mathrm{CW}_q)^{\otimes n\cdot 2^{\ell-1}}$. We set $N := n\cdot 2^{\ell-1}$ and note that the leveled partition of $T^{(\ell)}$ induces a partition on $(T^{(\ell)})^{\otimes n}$. We recall some basic terminology and notations with respect to the leveled-partition of $(T^{(\ell)})^{\otimes n}$.

Level-1 partition of $(\mathrm{CW}_q)^{\otimes N}$. In level-1, we view $(\mathrm{CW}_q)^{\otimes N}$ as the tensor $(T^{(1)})^{\otimes N}$ and consider the partition induced by the level-1 partition on $T^{(1)}$. Each level-1 X-variable block $X_{\hat{I}}$ is indexed by a sequence $\hat{I} = (\hat{I}_1, \ldots, \hat{I}_N)$ of length N in $\{0,1,2\}^N$. The variable block X_I is defined as

$$X_{\hat{I}} := X^{(1)}_{\hat{I}_1} \otimes \cdots \otimes X^{(1)}_{\hat{I}_N},$$

where $X^{(1)}_{I_t}$ for $t \in [N]$ is the level-1 partition of $T^{(1)}$. We call $X_{\hat{I}}$ a *level-1 variable block* and $\hat{I}$ its *level-1 index sequence*. The level-1 Y- and Z-variable blocks $Y_{\hat{J}}$ and $Z_{\hat{K}}$ are defined similarly for level-1 index sequences $\hat{J}, \hat{K} \in \{0,1,2\}^N$. Then notice that $X_{\hat{I}}, Y_{\hat{J}}, Z_{\hat{K}}$ form a nonzero subtensor of $(T^{(1)})^{\otimes N}$ if $\hat{I}_t + \hat{J}_t + \hat{K}_t = 2$ for all $t \in [N]$. So we can write $(T^{(1)})^{\otimes N}$ as a sum of subtensors

$$(T^{(1)})^{\otimes N} = \sum_{\substack{\hat{I},\hat{J},\hat{K}\in\{0,1,2\}^N \\ \hat{I}_t+\hat{J}_t+\hat{K}_t=2 \ \ \forall t\in[N]}} (T^{(1)})^{\otimes N}\big|_{X_{\hat{I}}, Y_{\hat{J}}, Z_{\hat{K}}}.$$

For convenience, we use $X_{\hat{I}}Y_{\hat{J}}Z_{\hat{K}}$ to denote the subtensor $(T^{(1)})^{\otimes N}|_{X_{\hat{I}}, Y_{\hat{J}}, Z_{\hat{K}}}$ and we call $X_{\hat{I}}Y_{\hat{J}}Z_{\hat{K}}$ a *level-1 triple*.

Level-ℓ partition of $(\mathrm{CW}_q)^{\otimes N}$. In level-$\ell$, we view $(\mathrm{CW}_q)^{\otimes N}$ as the tensor $(T^{(\ell)})^{\otimes n}$ where $n = N/2^{\ell-1}$ and consider the partition induced by the level-ℓ partition on $T^{(\ell)}$. Each level-1 X-variable block X_I is indexed by a sequence $I \in \{0, 1, \ldots, 2^\ell\}^n$ of length n. The variable block X_I is defined as

$$X_I := X^{(\ell)}_{I_1} \otimes \cdots \otimes X^{(\ell)}_{I_n}$$

where $X^{(\ell)}_i$ $(0 \le i \le 2^\ell)$ is the i-th part in the level-ℓ partition of $T^{(\ell)}$. We call X_I a *level-ℓ variable block* and I its *level-ℓ index sequence*. The level-ℓ Y- and Z-variable blocks Y_J and Z_K are defined similarly for level-ℓ index sequences $J, K \in \{0, 1, \ldots, 2^\ell\}^n$. Similarly, the level-$\ell$ variable blocks X_I, Y_J, Z_K form a nonzero subtensor of $(T^{(\ell)})^{\otimes n}$ when $I_t + J_t + K_t = 2^\ell$ for all $t \in [n]$. So we can write

$$(T^{(\ell)})^{\otimes n} = \sum_{\substack{\hat{I},\hat{J},\hat{K} \in \{0,1,2^\ell\}^n \\ I_t + J_t + K_t = 2^\ell \ \ \forall t \in [N]}} (T^{(\ell)})^{\otimes n}|_{X_I, Y_J, Z_K}.$$

For convenience, we use the notation $X_I Y_J Z_K$ to denote the subtensor $(T^{(\ell)})^{\otimes n}|_{X_I, Y_J, Z_K}$ and we call such $X_I Y_J Z_K$ a *level-ℓ triple*.

In addition, note that since the level-ℓ partition of $T^{(\ell)}$ is a coarsening of the partition induced by the level-1 partition of $T^{(1)}$, a level-1 variable block $X_{\hat{I}}$ is contained in a level-ℓ variable block X_I if the sequence $I' = (I'_1, \ldots, I'_n)$ formed by taking $I'_t = \sum_{i=1}^{2^{\ell-1}} \hat{I}_{(t-1)\cdot 2^{\ell-1}+i}$ satisfies $I'_t = I_t$ for all $t \in [n]$. Namely, if taking the sum of consecutive length-$2^{\ell-1}$ subsequences in $\hat{I}$ yields the sequence I, then $X_{\hat{I}}$ is contained in X_I. In this case, we use the notation $\hat{I} \in I$ and $X_{\hat{I}} \in X_I$.

3.9 Distributions and entropy. In this paper, we only consider distributions with a finite support. Let α be a distribution supported on a set S, we have $\alpha(s) \ge 0$ for all $s \in S$ and $\sum_{s \in S} \alpha(s) = 1$. The *entropy* of α, denoted as $H(\alpha)$, is defined as

$$H(\alpha) := -\sum_{\substack{s \in S \\ \alpha(s) > 0}} \alpha(s) \log \alpha(s),$$

where the log has base 2. We will frequently use the following well-known combinatorial fact.

LEMMA 3.1. *Let α be a distribution over the set $[s] = \{1, \ldots, s\}$. Let $N > 0$ be a positive integer, then we have*

$$\binom{N}{\alpha(1)N, \ldots, \alpha(s)N} = 2^{N(H(\alpha) \pm o(1))}.$$

For two distributions α and β over the sets S and S' respectively, we define the joint distribution $\alpha \times \beta$ as the distribution over $S \times S' = \{(s, s') \mid s \in S, s' \in S'\}$ such that

$$(\alpha \times \beta)(s, s') = \alpha(s) \cdot \beta(s').$$

When S and S' are sets of integer sequences, we will instead define $\alpha \times \beta$ as a distribution over all integer sequences that can be obtained by concatenating one sequence in S and one sequence in S', such that

$$(\alpha \times \beta)(s \circ s') = \alpha(s) \cdot \beta(s'),$$

where $s \circ s'$ denotes the concatenation of s and s'.

3.10 Complete split distributions. Motivated by the leveled partition of tensor powers of CW_q, we define the notion of complete split distributions to characterize the level-1 variable blocks contained in level-ℓ variable blocks.

DEFINITION 3.1 (Complete Split Distribution). *A* complete split distribution *for a level-ℓ constituent tensor $T_{i,j,k}$ with $i + j + k = 2^\ell$ is a distribution on all length $2^{\ell-1}$ sequences $(\hat{i}_1, \hat{i}_2, \ldots, \hat{i}_{2^{\ell-1}}) \in \{0, 1, 2\}^{2^{\ell-1}}$.*

For a level-1 index sequence $\hat{I} \in \{0,1,2\}^{2^{\ell-1}\cdot n}$, we say that it is *consistent* with a complete split distribution β if the proportion of any index sequence $(\hat{i}_1, \hat{i}_2, \ldots, \hat{i}_{2^{\ell-1}})$ in

$$\left\{ \left(\hat{I}_{(t-1)\cdot 2^{\ell-1}+p}\right)_{p=1}^{2^{\ell-1}} \;\middle|\; t \in [n] \right\}$$

equals $\beta(\hat{i}_1, \hat{i}_2, \ldots, \hat{i}_{2^{\ell-1}})$. Namely, for every $(\hat{i}_1, \ldots, \hat{i}_{2^{\ell-1}}) \in \{0,1,2\}^{2^{\ell-1}}$, we have

$$\left| \left\{ t \in [n] \;\middle|\; \left(\hat{I}_{(t-1)\cdot 2^{\ell-1}+p}\right)_{p=1}^{2^{\ell-1}} = (\hat{i}_1, \ldots, \hat{i}_{2^{\ell-1}}) \right\} \right| = \beta(\hat{i}_1, \hat{i}_2, \ldots, \hat{i}_{2^{\ell-1}}) \cdot n.$$

Notice that any level-1 index sequence $\hat{I} \in \{0,1,2\}^{2^{\ell-1}\cdot n}$ defines a complete split distribution by computing the proportions of each type of length-$2^{\ell-1}$ consecutive chunks present in $\hat{I}$. More specifically, we have the following definition.

DEFINITION 3.2. *Given a level-1 index sequence $\hat{I} \in \{0,1,2\}^{2^{\ell-1}\cdot n}$, its complete split distribution over $(\hat{i}_1, \ldots, \hat{i}_{2^{\ell-1}}) \in \{0,1,2\}^{2^{\ell-1}}$ is defined as*

$$\mathsf{split}\big(\hat{I}\big)\big(\hat{i}_1, \ldots, \hat{i}_{2^{\ell-1}}\big) = \frac{1}{n} \cdot \left| \left\{ t \in [n] \;\middle|\; \left(\hat{I}_{(t-1)\cdot 2^{\ell-1}+p}\right)_{p=1}^{2^{\ell-1}} = \big(\hat{i}_1, \ldots, \hat{i}_{2^{\ell-1}}\big) \right\} \right|.$$

Given a subset $S \subseteq [n]$, we can define the complete split distribution over $(\hat{i}_1, \ldots, \hat{i}_{2^{\ell-1}}) \in \{0,1,2\}^{2^{\ell-1}}$ given by $\hat{I}$ restricted to the subset S as

$$\mathsf{split}\big(\hat{I}, S\big)\big(\hat{i}_1, \ldots, \hat{i}_{2^{\ell-1}}\big) = \frac{1}{|S|} \cdot \left| \left\{ t \in S \;\middle|\; \left(\hat{I}_{(t-1)\cdot 2^{\ell-1}+p}\right)_{p=1}^{2^{\ell-1}} = \big(\hat{i}_1, \ldots, \hat{i}_{2^{\ell-1}}\big) \right\} \right|.$$

Given two complete split distributions β_1 and β_2 over the length-$2^{\ell-1}$ index sequences $\{0,1,2\}^{2^{\ell-1}}$, the L_∞ distances between β_1 and β_2 is defined to be

$$\|\beta_1 - \beta_2\|_\infty = \max_{\sigma \in \{0,1,2\}^{2^{\ell-1}}} |\beta_1(\sigma) - \beta_2(\sigma)|.$$

For any constant $\varepsilon > 0$ and a fixed complete split distribution β, we say that a level-1 index sequence $\hat{I} \in \{0,1,2\}^{2^{\ell-1}\cdot n}$ is consistent with β up to ε error if $\|\mathsf{split}(\hat{I}) - \beta\|_\infty \le \varepsilon$. When the ε is clear from context, we say that $\hat{I}$ is approximately consistent with β if it is consistent with β up to ε error.

DEFINITION 3.3. *For a level-ℓ constituent tensor $T_{i,j,k}$, an integer exponent N, a constant $\varepsilon \ge 0$, and three complete split distributions $\beta_X, \beta_Y, \beta_Z$ for the X-, Y-, Z-variables respectively, we define*

$$T^{\otimes N}_{i,j,k}[\beta_X, \beta_Y, \beta_Z, \varepsilon] \coloneqq \sum_{\substack{\text{level-1 triple } X_{\hat{I}}Y_{\hat{J}}Z_{\hat{K}} \text{ in } T^{\otimes N}_{i,j,k} \\ \hat{I} \text{ approximately consistent with } \beta_X \\ \hat{J} \text{ approximately consistent with } \beta_Y \\ \hat{K} \text{ approximately consistent with } \beta_Z}} X_{\hat{I}}Y_{\hat{J}}Z_{\hat{K}}.$$

It is a subtensor of $T^{\otimes N}_{i,j,k}$ over all level-1 X-, Y-, Z-variable blocks that are approximately consistent with β_X, β_Y, β_Z, respectively. When $\varepsilon = 0$, we will simplify the notation to $T^{\otimes N}_{i,j,k}[\beta_X, \beta_Y, \beta_Z]$.

3.11 Salem-Spencer sets. In the hashing step of the laser method, we make use of the existence of a large dense subset of $\mathbb{Z}_M$ that avoids 3-term arithmetic progressions. We recall the following past result.

THEOREM 3.3 ([28, 7]). *For every positive integer $M > 0$, there exists a subset $B \subseteq \mathbb{Z}_M$ of size*

$$|B| \ge M \cdot e^{-O(\sqrt{\log M})} = M^{1-o(1)}$$

that contains no nontrivial 3-term arithmetic progressions. Specifically, any $a, b, c \in B$ satisfy $a+b \equiv 2c \pmod M$ if and only if $a = b = c$.

4 Algorithm Outline

In the following, we will use $\kappa \geq 0$ to denote that we want to obtain an upper bound on $\omega(1, \kappa, 1)$.

In this section, we give the outline of our algorithm, which accepts $\mathrm{CW}_q^{\otimes N}$ as its input for a large enough N, and degenerates it into a collection of independent matrix multiplication tensors of the same size $\langle m, m^\kappa, m\rangle$. By the asymptotic sum inequality (Theorem 3.2), this will give an upper bound on $\omega(1, \kappa, 1)$.

4.1 Algorithm framework. The following notion of *interface tensor* acts as an interface of our algorithm between different levels. In general, each level of our algorithm takes an interface tensor as input (except the first level, which takes a large tensor power of CW_q), and degenerates it into independent copies of an interface tensor.

DEFINITION 4.1 (Interface tensor). *For a positive integer $\ell \geq 1$ and any constant $0 \leq \varepsilon \leq 1$, a level-$\ell$ ε-*interface tensor $\mathcal{T}^*$ *with parameter list*

$$\{(n_t, i_t, j_t, k_t, \beta_{X,t}, \beta_{Y,t}, \beta_{Z,t})\}_{t\in[s]}$$

is defined as

$$\mathcal{T}^* \coloneqq \bigotimes_{t=1}^{s} T_{i_t,j_t,k_t}^{\otimes n_t}[\beta_{X,t}, \beta_{Y,t}, \beta_{Z,t}, \varepsilon],$$

where $i_t + j_t + k_t = 2^\ell$ for every $t \in [s]$ (i.e., T_{i_t,j_t,k_t} is a level-ℓ constituent tensor) and $\beta_{X,t}, \beta_{Y,t}, \beta_{Z,t}$ are level-ℓ complete split distributions for X-, Y-, Z-variables respectively. We call each $T_{i_t,j_t,k_t}^{\otimes n_t}[\beta_{X,t}, \beta_{Y,t}, \beta_{Z,t}, \varepsilon]$ a term *of $\mathcal{T}^*$. When $\varepsilon = 0$, we will simply call $\mathcal{T}^*$ a level-ℓ interface tensor.*

Note that the same (i_t, j_t, k_t) can appear multiple times in the parameter list, with potentially different $n_k, \beta_{X,t}, \beta_{Y,t}, \beta_{Z,t}$. Also note that the tensor product of two level-ℓ ε-interface tensors is also a level-ℓ ε-interface tensor, whose parameter list is the concatenation of the parameter lists of the two level-ℓ ε-interface tensors.

The framework of our algorithm is as follows. First, we apply the *global stage* algorithm described in Section 5 on input $\left(\mathrm{CW}_q^{\otimes 2^{\ell^*}}\right)^{\otimes n}$ to degenerate it into independent copies of a level-ℓ^* ε_{ℓ^*}-interface tensor. Then we apply the *constituent tensor stage* algorithm described in Section 6 for $\ell = \ell^*, \ell^* - 1, \ldots, 2$ to obtain the tensor product between a matrix multiplication tensor and independent copies of a level-1 ε_1-interface tensor. More specifically, the constituent tensor stage algorithm takes as input a level-ℓ ε_ℓ-interface tensor and outputs the tensor product between a matrix multiplication tensor and independent copies of a level-$(\ell-1)$ $\varepsilon_{\ell-1}$-interface tensor, so we can keep applying the constituent tensor stage algorithm on each level-$(\ell-1)$ interface tensors that was outputted previously until we get a tensor product between a matrix multiplication tensor and independent copies of a level-1 ε_1-interface tensor. Finally, we show that each level-1 ε_1-interface tensor can be easily degenerated into a matrix multiplication tensor, so we obtain independent copies of matrix multiplication tensors of dimension $\langle m, m^\kappa, m\rangle$.

4.2 Algorithm outline. We first give a high-level outline of each step of the global stage algorithm. The constituent tensor stage algorithm will share similar high-level ideas.

The algorithm takes in $\left(\mathrm{CW}_q^{\otimes 2^{\ell-1}}\right)^{\otimes n}$ as input and outputs level-1-independent level-ℓ interface tensors as a degeneration of the input (for simplicity, we consider the $\varepsilon = 0$ case in this outline). In the algorithm, we define the notion of compatibility between level-1 blocks and level-ℓ triples with respect to some specified complete split distributions, so that if all level-1 blocks in the remaining tensor are compatible with *exactly one* level-ℓ triple, then the subtensors over each remaining triple are level-1-independent. So the goal of the algorithm is to zero out some level-ℓ and level-1 variable blocks such that each remaining level-1 block is compatible with a unique level-ℓ triple. The structure of the algorithm is similar to the global stage algorithm in [17] with the main modification being the generalization from split distributions to complete split distributions.

On input $\left(\mathrm{CW}_q^{\otimes 2^{\ell-1}}\right)^{\otimes n}$, we first view the tensor as the tensor product of three terms, where each term is called a region, i.e., we write $\left(\mathrm{CW}_q^{\otimes 2^{\ell-1}}\right)^{\otimes n}$ as $\bigotimes_{r\in[3]}\left(\mathrm{CW}_q^{\otimes 2^{\ell-1}}\right)^{\otimes A_r\cdot n}$ for some $A_1, A_2, A_3 \geq 0$ and $A_1 + A_2 + A_3 = 1$. Recall that we are only able to allow the sharing of level-ℓ variable blocks in one of X-, Y-, Z-dimensions, so each region will allow the sharing of level-ℓ variable blocks in different dimensions and we will perform the subsequent steps on the three regions separately. This step helps balance the number of remaining variable blocks in the three dimensions due to the asymmetric nature of the subsequent procedure.

From now on, we describe the procedure on the first region where we allow the sharing of level-ℓ Z-variable blocks. We perform the same procedure up to rotation of the three dimensions on the other two regions separately.

1. **Zero out according to α.** For a distribution α over the level-ℓ constituent subtensors and its induced marginals $\alpha_X, \alpha_Y, \alpha_Z$ in the X-, Y-, Z-dimensions, we zero out level-ℓ X-, Y-, Z-variable blocks that are not consistent with $\alpha_X, \alpha_Y, \alpha_Z$ respectively.

2. **Asymmetric hashing.** We use pairwise independent hash functions that hash level-ℓ index sequences to the set $\{0, \dots, M-1\}$ for some M which partitions the level-ℓ variable blocks into buckets based on its hash value. Within each bucket, we do asymmetric cleanup so that every level-ℓ X-variable block X_I or Y-variable block Y_J is contained in a unique level-ℓ triple $X_I Y_J Z_K$, while a level-ℓ Z-variable block Z_K could be contained in multiple level-ℓ triples.

3. **Compatibility zero-out I.** We define a notion of *compatibility* with respect to the complete split distributions between level-1 blocks and level-ℓ triples for a set of specified level-ℓ complete split distributions $\{\beta_{X,i,j,k}, \beta_{Y,i,j,k}, \beta_{Z,i,j,k}\}_{i+j+k=2^\ell}$ for the X-, Y-, Z-blocks. We zero out all the level-1 X- or Y-blocks that are not consistent with $\{\beta_{X,i,j,k}\}_{i+j+k=2^\ell}, \{\beta_{Y,i,j,k}\}_{i+j+k=2^\ell}$ respectively (we can only do this because every level-ℓ X-variable block X_I or Y-variable block Y_J is contained in a unique level-ℓ triple). We zero out all the level-1 Z-blocks that are incompatible with any level-ℓ triples.

4. **Compatibility zero-out II: unique triple.** After the compatibility zero-out I, every level-1 block is compatible with at least 1 level-ℓ triple and we want every level-1 block to be compatible with *exactly one* level-ℓ triple. So in this step, we zero out level-1 Z-blocks that are compatible with more than one level-ℓ triples. Note that the level-1 blocks zeroed out in this step will become holes.

5. **Usefulness zero-out.** Now that each remaining level-1 Z-block $Z_{\hat{K}}$ is contained in exactly one level-ℓ triple $X_I Y_J Z_K$, we can define the notion of whether a level-1 block is *useful* for the level-ℓ triple containing it as whether it is consistent with the complete split distributions $\{\beta_{Z,i,j,k}\}_{i+j+k=2^\ell}$. Note that we can only do this now because previously we do not have the property that every level-1 Z-block is in a unique level-ℓ triple. In this step we zero out the level-1 blocks that are not useful for the level-ℓ triple containing it.

6. **Fixing holes.** Now we have obtained level-1-independent level-ℓ interface tensors with holes. We use the following result which will be proved in Section 7 to fix the holes.

   COROLLARY 4.1 (Fixing holes in interface tensors). *Let T be a level-ℓ interface tensor with parameter list*

   $$\{(n_t, i_t, j_t, k_t, \beta_{X,t}, \beta_{Y,t}, \beta_{Z,t})\}_{t\in[s]}.$$

   Let $N = 2^{\ell-1} \cdot \sum_{t\in[s]} n_t$. Suppose $T_1, \dots, T_r$ are broken copies of T where $\le \frac{1}{8N}$ fraction of level-1 X-, Y- and Z-blocks are holes. If $r \ge 2^{C_1 N/\log N}$ for some large enough constant $C_1 > 0$, the direct sum $\bigoplus_{i=1}^r T_i$ can degenerate into an unbroken copy of T.

5 Global Stage

In the global stage, we take as input the tensor $\mathrm{CW}_q^{\otimes N}$ for $N = n \cdot 2^{\ell^*}$ and output independent copies of a level-ℓ^* interface tensor, where the output will be a degeneration of the input. For the rest of this section, we will use ℓ to denote ℓ^* for convenience.

Given α, which is a distribution over $\{(i,j,k) \in \mathbb{Z}_{\ge 0}^3 \mid i+j+k = 2^\ell\}$, and $\beta_{X,i,j,k}, \beta_{Y,i,j,k}, \beta_{Z,i,j,k}$, which are level-$\ell$ complete split distributions, we define the following quantities:

- α_X is the marginal distribution of α on the X-dimension, i.e., $\alpha_X(i) = \sum_{j,k} \alpha(i,j,k)$ for any i. We also similarly define α_Y and α_Z.
- D is the set of distributions whose marginal distributions on the three dimensions are $\alpha_X, \alpha_Y, \alpha_Z$ respectively, and let the penalty term $P_\alpha := \max_{\alpha' \in D} H(\alpha') - H(\alpha) \ge 0$.
- For every k, $\alpha(+,+,k) := \sum_{i>0, j>0} \alpha(i,j,k)$; for every j, $\alpha(+,j,+) := \sum_{i>0,k>0} \alpha(i,j,k)$; and for every i, $\alpha(i,+,+) := \sum_{j>0,k>0} \alpha(i,j,k)$.

- For every k, $\overline{\beta}_{Z,+,+,k} \coloneqq \frac{1}{\alpha(+,+,k)} \sum_{i>0,j>0} \alpha(i,j,k) \cdot \beta_{Z,i,j,k}$, and $\overline{\beta}_{Y,+,j,+}$ and $\overline{\beta}_{X,i,+,+}$ are defined similarly.
- $\overline{\beta}_{X,*,*,*} \coloneqq \sum_{i,j,k} \alpha(i,j,k) \cdot \beta_{X,i,j,k}$ and $\overline{\beta}_{Y,*,*,*}$ and $\overline{\beta}_{Z,*,*,*}$ are defined similarly.
- $\lambda_Z \coloneqq \sum_{i,j,k:i=0 \text{ or } j=0} \alpha(i,j,k) \cdot H(\beta_{Z,i,j,k}) + \sum_k \alpha(+,+,k) \cdot H(\overline{\beta}_{Z,+,+,k})$, and λ_X and λ_Y are defined similarly.

In the following proposition, we will have $\alpha^{(r)}, \beta^{(r)}_{X,i,j,k}, \beta^{(r)}_{Y,i,j,k}, \beta^{(r)}_{Z,i,j,k}$ for every $r \in [3]$. For every $r \in [3]$, we use superscript (r) on variables to denote that they are computed using values of $\alpha^{(r)}, \beta^{(r)}_{X,i,j,k}, \beta^{(r)}_{Y,i,j,k}, \beta^{(r)}_{Z,i,j,k}$.

PROPOSITION 5.1. *$\left(\mathrm{CW}_q^{\otimes 2^{\ell-1}}\right)^{\otimes n}$ can be degenerated into*

$$2^{(A_1E_1+A_2E_2+A_3E_3)n-o(n)}$$

independent copies of a level-ℓ interface tensor with parameter list

$$\left\{\left(n \cdot A_r \cdot \alpha^{(r)}(i,j,k), i, j, k, \beta^{(r)}_{X,i,j,k}, \beta^{(r)}_{Y,i,j,k}, \beta^{(r)}_{Z,i,j,k}\right)\right\}_{r\in[3],\, i+j+k=2^\ell}$$

where

- *$0 \le A_1, A_2, A_3 \le 1, A_1 + A_2 + A_3 = 1$;*
- *$\alpha^{(r)}$ for every $r \in [3]$ is a distribution over $\{(i,j,k) \in \mathbb{Z}^3_{\ge 0} \mid i+j+k = 2^\ell\}$;*
- *For every $W \in \{X,Y,Z\}$, $\beta^{(r)}_{W,i,j,k}$ for $r \in [3], i+j+k = 2^{\ell-1}$ is a level-ℓ complete split distribution;*
- $$\begin{aligned} E_1 &\coloneqq \min\left\{H(\alpha^{(1)}_X) - P^{(1)}_\alpha, H(\alpha^{(1)}_Y) - P^{(1)}_\alpha, H(\overline{\beta}^{(1)}_{Z,*,*,*}) - \lambda^{(1)}_Z\right\},\\ E_2 &\coloneqq \min\left\{H(\alpha^{(2)}_X) - P^{(2)}_\alpha, H(\alpha^{(2)}_Z) - P^{(2)}_\alpha, H(\overline{\beta}^{(2)}_{Y,*,*,*}) - \lambda^{(2)}_Y\right\},\\ E_3 &\coloneqq \min\left\{H(\alpha^{(3)}_Y) - P^{(3)}_\alpha, H(\alpha^{(3)}_Z) - P^{(3)}_\alpha, H(\overline{\beta}^{(3)}_{X,*,*,*}) - \lambda^{(3)}_X\right\}. \end{aligned}$$

REMARK 5.1. *Note that without loss of generality, we can assume that, for every r, i, j, k, and every $L \in \{0,1,2\}^{2^{\ell-1}}$,*

$$\beta^{(r)}_{X,i,0,k}(L) = \beta^{(r)}_{Z,i,0,k}(\vec{2} - L), \quad \beta^{(r)}_{Z,0,j,k}(L) = \beta^{(r)}_{Y,0,j,k}(\vec{2} - L), \quad \beta^{(r)}_{Y,i,j,0}(L) = \beta^{(r)}_{X,i,j,0}(\vec{2} - L),$$

where $\vec{2}$ denotes the length-$(2^{\ell-1})$ vector whose coordinates are all 2, and

$$\beta^{(r)}_{X,i,j,k}(L) = 0 \text{ if } \sum_t L_t \ne i, \quad \beta^{(r)}_{Y,i,j,k}(L) = 0 \text{ if } \sum_t L_t \ne j, \quad \beta^{(r)}_{Z,i,j,k}(L) = 0 \text{ if } \sum_t L_t \ne k,$$

because otherwise, the level-ℓ interface tensor will be the zero tensor and the lemma will follow trivially.

Next, we show Theorem 5.1, which is a corollary of Proposition 5.1.

THEOREM 5.1. *For any $\varepsilon > 0$, $2^{o(n)}$ independent copies of $(\mathrm{CW}_q^{\otimes 2^{\ell-1}})^{\otimes n}$ can be degenerated into*

$$2^{(A_1E_1+A_2E_2+A_3E_3-o_{1/\varepsilon}(1))n-o(n)}$$

independent copies of a level-ℓ ε-interface tensor with parameter list

$$\left\{\left(n \cdot A_r \cdot \alpha^{(r)}(i,j,k), i, j, k, \beta^{(r)}_{X,i,j,k}, \beta^{(r)}_{Y,i,j,k}, \beta^{(r)}_{Z,i,j,k}\right)\right\}_{r\in[3], i+j+k=2^\ell}$$

where the constraints are the same as those in Proposition 5.1.[3]

[3] $o_{1/\varepsilon}(1)$ denotes a function $f(\varepsilon)$ where $f(\varepsilon) \to 0$ as $\varepsilon \to 0$. We also use $o_{1/\varepsilon}(n)$ to denote $o_{1/\varepsilon}(1) \cdot n$.

Here, the differences with Proposition 5.1 are the followings:

- The input becomes multiple independent copies of $\left(\mathrm{CW}_q^{\otimes 2^{\ell-1}}\right)^{\otimes n}$.
- The output tensor becomes independent copies of some level-ℓ ε-interface tensor, instead of level-ℓ interface tensor in Proposition 5.1.
- There is a small $2^{o_{1/\varepsilon}(n)}$ factor loss in the number of independent copies of the level-ℓ ε-interface tensor we can keep.

The high-level idea of the proof is the following: for each copy of $\left(\mathrm{CW}_q^{\otimes 2^{\ell-1}}\right)^{\otimes n}$ in the input, we apply Proposition 5.1 where the target complete split distributions are slightly different in each application (up to ε in L_∞ distance with some specified complete split distributions). Finally, we merge the level-ℓ interface tensors into a level-ℓ ε-interface tensor.

Proof of Theorem 5.1. Let

$$\left\{\xi^{(r)}_{W,i,j,k}\right\}_{r\in[3],W\in\{X,Y,Z\},i+j+k=2^\ell}$$

be a set of level-ℓ complete split distributions whose L_∞ distance with

$$\left\{\beta^{(r)}_{W,i,j,k}\right\}_{r\in[3],W\in\{X,Y,Z\},i+j+k=2^\ell}$$

is at most ε. Furthermore, we require that all entries of $\xi^{(r)}_{W,i,j,k}$ are integral multiples of $\frac{1}{A_r\cdot\alpha(i,j,k)\cdot n}$. Let $\widetilde{\mathcal{D}}$ be the collection of such sets of complete split distributions. For every W,i,j,k, there are $O(n)$ choices for the value of each entry in $\xi^{(r)}_{W,i,j,k}$, and the total number of entries is $3^{2^{\ell-1}}=O(1)$ as ℓ is a constant. Thus, the number of $\xi^{(r)}_{W,i,j,k}$ is bounded by $\mathrm{poly}(n)=2^{o(n)}$, and consequently the number of $\left\{\beta^{(r)}_{W,i,j,k}\right\}_{r\in[3],W\in\{X,Y,Z\},i+j+k=2^\ell}$ (i.e., the size of $\widetilde{\mathcal{D}}$) is also bounded by $2^{o(n)}$. Also, it is not difficult to verify that the level-ℓ ε-interface tensor with parameter list

$$\left\{\left(n\cdot A_r\cdot\alpha^{(r)}(i,j,k),i,j,k,\beta^{(r)}_{X,i,j,k},\beta^{(r)}_{Y,i,j,k},\beta^{(r)}_{Z,i,j,k}\right)\right\}_{r\in[3],i+j+k=2^\ell} \tag{5.2}$$

is the sum of all level-ℓ interface tensors with parameter lists

$$\left\{\left(n\cdot A_r\cdot\alpha^{(r)}(i,j,k),i,j,k,\xi^{(r)}_{X,i,j,k},\xi^{(r)}_{Y,i,j,k},\xi^{(r)}_{Z,i,j,k}\right)\right\}_{r\in[3],i+j+k=2^\ell} \tag{5.3}$$

over all such $\left\{\xi^{(r)}_{W,i,j,k}\right\}\in\widetilde{\mathcal{D}}$.

Let E_1,E_2,E_3 be defined as in Proposition 5.1 applied to complete split distributions $\left\{\beta^{(r)}_{W,i,j,k}\right\}$, and let E'_1,E'_2,E'_3 be defined as in Proposition 5.1 but applied to some complete split distributions $\left\{\xi^{(r)}_{W,i,j,k}\right\}\in\widetilde{\mathcal{D}}$. By Proposition 5.1, each copy of $\left(\mathrm{CW}_q^{\otimes 2^{\ell-1}}\right)^{\otimes n}$ can be degenerated into $2^{(A_1E'_1+A_2E'_2+A_3E'_3)n-o(n)}$ independent copies of the level-ℓ interface tensor with parameter list as in (5.3).

It is not difficult to see that $A_1E_1+A_2E_2+A_3E_3$ is continuous with respect to $\left\{\beta^{(r)}_{W,i,j,k}\right\}$, and because the L_∞ distance between $\left\{\beta^{(r)}_{W,i,j,k}\right\}$ and $\left\{\xi^{(r)}_{W,i,j,k}\right\}$ is at most ε, we get that $A_1E'_1+A_2E'_2+A_3E'_3\geq A_1E_1+A_2E_2+A_3E_3-o_{1/\varepsilon}(1)$.

Thus, $2^{o(n)}$ independent copies of $\left(\mathrm{CW}_q^{\otimes 2^{\ell-1}}\right)^{\otimes n}$ can be degenerated into $2^{(A_1E_1+A_2E_2+A_3E_3-o_{1/\varepsilon}(1))n-o(n)}$ independent copies of a direct sum of all level-ℓ interface tensor with parameter list

$$\left\{\left(n\cdot A_r\cdot\alpha^{(r)}(i,j,k),i,j,k,\xi^{(r)}_{X,i,j,k},\xi^{(r)}_{Y,i,j,k},\xi^{(r)}_{Z,i,j,k}\right)\right\}_{r\in[3],i+j+k=2^\ell}$$

over all such $\left\{\xi^{(r)}_{W,i,j,k}\right\}\in\widetilde{\mathcal{D}}$, and because a direct sum of some tensors can be degenerated into the sum of these tensors, the theorem follows. ∎

The remainder of this section aims to show and analyze an algorithm that proves Proposition 5.1.

5.1 Dividing into regions. Similar to [17], we consider

$$\left(\mathrm{CW}_q^{\otimes 2^{\ell-1}}\right)^{\otimes n} \equiv \left(\mathrm{CW}_q^{\otimes 2^{\ell-1}}\right)^{\otimes A_1\cdot n} \otimes \left(\mathrm{CW}_q^{\otimes 2^{\ell-1}}\right)^{\otimes A_2\cdot n} \otimes \left(\mathrm{CW}_q^{\otimes 2^{\ell-1}}\right)^{\otimes A_3\cdot n}$$

for $A_1, A_2, A_3 \geq 0$ and $A_1 + A_2 + A_3 = 1$. We call each of the three factors of the above tensor product a *region*. For $r \in [3]$, we denote the r-th region as

$$\mathcal{T}^{(r)} \coloneqq \left(\mathrm{CW}_q^{\otimes 2^{\ell-1}}\right)^{\otimes A_r\cdot n}.$$

The idea is to apply asymmetric hashing on the three regions separately. We will use asymmetric hashing that shares level-ℓ Z-blocks in the first region, Y-blocks in the second region, and X-blocks in the third region. Each region will be degenerated into independent copies of a level-ℓ interface tensor and the output will be the tensor product of the independent copies of the three level-ℓ interface tensors from the three regions. Thus we can analyze each region independently and we only give the detailed analysis on the first region as the analysis for the other two regions follow by symmetry.

From now on, we will describe the analysis on $\mathcal{T}^{(1)}$ in which the level-ℓ Z-variable blocks are shared and we will omit the superscript (1) on all variables for conciseness.

5.2 Asymmetric hashing. Recall that α is a distribution on $\{(i,j,k) \in \mathbb{Z}^3_{\geq 0} \mid i+j+k = 2^\ell\}$, i.e., it can be viewed as a distribution on level-ℓ constituent tensors. Recall that α induces marginal distributions $\alpha_X, \alpha_Y, \alpha_Z$. We first zero out X-, Y-, Z-blocks that are not consistent with the marginals $\alpha_X, \alpha_Y, \alpha_Z$ respectively. Let N_{BX} be the number of remaining level-ℓ X-blocks, and it is not difficult to see that

$$N_{\mathrm{BX}} = 2^{H(\alpha_X)\cdot A_1 n \pm o(n)}. \tag{5.4}$$

Similarly, let N_{BY} and N_{BZ} be the number of remaining Y- and Z-blocks, and we have

$$N_{\mathrm{BY}} = 2^{H(\alpha_Y)\cdot A_1 n \pm o(n)}, \quad N_{\mathrm{BZ}} = 2^{H(\alpha_Z)\cdot A_1 n \pm o(n)}. \tag{5.5}$$

Let N_α be the number of remaining block triples that are consistent with α. We have

$$N_\alpha = 2^{H(\alpha)\cdot A_1 n \pm o(n)}. \tag{5.6}$$

Finally, let $N_{\alpha_X,\alpha_Y,\alpha_Z}$ be the number of remaining block triples $X_I Y_J Z_K$.

Claim 5.1. $N_{\alpha_X,\alpha_Y,\alpha_Z} = 2^{(H(\alpha)+P_\alpha)\cdot A_1 n \pm o(n)}$.

Proof. Recall that $P_\alpha = \max_{\alpha' \in D} H(\alpha') - H(\alpha)$ where D is the set of distributions whose marginal distributions on the three dimensions are $\alpha_X, \alpha_Y, \alpha_Z$ respectively.

As we zeroed out X-, Y-, Z-blocks based on $\alpha_X, \alpha_Y, \alpha_Z$ respectively, all remaining block triples are consistent with one of the distributions $\alpha' \in D$. Additionally, $\alpha'(i,j,k) \cdot A_1 \cdot n$ must be an integer for every i,j,k. Let us denote the set of distributions satisfying such properties as D'.

Thus, $N_{\alpha_X,\alpha_Y,\alpha_Z} = \sum_{\alpha' \in D'} 2^{H(\alpha')\cdot A_1 n \pm o(n)}$. As $|D'| = \mathrm{poly}(n)$, we have that

$$N_{\alpha_X,\alpha_Y,\alpha_Z} = 2^{(\max_{\alpha' \in D'} H(\alpha'))\cdot A_1 n \pm o(n)}.$$

When n approaches ∞, the difference between $\max_{\alpha' \in D'} H(\alpha')$ and $\max_{\alpha' \in D} H(\alpha')$ will approach 0, as the entropy function H is continuous. Thus,

$$N_{\alpha_X,\alpha_Y,\alpha_Z} = 2^{(\max_{\alpha' \in D} H(\alpha'))\cdot A_1 n \pm o(n)} = 2^{(H(\alpha)+P_\alpha)\cdot A_1 n \pm o(n)}. \qquad \square$$

Let $M \in [M_0, 2M_0]$ be a prime number for some integer M_0. The value of M_0 is yet to be fixed, but we first require that

$$M_0 \geq 8 \cdot \max\left\{ \frac{N_{\alpha_X,\alpha_Y,\alpha_Z}}{N_{\mathrm{BX}}}, \frac{N_{\alpha_X,\alpha_Y,\alpha_Z}}{N_{\mathrm{BY}}} \right\}. \tag{5.7}$$

One additional term that lower bounds M_0 will be mentioned later.

We independently pick uniformly random elements $b_0, \{w_t\}_{t=0}^n \in \{0, \dots, M-1\}$, and define the following hash functions $h_X, h_Y, h_Z : \{0, \dots, 2^\ell\}^n \to \{0, \dots, M-1\}$:

$$h_X(I) = b_0 + \left(\sum_{t=1}^{n} w_t \cdot I_t \right) \bmod M,$$

$$h_Y(J) = b_0 + \left(w_0 + \sum_{t=1}^{n} w_t \cdot J_t \right) \bmod M,$$

$$h_Z(K) = b_0 + \frac{1}{2}\left(w_0 + \sum_{t=1}^{n} w_t \cdot (2^\ell - K_t) \right) \bmod M.$$

Let B be a Salem-Spencer subset of $\{0, \dots, M-1\}$ that has size $M^{1-o(1)}$ and does not contain any nontrivial 3-term arithmetic progressions (modulo M). Then we zero out all blocks X_I with $h_X(I) \notin B$, Y_J with $h_Y(J) \notin B$, and Z_K with $h_Z(K) \notin B$.

For every block triple $X_I Y_J Z_K$ in $\mathcal{T}$, we have that $X_t + Y_t + Z_t = 2^\ell$ for every $t \in [n]$. Therefore, it is not difficult to verify that $h_X(I) + h_Y(J) \equiv 2h_Z(K) \pmod M$. In order for $h_X(I), h_Y(J), h_Z(K) \in B$, we must have $h_X(I) = h_Y(J) = h_Z(K) = b$ for some b, because B does not contain any nontrivial 3-term arithmetic progression (modulo M). We say that triples $X_I Y_J Z_K$ with $h_X(I) = h_Y(J) = h_Z(K) = b$ are contained in bucket b.

For every bucket b, if it contains two level-ℓ triples $X_I Y_J Z_K$ and $X_I Y_{J'} Z_{K'}$ that share the same X-block, then we zero out X_I. Similarly, if a bucket contains two level-ℓ triples $X_I Y_J Z_K$ and $X_{I'} Y_J Z_{K'}$ that share the same level-ℓ Y-block, then we zero out Y_J. We repeatedly perform the previous zeroing-outs so that eventually, all remaining triples in the same bucket do not share X- or Y-blocks. As each level-ℓ block triple in $\mathcal{T}$ must belong to some bucket, we get that all remaining triples do not share X- or Y-blocks, i.e., each level-ℓ block X_I or Y_J is in a unique level-ℓ block triple. For every level-ℓ block X_I (or Y_J), we check whether the unique triple containing it is consistent with the distribution α; if not, we zero out X_I (or Y_J). We call the tensor after this step $\mathcal{T}_{\text{hash}}$.

Claim 5.2 (Implicit in [13], see also [17]). *For a block triple $X_I Y_J Z_K \in \mathcal{T}$, and for every $b \in \{0, \dots, M-1\}$,*

$$\Pr\Big[h_X(I) = h_Y(J) = h_Z(K) = b\Big] = \frac{1}{M^2}.$$

Furthermore, for two different block triples $X_I Y_J Z_K, X_I Y_{J'} Z_{K'} \in \mathcal{T}$ that share the same X-block, and for every $b \in \{0, \dots, M-1\}$,

$$\Pr\Big[h_X(I) = h_Y(J') = h_Z(K') = b \;\Big|\; h_X(I) = h_Y(J) = h_Z(K) = b\Big] = \frac{1}{M}.$$

This also holds analogously for different block triples that share the same Y-block or Z-block.

Claim 5.3. *For every $b \in B$ and for every level-ℓ block triple $X_I Y_J Z_K \in \mathcal{T}$ that is consistent with α, the probability that $X_I Y_J Z_K$ remains in $\mathcal{T}_{\text{hash}}$ conditioned on $h_X(I) = h_Y(J) = h_Z(K) = b$ is $\geq \frac{3}{4}$.*

Proof. The only way that $X_I Y_J Z_K$ does not remain in $\mathcal{T}_{\text{hash}}$ conditioned on $h_X(I) = h_Y(J) = h_Z(K) = b$ is when some other block triples that share the same X-block or the same Y-block are hashed to the same bucket b.

Right before the hashing step, the total number of block triples remaining is $N_{\alpha_X, \alpha_Y, \alpha_Z}$, and the number of X-blocks is N_{BX}. By symmetry, each X-block is in the same number of block triples, which is $\frac{N_{\alpha_X, \alpha_Y, \alpha_Z}}{N_{\text{BX}}}$. Thus, the total number of block triples that share the same X-block as $X_I Y_J Z_K$ is $\frac{N_{\alpha_X, \alpha_Y, \alpha_Z}}{N_{\text{BX}}} - 1$. For each of them, the probability that they are hashed to the same bucket b with $X_I Y_J Z_K$ is $\frac{1}{M}$ by Claim 5.2. Therefore, by union bound, the probability that any of them is hashed to the same bucket with $X_I Y_J Z_K$ is at most

$$\frac{N_{\alpha_X, \alpha_Y, \alpha_Z}}{M \cdot N_{\text{BX}}} \leq \frac{N_{\alpha_X, \alpha_Y, \alpha_Z}}{M_0 \cdot N_{\text{BX}}} \overset{Eq.\ (5.7)}{\leq} \frac{1}{8}.$$

Similarly, the probability that any block triple that shares the same level-ℓ Y-block is mapped to the same bucket as $X_I Y_J Z_K$ is at most $\frac{1}{8}$. By union bound, the probability that $X_I Y_J Z_K$ will be zeroed out is $\leq \frac{1}{4}$. ∎

Claim 5.4. *The expected number of level-ℓ block triples in $\mathcal{T}_{\text{hash}}$ is at least*

$$N_\alpha \cdot M_0^{-1-o(1)}.$$

Proof. For every level-ℓ block triple $X_I Y_J Z_K \in \mathcal{T}$ that is consistent with α, and for every $b \in B$, the probability that $h_X(I) = h_Y(J) = h_Z(K) = b$ is $\frac{1}{M^2}$ by Claim 5.2. Also, by Claim 5.3, $X_I Y_J Z_K$ will remain in $\mathcal{T}_{\text{hash}}$ with probability $\geq \frac{3}{4} \cdot \frac{1}{M^2}$.

Summing over all block triples $X_I Y_J Z_K$ and all $b \in B$, we get that the expected number of block triples in $\mathcal{T}_{\text{hash}}$ is at least

$$N_\alpha \cdot |B| \cdot \frac{3}{4} \cdot \frac{1}{M^2} = N_\alpha \cdot M_0^{-1-o(1)}.$$

∎

5.3 Compatibility zero-out I. Recall that $\{\beta_{X,i,j,k}, \beta_{Y,i,j,k}, \beta_{Z,i,j,k}\}_{i+j+k=2^\ell}$ are level-ℓ complete split distributions for the X-, Y-, Z-blocks.

Let

$$S^{(I,J,K)}_{i,j,k} \coloneqq \{t \in [n] \mid I_t = i, J_t = j, K_t = k\},$$

and

$$S^{(K)}_{*,*,k} \coloneqq \{t \in [n] \mid K_t = k\}.$$

If clear from the context, we will drop the superscript (I, J, K) or (K).

Recall that in $\mathcal{T}_{\text{hash}}$, every level-$\ell$ block X_I is in a unique block triple $X_I Y_J Z_K$. For every level-1 block $X_{\hat{I}} \in X_I$, we will zero out $X_{\hat{I}}$ if $\mathsf{split}(\hat{I}, S_{i,j,k}) \neq \beta_{X,i,j,k}$ for any i, j, k (recall the definition of $\mathsf{split}(\hat{I}, S_{i,j,k})$ in Definition 3.2). Similarly, every level-ℓ block Y_J is in a unique block triple, and we zero out every $Y_{\hat{J}} \in Y_J$ where $\mathsf{split}(\hat{J}, S_{i,j,k}) \neq \beta_{Y,i,j,k}$ for any i, j, k.

We can not perform the same zeroing out for Z-variables, because in $\mathcal{T}_{\text{hash}}$ each level-ℓ Z-block is not in a unique block triple and $S_{i,j,k}$ is not well-defined just given the Z-block. Instead, for every level-1 block $Z_{\hat{K}} \in Z_K$, we zero out $Z_{\hat{K}}$ if $\mathsf{split}(\hat{K}, S_{*,*,k}) \neq \overline{\beta}_{Z,*,*,k}$ for any k, where

$$\overline{\beta}_{Z,*,*,k} = \frac{1}{\sum_{i+j=2^\ell-k} \alpha(i,j,k)} \sum_{i+j=2^\ell-k} \alpha(i,j,k) \cdot \beta_{Z,i,j,k}$$

is the average complete split distribution for constituent tensors whose third coordinate is k.

We call the tensor after the previous zeroing-outs $\mathcal{T}_{\text{comp}}$.

Next, we are ready to define the notion of compatibility. The notion is adapted from [17], which is a crucial ingredient in their analysis (and ours).

Definition 5.1 (Compatibility). *For some I, J, K, a level-1 block $Z_{\hat{K}} \in Z_K$ is compatible with a level-ℓ triple $X_I Y_J Z_K$ if*

1. *For every $(i, j, k) \in \mathbb{Z}^3_{\geq 0}$ with $i + j + k = 2^\ell, i = 0$ or $j = 0$, $\mathsf{split}(\hat{K}, S_{i,j,k}) = \beta_{Z,i,j,k}$.*
2. *For every index $k \in \{0, 1, \ldots, 2^\ell\}$, $\mathsf{split}(\hat{K}, S_{*,*,k}) = \overline{\beta}_{Z,*,*,k}$.*

Claim 5.5. *In $\mathcal{T}_{\text{comp}}$, for every level-1 block triple $X_{\hat{I}} Y_{\hat{J}} Z_{\hat{K}}$ and the level-ℓ block triple $X_I Y_J Z_K$ that contains it, $Z_{\hat{K}}$ is compatible with $X_I Y_J Z_K$.*

Proof. First of all, Item 2 is clearly satisfied, because we zeroed out every $\hat{K}$ with $\mathsf{split}(\hat{K}, S_{*,*,k}) \neq \overline{\beta}_{Z,*,*,k}$ for any k. Next, we show that Item 1 is also satisfied.

Recall that we zeroed out all $X_{\hat{I}}$ where $\mathsf{split}(\hat{I}, S_{i,j,k}) \neq \beta_{X,i,j,k}$ for any i, j, k. Let $(i, j, k) \in \mathbb{Z}^3_{\geq 0}$ where $i + j + k = 2^\ell$ and $j = 0$. As $X_{\hat{I}} Y_{\hat{J}} Z_{\hat{K}}$ remains in $\mathcal{T}_{\text{comp}}$, $\mathsf{split}(\hat{I}, S_{i,j,k}) = \beta_{X,i,j,k}$. Because $j = 0$, $J_t = 0$ for every $t \in S_{i,j,k}$, which implies that $\left(\hat{J}_{(t-1)\cdot 2^{\ell-1}+1}, \hat{J}_{(t-1)\cdot 2^{\ell-1}+2}, \ldots, \hat{J}_{t\cdot 2^{\ell-1}}\right) = \vec{0}$. As $\hat{I}_{\hat{t}} + \hat{J}_{\hat{t}} + \hat{K}_{\hat{t}} = 2$ for every $\hat{t}$, we have that

$$\left(\hat{K}_{(t-1)\cdot 2^{\ell-1}+1}, \hat{K}_{(t-1)\cdot 2^{\ell-1}+2}, \ldots, \hat{K}_{t\cdot 2^{\ell-1}}\right) = \vec{2} - \left(\hat{I}_{(t-1)\cdot 2^{\ell-1}+1}, \hat{I}_{(t-1)\cdot 2^{\ell-1}+2}, \ldots, \hat{I}_{t\cdot 2^{\ell-1}}\right)$$

for every $t \in S_{i,j,k}$. Thus, for every $L \in \{0,1,2\}^{2^{\ell-1}}$, the proportion of L appearing in $\big(\hat{I}_{(t-1)\cdot 2^{\ell-1}+1}, \hat{I}_{(t-1)\cdot 2^{\ell-1}+2}, \ldots, \hat{I}_{t\cdot 2^{\ell-1}}\big)$ over $t \in S_{i,j,k}$ is exactly the proportion of $\vec{2} - L$ appearing in $\big(\hat{K}_{(t-1)\cdot 2^{\ell-1}+1}, \hat{K}_{(t-1)\cdot 2^{\ell-1}+2}, \ldots, \hat{K}_{t\cdot 2^{\ell-1}}\big)$. In other words, $\mathsf{split}(\hat{K}, S_{i,j,k})(L) = \mathsf{split}(\hat{I}, S_{i,j,k})(\vec{2} - L) = \beta_{X,i,j,k}(\vec{2} - L)$. By Remark 5.1, this implies that $\mathsf{split}(\hat{K}, S_{i,j,k}) = \beta_{Z,i,j,k}$.

We can show that $\mathsf{split}(\hat{K}, S_{i,j,k}) = \beta_{Z,i,j,k}$ with $i = 0$ similarly. ∎

5.4 Compatibility zero-out II: unique triple. In this step, we zero out level-1 Z-blocks that are compatible with more than one level-ℓ triples. To do so, we check if each level-1 Z-block $Z_{\hat{K}}$ is compatible with multiple level-ℓ triples. If so, we zero it out and it becomes a "hole". Note that after this step, each remaining level-1 Z-block $Z_{\hat{K}} \in Z_K$ is compatible with a unique level-ℓ triple (X_I, Y_J, Z_K) containing it.

5.5 Usefulness zero-out. Next, we further zero out some level-1 Z-blocks using the following definition of *usefulness*.

DEFINITION 5.2 (Usefulness). *For a level-1 block $Z_{\hat{K}}$ and a level-ℓ triple $X_I Y_J Z_K$ containing it, if for all (i,j,k) we have $\mathsf{split}(\hat{K}, S_{i,j,k}) = \beta_{Z,i,j,k}$, then we say that $Z_{\hat{K}}$ is* useful *for $X_I Y_J Z_K$.*

For each $Z_{\hat{K}}$, it appears in a unique triple $X_I Y_J Z_K$ by the previous zeroing out. Furthermore, if $Z_{\hat{K}}$ is not useful for this triple, we zero out $Z_{\hat{K}}$. We call the current tensor $\mathcal{T}_{\text{useful}}$.

If there is no hole, then the subtensor of the remaining tensor over $X_I Y_J Z_K$ is isomorphic to

$$\mathcal{T}^* = \bigotimes_{i+j+k=2^\ell} T_{i,j,k}^{\otimes A_1 \cdot \alpha(i,j,k)\cdot n}[\beta_{X,i,j,k}, \beta_{Y,i,j,k}, \beta_{Z,i,j,k}],$$

i.e., it is the level-ℓ interface tensor with parameter list

$$\{(A_1 \cdot \alpha(i,j,k)\cdot n, i, j, k, \beta_{X,i,j,k}, \beta_{Y,i,j,k}, \beta_{Z,i,j,k})\}_{i+j+k=2^\ell}.$$

More formally:

CLAIM 5.6. *For any level-ℓ block triple $X_I Y_J Z_K$ contained in $\mathcal{T}_{\text{comp}}$ (or equivalently, $\mathcal{T}_{\text{hash}}$), the subtensor of $\mathcal{T}_{\text{useful}}$ restricted to blocks X_I, Y_J, Z_K is a subtensor of $\mathcal{T}^*$, where the missing variables in this subtensor are exactly those in level-1 blocks $Z_{\hat{K}}$ that are compatible with multiple level-ℓ triples in $\mathcal{T}_{\text{comp}}$.*

Proof. Initially,

$$\mathcal{T}_{\text{hash}}|_{X_I Y_J Z_K} \equiv \bigotimes_{i+j+k=2^\ell} T_{i,j,k}^{\otimes A_1 \cdot \alpha(i,j,k)\cdot n}.$$

To show $\mathcal{T}_{\text{useful}}|_{X_I Y_J Z_K}$ is a subtensor of $\mathcal{T}^*$, it suffices to show that the level-1 X-blocks (Y-blocks or Z-blocks resp.) remaining in $\mathcal{T}_{\text{useful}}|_{X_I Y_J Z_K}$ have the property that $\mathsf{split}(\hat{I}, S_{i,j,k}) = \beta_{X,i,j,k}$ ($\mathsf{split}(\hat{J}, S_{i,j,k}) = \beta_{Y,i,j,k}$ or $\mathsf{split}(\hat{K}, S_{i,j,k}) = \beta_{Z,i,j,k}$ resp.) for every i,j,k. This is true because we enforced these constraints on X- and Y-blocks in the compatibility zeroing-out step, and enforced the constraints on Z-blocks by zeroing out $Z_{\hat{K}}$ that is not useful for the unique level-ℓ triple that contains it. Furthermore, these are the only constraints we have on the level-1 X- and Y-blocks, so the set of X- and Y-variables in $\mathcal{T}_{\text{useful}}|_{X_I Y_J Z_K}$ is the same as that in $\mathcal{T}^*$. It remains to analyze which level-1 Z-blocks are missing in $\mathcal{T}_{\text{useful}}|_{X_I Y_J Z_K}$.

There are three constraints we enforced on level-1 Z-blocks:

1. In the compatibility zeroing-out, we enforced that for every index $k \in \{0,1,\ldots,2^\ell\}$, $\mathsf{split}(\hat{K}, S_{*,*,k}) = \overline{\beta}_{Z,*,*,k}$.
2. In the unique triple zeroing-out, we zeroed out $Z_{\hat{K}}$ that is compatible with multiple level-ℓ triples.
3. In the unique triple zeroing-out, we zeroed out $Z_{\hat{K}}$ that is not useful for the unique level-ℓ triple $X_I Y_J Z_K$ that contains it. Thus, we will have that $\mathsf{split}(\hat{K}, S_{i,j,k}) = \beta_{Z,i,j,k}$ for every i,j,k if $Z_{\hat{K}}$ remains.

The third constraint implies the first constraint, because if the third constraint holds, then for every k,

$$\begin{aligned}\mathsf{split}(\hat{K}, S_{*,*,k}) &= \frac{1}{\sum_{i,j}\alpha(i,j,k)} \sum_{i+j=2^\ell-k} \alpha(i,j,k)\cdot \mathsf{split}(\hat{K}, S_{i,j,k}) \\ &= \frac{1}{\sum_{i,j}\alpha(i,j,k)} \sum_{i+j=2^\ell-k} \alpha(i,j,k)\cdot \beta_{Z,i,j,k} = \overline{\beta}_{Z,*,*,k}.\end{aligned}$$

Therefore, we can ignore the first condition. As a result, the set of level-1 Z-blocks not in $\mathcal{T}_{\text{useful}}|_{X_I Y_J Z_K}$ but in $\mathcal{T}^*$ are exactly those that are compatible with multiple block triples in $\mathcal{T}_{\text{comp}}$. ∎

Also, note that for different remaining block triples $X_I Y_J Z_K$, $\mathcal{T}_{\text{useful}}|_{X_I Y_J Z_K}$ are level-1-independent, i.e., they do not share the same level-1 blocks. This is because X_I and Y_J are already in unique level-ℓ triples in $\mathcal{T}_{\text{hash}}$; for every level-1 block $Z_{\hat{K}}$, Claim 5.5 shows that $Z_{\hat{K}}$ is compatible with every level-ℓ triple $X_I Y_J Z_K$ containing it, and then we zeroed out $Z_{\hat{K}}$ that are compatible with multiple triples. Thus, every remaining $Z_{\hat{K}}$ in $\mathcal{T}_{\text{useful}}$ is contained a unique level-ℓ triple as well. As a result, we can write

$$\mathcal{T}_{\text{useful}} = \bigoplus_{X_I Y_J Z_K \text{ remaining}} \mathcal{T}_{\text{useful}}|_{X_I Y_J Z_K} \tag{5.8}$$

as a direct sum of broken copies of $\mathcal{T}^*$.

5.6 Fixing holes. Next, we analyze the fraction of holes in the broken copies of $\mathcal{T}^*$ contained in $\mathcal{T}_{\text{useful}}$. To do so, we define the following notion of *typicalness*, which will then be used to define the quantity p_{comp}:

Definition 5.3 (Typicalness). *A level-1 Z-block $Z_{\hat{K}}$ in some level-ℓ Z-block Z_K is* typical *if* $\mathsf{split}(\hat{K}, S_{*,*,k}) = \overline{\beta}_{Z,*,*,k}$ *for every k. When Z_K is consistent with α_Z, this condition can be equivalently written as* $\mathsf{split}(\hat{K}, [A_1 n]) = \overline{\beta}_{Z,*,*,*}$, *where we recall that* $\overline{\beta}_{Z,*,*,*} = \sum_{i,j,k}\alpha(i,j,k)\cdot\beta_{Z,i,j,k}$.

Definition 5.4 (p_{comp}). *For fixed $Z_{\hat{K}}$ and Z_K where $Z_{\hat{K}} \in Z_K$ and $Z_{\hat{K}}$ is typical, p_{comp} is the probability that a uniformly random block triple $X_I Y_J Z_K$ consistent with α is compatible with $\hat{K}$.*

By symmetry, this probability is the same for different $Z_{\hat{K}}$ and Z_K where $Z_{\hat{K}} \in Z_K$ and $Z_{\hat{K}}$ is typical, so p_{comp} is well-defined. Since holes only arise when some $Z_{\hat{K}}$ is compatible with multiple triples, the value of p_{comp} is closely related to the fraction of holes, and is given by the following claim.

Claim 5.7. *The value of p_{comp} is*

$$2^{\left(\lambda_Z - H(\overline{\beta}_{Z,*,*,*}) + H(\alpha_Z)\right)A_1\cdot n \pm o(n)},$$

where we recall that

$$\lambda_Z = \sum_{i,j,k: i=0 \text{ or } j=0} \alpha(i,j,k)\cdot H(\beta_{Z,i,j,k}) + \sum_k \alpha(+,+,k)\cdot H(\overline{\beta}_{Z,+,+,k}),$$

$$\alpha(+,+,k) = \sum_{i,j>0}\alpha(i,j,k), \quad \overline{\beta}_{Z,+,+,k} = \frac{1}{\alpha(+,+,k)}\sum_{i,j>0}\alpha(i,j,k)\cdot\beta_{Z,i,j,k},$$

and

$$\overline{\beta}_{Z,*,*,*} = \sum_{i,j,k}\alpha(i,j,k)\cdot\beta_{Z,i,j,k}.$$

Proof. By symmetry, it suffices to compute the following two quantities, and p_{comp} will be the ratio between them: (1) the number of tuples $(I,J,K,\hat{K})$ where $X_I Y_J Z_K$ is consistent with α, $\hat{K} \in K$, $Z_{\hat{K}}$ is typical, and $Z_{\hat{K}}$ is compatible with $X_I Y_J Z_K$; (2) the number of $(I,J,K,\hat{K})$ where $X_I Y_J Z_K$ is consistent with α, $\hat{K} \in K$, and $Z_{\hat{K}}$ is typical.

We first compute the second quantity. First, the number of typical $Z_{\hat{K}}$ is $2^{H(\overline{\beta}_{Z,*,*,*})\cdot A_1 \cdot n \pm o(n)}$. Each of these $Z_{\hat{K}}$ uniquely determines a level-ℓ block Z_K. Also, for each Z_K, the number of block triples $X_I Y_J Z_K$ consistent with α is $\frac{N_\alpha}{N_{\mathrm{BZ}}} = 2^{(H(\alpha)-H(\alpha_Z))\cdot A_1 \cdot n \pm o(n)}$. Therefore, the second quantity is

$$2^{(H(\overline{\beta}_{Z,*,*,*})+H(\alpha)-H(\alpha_Z))\cdot A_1 \cdot n \pm o(n)}. \tag{5.9}$$

Next, we compute the first quantity, which is the number of $(I, J, K, \hat{K})$ where $X_I Y_J Z_K$ is consistent with α, $\hat{K} \in K$, $Z_{\hat{K}}$ is typical, and $Z_{\hat{K}}$ is compatible with $X_I Y_J Z_K$. By Item 2 in Definition 5.1, if $Z_{\hat{K}}$ is compatible with any level-ℓ block triple, then it is typical. Thus, we can drop the condition that $Z_{\hat{K}}$ is typical, and equivalently count the number of $(I, J, K, \hat{K})$ where $X_I Y_J Z_K$ is consistent with α, $\hat{K} \in K$, and $Z_{\hat{K}}$ is compatible with $X_I Y_J Z_K$.

First, the number of block triples $X_I Y_J Z_K$ consistent with α is N_α. Then, for each such block triple, we count the number of $Z_{\hat{K}} \in Z_K$ that is compatible with it. If we fix some $X_I Y_J Z_K$, then we also have fixed the values of $S_{i,j,k}$ for all i, j, k. Then we can rewrite the condition for $Z_{\hat{K}}$ being compatible with $X_I Y_J Z_K$ equivalently as follows:

Definition 5.5 (Compatibility'). *For level-ℓ triple $X_I Y_J Z_K$ consistent with α, a level-1 block $Z_{\hat{K}} \in Z_K$ is compatible with $X_I Y_J Z_K$ if*

- *For every $\{(i,j,k) \in \mathbb{Z}^3_{\geq 0} \mid i+j+k = 2^\ell,\ i = 0 \text{ or } j = 0\}$, $\mathsf{split}(\hat{K}, S_{i,j,k}) = \beta_{Z,i,j,k}$. (This is exactly Item 1 in Definition 5.1).*
- *For every k, let $S_{+,+,k} \coloneqq \bigcup_{i>0,j>0} S_{i,j,k}$. Then $\mathsf{split}(\hat{K}, S_{+,+,k}) = \overline{\beta}_{Z,+,+,k}$.*

Item 1 and the second condition above imply the original condition $\mathsf{split}(\hat{K}, S_{*,*,k}) = \overline{\beta}_{Z,*,*,k}$ in Item 2, because

$$\begin{aligned}
\mathsf{split}(\hat{K}, S_{*,*,k}) &= \frac{1}{\sum_{i,j\geq 0} \alpha(i,j,k)} \sum_{i,j\geq 0} \alpha(i,j,k) \cdot \mathsf{split}(\hat{K}, S_{i,j,k}) \\
&= \frac{1}{\sum_{i,j\geq 0} \alpha(i,j,k)} \left(\sum_{i,j>0} \alpha(i,j,k) \cdot \mathsf{split}(\hat{K}, S_{i,j,k}) + \sum_{i=0 \text{ or } j=0} \alpha(i,j,k) \cdot \mathsf{split}(\hat{K}, S_{i,j,k}) \right) \\
&= \frac{1}{\sum_{i,j\geq 0} \alpha(i,j,k)} \left(\sum_{i,j>0} \alpha(i,j,k) \cdot \mathsf{split}(\hat{K}, S_{+,+,k}) + \sum_{i=0 \text{ or } j=0} \alpha(i,j,k) \cdot \beta_{Z,i,j,k} \right) \\
&= \frac{1}{\sum_{i,j\geq 0} \alpha(i,j,k)} \left(\sum_{i,j>0} \alpha(i,j,k) \cdot \overline{\beta}_{Z,+,+,k} + \sum_{i=0 \text{ or } j=0} \alpha(i,j,k) \cdot \beta_{Z,i,j,k} \right) \\
&= \frac{1}{\sum_{i,j\geq 0} \alpha(i,j,k)} \sum_{i,j\geq 0} \alpha(i,j,k) \cdot \beta_{Z,i,j,k} \ = \ \overline{\beta}_{Z,*,*,k}.
\end{aligned}$$

Similarly, Item 1 and Item 2 together imply the second condition in Definition 5.5. Therefore, Definition 5.5 is an equivalent definition of compatibility.

In Definition 5.5, there are constraints on the complete split distributions of $\hat{K}$ on some disjoint subsets of $[A_1 n]$. Therefore, we can count the number of valid subsequences of $\hat{K}$ for each of these subsets of indices, and multiply them together. For every $(i,j,k) \in \mathbb{Z}^3_{\geq 0}$ where $i+j+k = 2^\ell$ while $i = 0$ or $j = 0$, we require that $\mathsf{split}(\hat{K}, S_{i,j,k}) = \beta_{Z,i,j,k}$, so the number of possibilities of $\hat{K}$ on the subset of indices $S_{i,j,k}$ is $2^{H(\beta_{Z,i,j,k})\cdot |S_{i,j,k}| \pm o(n)} = 2^{H(\beta_{Z,i,j,k})\cdot \alpha(i,j,k)\cdot A_1 n \pm o(n)}$. For every k, we require that $\mathsf{split}(\hat{K}, S_{+,+,k}) = \overline{\beta}_{Z,+,+,k}$, so the number of possibilities of $\hat{K}$ on $S_{+,+,k}$ is $2^{H(\overline{\beta}_{Z,+,+,k})\cdot |S_{+,+,k}| \pm o(n)} = 2^{H(\overline{\beta}_{Z,+,+,k})\cdot \alpha(+,+,k)\cdot A_1 n \pm o(n)}$. Overall, the number of possible compatible $\hat{K}$, multiplied by the number of block triples $X_I Y_J Z_K$, is

$$N_\alpha \cdot \prod_{\substack{i,j,k \\ i=0 \text{ or } j=0}} 2^{H(\overline{\beta}_{Z,i,j,k})\cdot \alpha(i,j,k)\cdot A_1 n \pm o(n)} \cdot \prod_k 2^{H(\overline{\beta}_{Z,+,+,k})\cdot \alpha(+,+,k)\cdot A_1 n \pm o(n)} = 2^{(H(\alpha)+\lambda_Z)\cdot A_1 n \pm o(n)}. \tag{5.10}$$

Finally, as mentioned, p_{comp} is the ratio between (5.10) and (5.9), so

$$p_{\text{comp}} = \left(2^{(H(\alpha)+\lambda_Z)\cdot A_1 n \pm o(n)}\right) \Big/ \left(2^{(H(\bar{\beta}_{Z,*,*,*})+H(\alpha)-H(\alpha_Z))\cdot A_1\cdot n\pm o(n)}\right) = 2^{\left(\lambda_Z - H(\bar{\beta}_{Z,*,*,*})+H(\alpha_Z)\right)A_1\cdot n\pm o(n)}$$

as desired. ∎

Claim 5.8. *For every $b \in B$, every level-ℓ block triple $X_I Y_J Z_K$ consistent with α, and for each typical $Z_{\hat{K}} \in Z_K$, the probability that $Z_{\hat{K}}$ is compatible with multiple triples in $\mathcal{T}_{\text{comp}}$ is at most*

$$\frac{N_\alpha \cdot p_{\text{comp}}}{N_{\text{BZ}} \cdot M_0},$$

conditioned on $h_X(I) = h_Y(J) = h_Z(K) = b$.

Proof. By the definition of p_{comp}, the total number of level-ℓ block triples $X_{I'}Y_{J'}Z_K$ that is compatible with $Z_{\hat{K}}$ is $\frac{N_\alpha}{N_{\text{BZ}}}\cdot p_{\text{comp}}$. For each $X_{I'}Y_{J'}Z_K$ different from $X_I Y_J Z_K$, the probability that $X_{I'}Y_{J'}Z_K$ is mapped to the same bucket b as $X_I Y_J Z_K$ is $\frac{1}{M}$ by Claim 5.2. Thus, by the union bound, the probability that any of them is mapped to the same bucket as $X_I Y_J Z_K$ is upper bounded by $\frac{N_\alpha}{N_{\text{BZ}}}\cdot p_{\text{comp}}\cdot\frac{1}{M} \le \frac{N_\alpha\cdot p_{\text{comp}}}{N_{\text{BZ}}\cdot M_0}$. Furthermore, if none of them are mapped to the same bucket as $X_I Y_J Z_K$, then $Z_{\hat{K}}$ is compatible with a unique triple $X_I Y_J Z_K$ in $\mathcal{T}_{\text{comp}}$, so the claim follows. ∎

Recall that we require M_0 to be at least $8\cdot\max\left\{\frac{N_{\alpha_X,\alpha_Y,\alpha_Z}}{N_{\text{BX}}}, \frac{N_{\alpha_X,\alpha_Y,\alpha_Z}}{N_{\text{BY}}}\right\}$. Now, we add another (and final) constraint: $M_0 \ge \frac{N_\alpha\cdot p_{\text{comp}}}{N_{\text{BZ}}}\cdot 80N$. That is, we will set M_0 to be

$$\begin{aligned}&\max\left\{\frac{8N_{\alpha_X,\alpha_Y,\alpha_Z}}{N_{\text{BX}}}, \frac{8N_{\alpha_X,\alpha_Y,\alpha_Z}}{N_{\text{BY}}}, \frac{N_\alpha\cdot p_{\text{comp}}}{N_{\text{BZ}}}\cdot 80N\right\}\\ &= 2^{\max\{H(\alpha)-P_\alpha-H(\alpha_X),\, H(\alpha)-P_\alpha-H(\alpha_Y),\, H(\alpha)+\lambda_Z-H(\bar{\beta}_{Z,*,*,*})\}\cdot A_1\cdot n\pm o(n)}.\end{aligned}$$

Now, for every $b \in B$ and every level-ℓ block triple $X_I Y_J Z_K$ that is consistent with α with $h_X(I) = h_Y(J) = h_Z(K) = b$,

1. by Claim 5.3, it remains in $\mathcal{T}_{\text{hash}}$ with probability $\ge \frac{3}{4}$;
2. by Claim 5.8, linearity of expectation and Markov's inequality, among $Z_{\hat{K}} \in Z_K$ that is useful for $X_I Y_J Z_K$ (this implies that $Z_{\hat{K}}$ is typical, so we could apply Claim 5.8), the fraction of $Z_{\hat{K}}$ that becomes a hole in $\mathcal{T}_{\text{useful}}$ is at most $10/80N = \frac{1}{8N}$ with probability at least $9/10$.

Therefore, by the union bound, with constant probability, the subtensor of $\mathcal{T}_{\text{useful}}$ over X_I, Y_J, Z_K is a copy of $\mathcal{T}^*$ whose fraction of holes does not exceed $1/8N$. The expected number of $X_I Y_J Z_K$ with $h_X(I) = h_Y(J) = h_Z(K) = b$ over all $b \in B$ is $N_\alpha \cdot M^{-1-o(1)}$, so overall, $\mathcal{T}_{\text{useful}}$ contains $N_\alpha \cdot M^{-1-o(1)}$ copies of $\mathcal{T}^*$ whose fraction of holes is $1/8N$.

By Corollary 4.1, we can degenerate them into $N_\alpha \cdot M^{-1-o(1)}$ unbroken copies of $\mathcal{T}^*$.

5.7 Summary. So far, we have degenerated $\left(\text{CW}_q^{\otimes 2^{\ell-1}}\right)^{\otimes A_1\cdot n}$ into $\ge N_\alpha \cdot M_0^{-1-o(1)}$ copies of a level-ℓ interface tensor $\mathcal{T}^*$ with parameter list

$$\left\{\left(n\cdot A_1\cdot\alpha^{(1)}(i,j,k), i, j, k, \beta^{(1)}_{X,i,j,k}, \beta^{(1)}_{Y,i,j,k}, \beta^{(1)}_{Z,i,j,k}\right)\right\}_{i+j+k=2^\ell}.$$

By plugging in the bounds of N_α and M_0, we see that the number of copies we obtained (in the first region) is

$$2^{A_1 n\cdot\min\left\{H(\alpha^{(1)}_X)-P^{(1)}_\alpha,\, H(\alpha^{(1)}_Y)-P^{(1)}_\alpha,\, H(\bar{\beta}^{(1)}_{Z,*,*,*})-\lambda^{(1)}_Z\right\}-o(n)}.$$

By symmetry, we can apply the same method to the second and third region, where for the second region we perform asymmetric hashing that shares Y-variable blocks, and for the third region we perform asymmetric hashing that shares X-blocks. Taking the tensor product of these results returned by our method on the three regions concludes the proof.

6 Constituent Stage

In the constituent stage for level-ℓ for some $\ell > 1$, the input is an s-term level-ℓ ε-interface tensor with parameters

$$\{(n_t, i_t, j_t, k_t, \beta_{X,t}, \beta_{Y,t}, \beta_{Z,t})\}_{t\in[s]}$$

that meet the following constraints:

1. For every $t \in [s]$, if $\hat{i}_1 + \hat{i}_2 + \cdots + \hat{i}_{2^{\ell-1}} \neq i_t$, then $\beta_{X,t}(\hat{i}_1, \hat{i}_2, \ldots, \hat{i}_{2^{\ell-1}}) = 0$. Similar constraints hold for $\beta_{Y,t}$ and $\beta_{Z,t}$.
2. For every $t \in [s]$ with $j_t = 0$, and every $\hat{i}_1, \hat{i}_2, \ldots, \hat{i}_{2^{\ell-1}}$,
$$\beta_{X,t}(\hat{i}_1, \hat{i}_2, \ldots, \hat{i}_{2^{\ell-1}}) = \beta_{Z,t}(2 - \hat{i}_1, 2 - \hat{i}_2, \ldots, 2 - \hat{i}_{2^{\ell-1}}).$$
Similar relations hold between $\beta_{X,t}$ and $\beta_{Y,t}$ where $k_t = 0$ and between $\beta_{Y,t}$ and $\beta_{Z,t}$ where $i_t = 0$.

Additionally, we let $n = \sum_t n_t$ and $N = 2^{\ell-1} \cdot n$. The goal of this stage is to degenerate the input to the tensor product between a matrix multiplication tensor and multiple independent copies of a level-$(\ell-1)$ ε'-interface tensor for some $\varepsilon' > 0$.

Before we apply the laser method, let us handle the terms $t \in [s]$ in the level-ℓ ε-interface tensor where $i_t = 0$, $j_t = 0$ or $k_t = 0$, which are already matrix multiplication tensors. The proof idea of the following theorem is similar to the proof idea of a result in [33], who showed the version of the following theorem without complete split distributions.

THEOREM 6.1. *If $k_t = 0$, then*

$$T^{\otimes n_t}_{i_t,j_t,k_t}[\beta_{X,t}, \beta_{Y,t}, \beta_{Z,t}, \varepsilon] \equiv \langle 1, M, 1\rangle,$$

where

$$M = 2^{n_t(H(\beta_{X,t})\pm o_{1/\varepsilon}(1))\pm o(n)} \cdot q^{n_t \sum_{(\hat{i}_1,\hat{i}_2,\ldots,\hat{i}_{2^{\ell-1}})} \beta_{X,t}(\hat{i}_1,\hat{i}_2,\ldots,\hat{i}_{2^{\ell-1}}) \sum_{p=1}^{2^{\ell-1}} [\hat{i}_p=1]}.$$

Similar results hold when $i_t = 0$ or $j_t = 0$.

Proof. As $k_t = 0$, there is only one Z-variable z_0 in the given tensor. Also, for each fixed X-variable x, there is a unique Y-variable y so that xyz_0 is a term in the given tensor (this is because it is a subtensor of $\mathrm{CW}_q^{\otimes N}$), and vice versa. Thus, the given tensor is isomorphic to an inner product tensor $\langle 1, M, 1\rangle$ for some $M \geq 0$. It remains to calculate the number of X-variables in the given tensor. The X-variables are distributed among several level-1 X-blocks. Fixing a complete split distribution $\xi_{X,t}$ whose L_∞ distance to $\beta_{X,t}$ is within ε, the number of level-1 X-blocks in $T^{\otimes n_t}_{i_t,j_t,k_t}$ that conform with $\xi_{X,t}$ is

$$2^{n_t H(\xi_{X,t})\pm o(n)} = 2^{n_t(H(\beta_{X,t})\pm o_{1/\varepsilon}(1))\pm o(n)}. \tag{6.11}$$

In each of these level-1 blocks, say $X_{\hat{I}}$, the number of X-variables is

$$\begin{aligned} q^{\sum_{p=1}^{n_t\cdot 2^{\ell-1}} [\hat{I}_p=1]} &= q^{n_t \sum_{(\hat{i}_1,\ldots,\hat{i}_{2^{\ell-1}})} \xi_{X,t}(\hat{i}_1,\ldots,\hat{i}_{2^{\ell-1}}) \sum_{p=1}^{2^{\ell-1}} [\hat{i}_p=1]} \\ &= q^{n_t \sum_{(\hat{i}_1,\ldots,\hat{i}_{2^{\ell-1}})} (\beta_{X,t}(\hat{i}_1,\ldots,\hat{i}_{2^{\ell-1}})\pm o_{1/\varepsilon}(1)) \sum_{p=1}^{2^{\ell-1}} [\hat{i}_p=1]}. \end{aligned} \tag{6.12}$$

The product of (6.11) and (6.12) gives the number of X-variables belonging to level-1 X-blocks that are consistent with a certain $\xi_{X,t}$; taking summation over all $\xi_{X,t}$ (there are poly(n) of which) proves the lemma. ∎

Next, we assume that we already used Theorem 6.1 to handle terms with $i_t = 0$, $j_t = 0$ or $k_t = 0$, and assume without loss of generality that we are left with the first s' terms for some $s' \leq s$.

For a triple of level-ℓ complete split distributions $(\beta_X, \beta_Y, \beta_Z)$ associated with the tensor power of the constituent tensor T_{i_t,j_t,k_t}, we define a distribution γ_X on $\{0, \ldots, 2^{\ell-1}\}^2$ as follows:

$$\gamma_X(l_X, r_X) \coloneqq \sum_{\substack{(\hat{i}_1,\hat{i}_2,\ldots,\hat{i}_{2^{\ell-1}}):\\ \hat{i}_1+\cdots+\hat{i}_{2^{\ell-2}}=l_X,\\ \hat{i}_{2^{\ell-2}+1}+\cdots+\hat{i}_{2^{\ell-1}}=r_X}} \beta_X(\hat{i}_1, \hat{i}_2, \ldots, \hat{i}_{2^{\ell-1}}).$$

It describes how every level-ℓ index i_t splits into two level-$(\ell-1)$ indices. We similarly define γ_Y and γ_Z.

Let α be a distribution on possible combinations of (l_X, l_Y, l_Z) such that the marginals of α are consistent with $\gamma_X(l_X, i-l_X)$, $\gamma_Y(l_Y, j-l_Y)$, $\gamma_Z(l_Z, k-l_Z)$. Moreover, let $\beta_{X,i',j',k'}$, $\beta_{Y,i',j',k'}$, $\beta_{Z,i',j',k'}$ be level-$(\ell-1)$ complete split distributions. We then define the following quantities:

- D is the set of distributions whose marginal distributions on the three dimensions are consistent with $\gamma_X(l_X, i-l_X)$, $\gamma_Y(l_Y, j-l_Y)$, $\gamma_Z(l_Z, k-l_Z)$ respectively, and let the penalty term $P_\alpha \coloneqq \max_{\alpha' \in D} H(\alpha') - H(\alpha) \geq 0$.
- For every k', $\alpha(+,+,k') \coloneqq \sum_{i'>0,j'>0} \alpha(i',j',k')$; for every j', $\alpha(+,j',+) \coloneqq \sum_{i'>0,k'>0} \alpha(i',j',k')$; and for every i', $\alpha(i',+,+) \coloneqq \sum_{j'>0,k'>0} \alpha(i',j',k')$.
- For every k', $\alpha(<,<,k') \coloneqq \sum_{i'<i_t,j'<j_t} \alpha(i',j',k')$; for every j, $\alpha(<,j',<) \coloneqq \sum_{i'<i_t,k'<k_t} \alpha(i',j',k')$; and for every i', $\alpha(i',<,<) \coloneqq \sum_{j'<j_t,k'<k_t} \alpha(i',j',k')$.
- For every k', $\overline{\beta}_{Z,+,+,k'} \coloneqq \frac{1}{\alpha(+,+,k')} \sum_{i'>0,j'>0} \alpha(i',j',k') \cdot \beta_{Z,i',j',k'}$, while $\overline{\beta}_{Y,+,j',+}$ and $\overline{\beta}_{X,i',+,+}$ are defined similarly.
- $\lambda_Z \coloneqq \sum_{i',j',k':i'=0 \text{ or } j'=0} \big(\alpha(i',j',k') + \alpha(i_t-i', j_t-j', k_t-k')\big) \cdot H(\beta_{Z,i',j',k'})$
  $+ \sum_{k'} \big(\alpha(+,+,k') + \alpha(<,<,k_t-k')\big) \cdot H(\overline{\beta}_{Z,+,+,k_t-k'})$, while λ_X and λ_Y are defined similarly.

In the following proposition, we will use the above definitions for different $t \in [s']$ and $r \in [3]$. We will use t in the subscripts and (r) in the superscripts on variables to denote that they are computed using values of $\alpha_t^{(r)}, \beta_{X,t}^{(r)}, \beta_{Y,t}^{(r)}, \beta_{Z,t}^{(r)}, \{\beta_{X,t,i',j',k'}^{(r)}\}_{i',j',k'}, \{\beta_{Y,t,i',j',k'}^{(r)}\}_{i',j',k'}, \{\beta_{Z,t,i',j',k'}^{(r)}\}_{i',j',k'}$.

PROPOSITION 6.1. *An s'-term level-ℓ ε-interface tensor with parameters*

$$\{(n_t, i_t, j_t, k_t, \beta_{X,t}, \beta_{Y,t}, \beta_{Z,t})\}_{t \in [s']}$$

for $\varepsilon > 0$, $i_t, j_t, k_t > 0 \; \forall \, t \in [s']$ can be degenerated into

$$2^{(E_1+E_2+E_3)-o(n)-o_{1/\varepsilon}(n)}$$

independent copies of a level-$(\ell-1)$ interface tensor with parameter list

$$\left\{\left(n_t \cdot A_{t,r} \cdot \big(\alpha_t^{(r)}(i',j',k') + \alpha_t^{(r)}(i_t-i', j_t-j', k_t-k')\big), i', j', k', \beta_{X,t,i',j',k'}^{(r)}, \beta_{Y,t,i',j',k'}^{(r)}, \beta_{Z,t,i',j',k'}^{(r)}\right)\right\}$$

for $t \in [s']$, $r \in [3]$, $i'+j'+k' = 2^{\ell-1}$, $0 \leq i' \leq i_t$, $0 \leq j' \leq j_t$, $0 \leq k' \leq k_t$, where

- *$0 \leq A_{t,1}, A_{t,2}, A_{t,3} \leq 1$ and $A_{t,1} + A_{t,2} + A_{t,3} = 1$ for every $t \in [s']$;*
- *For every t, and for every $W \in \{X,Y,Z\}$, $A_{t,1}\beta_{W,t}^{(1)} + A_{t,2}\beta_{W,t}^{(2)} + A_{t,3}\beta_{W,t}^{(3)} = \beta_{W,t}$ ($\beta_{W,t}^{(r)}$ are intermediate variables that will be used later);*
- *For every $W \in \{X,Y,Z\}$, $r \in [3]$ and $i'+j'+k' = 2^{\ell-1}$, $\beta_{W,t,i',j',k'}^{(r)}$ is a level-$(\ell-1)$ complete split distribution;*
- *For every $W \in \{X,Y,Z\}$, $t \in [s']$ and $r \in [3]$,*

$$\beta_{W,t}^{(r)} = \sum_{i',j',k'} \alpha_t^{(r)}(i',j',k') \cdot \left(\beta_{W,t,i',j',k'}^{(r)} \times \beta_{W,t,i_t-i',j_t-j',k_t-k'}^{(r)}\right);$$

- $E_1 \coloneqq \min\Bigg\{ \sum_{t\in[s']} A_{t,1}\cdot n_t\cdot\Big(H(\gamma^{(1)}_{X,t})-P^{(1)}_{\alpha,t}\Big),\ \sum_{t\in[s']} A_{t,1}\cdot n_t\cdot\Big(H(\gamma^{(1)}_{Y,t})-P^{(1)}_{\alpha,t}\Big),\ \sum_{t\in[s']} A_{t,1}\cdot n_t\cdot\Big(H(\beta^{(1)}_{Z,t})-\lambda^{(1)}_{Z,t}\Big)\Bigg\},$

$E_2 \coloneqq \min\Bigg\{ \sum_{t\in[s']} A_{t,2}\cdot n_t\cdot\Big(H(\gamma^{(2)}_{X,t})-P^{(2)}_{\alpha,t}\Big),\ \sum_{t\in[s']} A_{t,2}\cdot n_t\cdot\Big(H(\gamma^{(2)}_{Z,t})-P^{(2)}_{\alpha,t}\Big),\ \sum_{t\in[s']} A_{t,2}\cdot n_t\cdot\Big(H(\beta^{(2)}_{Y,t})-\lambda^{(2)}_{Y,t}\Big)\Bigg\},$

$E_3 \coloneqq \min\Bigg\{ \sum_{t\in[s']} A_{t,3}\cdot n_t\cdot\Big(H(\gamma^{(3)}_{Y,t})-P^{(3)}_{\alpha,t}\Big),\ \sum_{t\in[s']} A_{t,3}\cdot n_t\cdot\Big(H(\gamma^{(3)}_{Z,t})-P^{(3)}_{\alpha,t}\Big),\ \sum_{t\in[s']} A_{t,3}\cdot n_t\cdot\Big(H(\beta^{(3)}_{X,t})-\lambda^{(3)}_{X,t}\Big)\Bigg\}.$

Given Proposition 6.1, we obtain the following theorem, whose proof is essentially the same as that of Theorem 5.1.

THEOREM 6.2. *$2^{o(n)}$ independent copies of s'-term level-ℓ 3ε-interface tensor with parameters*

$$\{(n_t, i_t, j_t, k_t, \beta_{X,t}, \beta_{Y,t}, \beta_{Z,t})\}_{t\in[s']}$$

for $\varepsilon > 0, i_t, j_t, k_t > 0\ \forall\ t \in [s']$ can be degenerated into

$$2^{(E_1+E_2+E_3)-o(n)-o_{1/\varepsilon}(n)}$$

independent copies of a level-$(\ell-1)$ ε-interface tensor with parameter list

$$\Big\{\Big(n_t\cdot A_{t,r}\cdot\big(\alpha^{(r)}_t(i',j',k')+\alpha^{(r)}_t(i_t-i',j_t-j',k_t-k')\big), i', j', k', \beta^{(r)}_{X,t,i',j',k'}, \beta^{(r)}_{Y,t,i',j',k'}, \beta^{(r)}_{Z,t,i',j',k'}\Big)\Big\}$$

for $t \in [s']$, $r \in [3]$, $i'+j'+k' = 2^{\ell-1}$, $0 \le i' \le i_t$, $0 \le j' \le j_t$, $0 \le k' \le k_t$, where the constraints are the same as those in Proposition 6.1.

Proof. Similar to Theorem 5.1, for every set of complete split distributions $\{\xi^{(r)}_{W,t,i',j',k'}\}_{W,t,r,i',j',k'}$ that is at most ε away in L_∞ distance from $\{\beta^{(r)}_{W,t,i',j',k'}\}_{W,t,r,i',j',k'}$, we take an independent copy of the input interface tensor, and degenerate it to independent copies of the output interface tensor with the specified complete split distributions. Let

$$\xi^{(r)}_{W,t} = \sum_{i',j',k'} \alpha^{(r)}_t(i',j',k')\cdot\Big(\xi^{(r)}_{W,t,i',j',k'}\times\xi^{(r)}_{W,t,i_t-i',j_t-j',k_t-k'}\Big) \quad (\forall\, W\in\{X,Y,Z\},\, r\in[3],\, t\in[s']) \tag{6.13}$$

and $\xi_{W,t} = A_{t,1}\xi^{(1)}_{W,t} + A_{t,2}\xi^{(2)}_{W,t} + A_{t,3}\xi^{(3)}_{W,t}$ be determined by the considered complete split distributions $\{\xi_{W,t,i',j',k'}\}$. According to Proposition 6.1, an ε-interface tensor $\mathcal{T}$ with parameter list $\{(n_t, i_t, j_t, k_t, \xi_{X,t}, \xi_{Y,t}, \xi_{Z,t})\}_{t\in[s']}$ can degenerate to $2^{E_1+E_2+E_3-o(n)-o_{1/\varepsilon}(n)}$ copies of the target interface tensor. Summing up a copy of the outcome tensor for each $\{\xi^{(r)}_{W,t,i',j',k'}\}_{W,t,r,i',j',k'}$ will give the output ε-interface tensor, so we can get $2^{E_1+E_2+E_3-o(n)-o_{1/\varepsilon}(n)}$ independent copies of the output tensor in total.

It remains to show that $\mathcal{T}$ is a subtensor of the input interface tensor, i.e., a 3ε-interface tensor with parameters $\{(n_t, i_t, j_t, k_t, \beta_{X,t}, \beta_{Y,t}, \beta_{Z,t})\}_{t\in[s']}$. On the right-hand side of (6.13), the two complete split distributions have at most ε distance from $\beta^{(r)}_{W,t,i',j',k'}$ and $\beta^{(r)}_{W,t,i_t-i',j_t-j',k_t-k'}$, so their product has $\le 2\varepsilon$ distance[4] from

[4]The distance is at most 2ε for the following reason: first, we change $\beta^{(r)}_{W,t,i',j',k'}$ to $\xi^{(r)}_{W,t,i',j',k'}$, which introduces an additive ε

$\beta^{(r)}_{W,t,i',j',k'} \times \beta^{(r)}_{W,t,i_t-i',j_t-j',k_t-k'}$; since the coefficients $\alpha^{(r)}_t(i',j',k')$ sum up to 1, we know that the left-hand side $\xi^{(r)}_{W,t}$ has at most 2ε distance from $\beta^{(r)}_{W,t}$ as well, and the same holds between $\xi_{W,t}$ and $\beta_{W,t}$. Thus the ε-interface tensor with complete split distributions $\{\xi_{W,t}\}_{W,t}$ is contained in the 3ε-interface tensor with $\{\beta_{W,t}\}_{W,t}$ as a subtensor. Then we conclude the proof. ∎

The remainder of this section aims to prove Proposition 6.1.

6.1 Dividing into regions. For each of the s' terms, say the t-th term, we pick three real numbers $A_{t,1}, A_{t,2}, A_{t,3} \geq 0$ where $A_{t,1} + A_{t,2} + A_{t,3} = 1$, that aims to divide the t-th term in the input level-ℓ ε-interface tensor into three regions of sizes $A_{t,1}n_t, A_{t,2}n_t$ and $A_{t,3}n_t$ respectively. We also pick three different complete split distributions $\beta^{(1)}_{X,t}, \beta^{(2)}_{X,t}, \beta^{(3)}_{X,t}$, with the constraint

$$\beta^{(1)}_{X,t}A_{t,1} + \beta^{(2)}_{X,t}A_{t,2} + \beta^{(3)}_{X,t}A_{t,3} = \beta_{X,t}. \tag{6.14}$$

We also pick $\beta^{(r)}_{Y,t}$ and $\beta^{(r)}_{Z,t}$ for $r \in [3]$ with similar constraints. Similar to Remark 5.1, we assume without loss of generality that, for every t, r and every $L \in \{0,1,2\}^{2^{\ell-1}}$,

$$\beta^{(r)}_{X,t}(L) = \beta^{(r)}_{Z,t}(\vec{2} - L) \text{ if } j_t = 0, \quad \beta^{(r)}_{Z,t}(L) = \beta^{(r)}_{Y,t}(\vec{2} - L) \text{ if } i_t = 0, \quad \beta^{(r)}_{Y,t}(L) = \beta^{(r)}_{X,t}(\vec{2} - L) \text{ if } k_t = 0$$

where $\vec{2}$ denotes the length-$(2^{\ell-1})$ vector whose coordinates are all 2, and

$$\beta^{(r)}_{X,t}(L) = 0 \text{ if } \sum_t L_t \neq i_t, \quad \beta^{(r)}_{Y,t}(L) = 0 \text{ if } \sum_t L_t \neq j_t, \quad \beta^{(r)}_{Z,t}(L) = 0 \text{ if } \sum_t L_t \neq k_t.$$

For any level-1 X-block, if the portion of it in the r-th region of the t-th term is not ε-approximate consistent with $\beta^{(r)}_{X,t}$, we zero it out. We similarly handle level-1 Y-blocks and Z-blocks. It is not hard to see the following.

Claim 6.1. *After the previous zeroing-out, we obtain a tensor that is isomorphic to*

$$\bigotimes_{r=1}^{3} \bigotimes_{t=1}^{s'} T^{\otimes A_{t,r}n_t}_{i_t,j_t,k_t}[\beta^{(r)}_{X,t}, \beta^{(r)}_{Y,t}, \beta^{(r)}_{Z,t}, \varepsilon].$$

Proof. We only need to show that for a fixed t,

$$T^{\otimes n_t}_{i_t,j_t,k_t}[\beta_{X,t}, \beta_{Y,t}, \beta_{Z,t}, \varepsilon] \trianglerighteq \bigotimes_{r=1}^{3} T^{\otimes A_{t,r}n_t}_{i_t,j_t,k_t}[\beta^{(r)}_{X,t}, \beta^{(r)}_{Y,t}, \beta^{(r)}_{Z,t}, \varepsilon] \tag{6.15}$$

by performing the above zeroing-out rule, i.e., zeroing out every level-1 X-block whose portion in the r-th region is not ε-approximate consistent with $\beta^{(r)}_{X,t}$, and doing similarly for Y- and Z-blocks. Suppose some level-1 X-block belongs to the right-hand side and has complete split distributions $\xi^{(1)}_{X,t}, \xi^{(2)}_{X,t}, \xi^{(3)}_{X,t}$ in three regions respectively, each of which is at most ε away from $\beta^{(1)}_{X,t}, \beta^{(2)}_{X,t}, \beta^{(3)}_{X,t}$ in L_∞ distance. Then, its average complete split distribution $\xi_{X,t} \coloneqq A_1\xi^{(1)}_{X,t} + A_2\xi^{(2)}_{X,t} + A_3\xi^{(3)}_{X,t}$ has at most ε distance from $\beta_{X,t}$, which means that the considered level-1 X-block also belong to the left-hand side. It is the same for Y- and Z-blocks, so the right-hand side of (6.15) is a subtensor of the left-hand side, i.e., Eq. (6.15) holds, which further implies the claim. ∎

In the following, we will focus on the first region $r = 1$, in which we will apply asymmetric hashing that allows the sharing of Z-blocks. Let

$$\mathcal{T}^{(1)} \coloneqq \bigotimes_{t=1}^{s'} T^{\otimes A_{t,1}n_t}_{i_t,j_t,k_t}[\beta^{(1)}_{X,t}, \beta^{(1)}_{Y,t}, \beta^{(1)}_{Z,t}, \varepsilon].$$

We will omit the superscript (1) on all variables for conciseness.

error (as the right hand side in Eq. (6.13) is a weighted average of the entries of $\xi^{(r)}_{W,t,i',j',k'}$); then we change $\beta^{(r)}_{W,t,i_t-i',j_t-j',k_t-k'}$ to $\xi^{(r)}_{W,t,i_t-i',j_t-j',k_t-k'}$, which introduces another additive ε error.

6.2 Asymmetric hashing. Next, we apply hashing similarly to the global stage. For every $t \in [s']$, recall that α_t is a distribution on $\{(i', j', k') \in \mathbb{Z}^3_{\geq 0} : i' + j' + k' = 2^{\ell-1}\}$. Additionally, the marginal distributions of $\alpha_t(i', j', k')$ on the three dimensions are the same as $\gamma_{X,t}(i', i_t - i')$, $\gamma_{Y,t}(j', j_t - j')$, $\gamma_{Z,t}(k', k_t - k')$, respectively.

Each level-$(\ell-1)$ index sequence is partitioned into s' parts, where each part corresponds to one term in $\mathcal{T}$. The t-th part is a length-$(2n_t)$ $\{0, \ldots, 2^{\ell-1}\}$-sequence, which can also be viewed as a length-(n_t) $\{0, \ldots, 2^{\ell-1}\}^2$-sequence by combining pairs of adjacent numbers. If the t-th part of a level-$(\ell-1)$ X-index sequence is not consistent with the distribution $\gamma_{X,t}$ for any t, we zero out the corresponding level-$(\ell-1)$ X-block. We similarly handle the Y- and Z-blocks.

Let N_{BX} be the number of remaining level-$(\ell-1)$ X-blocks, and it is not difficult to see that

$$N_{\mathrm{BX}} = 2^{\sum_t H(\gamma_{X,t}) \cdot A_{t,1} n_t \pm o(n)}. \tag{6.16}$$

Similarly, let N_{BY} and N_{BZ} be the number of remaining Y- and Z-blocks, and we have

$$N_{\mathrm{BY}} = 2^{\sum_t H(\gamma_{Y,t}) \cdot A_{t,1} n_t \pm o(n)}, \quad N_{\mathrm{BZ}} = 2^{\sum_t H(\gamma_{Z,t}) \cdot A_{t,1} n_t \pm o(n)}. \tag{6.17}$$

Let N_α be the number of remaining block triples that are consistent with $\{\alpha_t\}_{t \in [s']}$. We have

$$N_\alpha = 2^{\sum_t H(\alpha_t) \cdot A_{t,1} n_t \pm o(n)}. \tag{6.18}$$

Finally, let $N_{\alpha_X, \alpha_Y, \alpha_Z}$ be the number of remaining level-$(\ell-1)$ block triples $X_I Y_J Z_K$.

Claim 6.2. *$N_{\alpha_X, \alpha_Y, \alpha_Z} = 2^{\sum_t (H(\alpha_t) + P_{\alpha,t}) \cdot A_{t,1} n_t \pm o(n)}$, where we recall that $P_{\alpha,t} \coloneqq \max_{\alpha'_t \in D_t} H(\alpha'_t) - H(\alpha_t)$ in which D_t is the set of distributions sharing the same marginals as α_t.*

Proof. Fixing a series of distributions $\alpha'_t \in D_t$ $(t = 1, 2, \ldots, s')$, the number of level-$(\ell-1)$ block triples consistent with $\{\alpha'_t\}_{t \in [s']}$ equals

$$2^{\sum_t H(\alpha'_t) \cdot A_{t,1} n_t \pm o(n)} \leq 2^{\sum_t \max_{\alpha''_t \in D_t} H(\alpha''_t) \cdot A_{t,1} n_t \pm o(n)} \leq 2^{\sum_t (H(\alpha_t) + P_{\alpha,t}) \cdot A_{t,1} n_t \pm o(n)}.$$

Taking summation over all $\mathrm{poly}(n) = 2^{o(n)}$ series of distributions $\{\alpha'_t\}_{t \in [s']}$ will prove the claim. ∎

Let $M \in [M_0, 2M_0]$ be a prime number for some integer M_0. Similar as before, the value of M_0 is yet to be fixed, but we first require that

$$M_0 \geq 8 \cdot \max\left\{ \frac{N_{\alpha_X, \alpha_Y, \alpha_Z}}{N_{\mathrm{BX}}}, \frac{N_{\alpha_X, \alpha_Y, \alpha_Z}}{N_{\mathrm{BY}}} \right\}. \tag{6.19}$$

We independently pick uniformly random elements $b_0, \{w_p\}_{p=0}^{2n} \in \{0, \ldots, M-1\}$, and define the following hash functions $h_X, h_Y, h_Z : \{0, \ldots, 2^{\ell-1}\}^n \to \{0, \ldots, M-1\}$:

$$h_X(I) = b_0 + \left(\sum_{p=1}^{2n} w_p \cdot I_p \right) \bmod M,$$

$$h_Y(J) = b_0 + \left(w_0 + \sum_{p=1}^{2n} w_p \cdot J_p \right) \bmod M,$$

$$h_Z(K) = b_0 + \frac{1}{2} \left(w_0 + \sum_{p=1}^{2n} w_p \cdot (2^{\ell-1} - K_p) \right) \bmod M.$$

Next, for a Salem-Spencer subset B of $\{0, \ldots, M-1\}$ that has size $M^{1-o(1)}$, we zero out all level-$(\ell-1)$ blocks X_I with $h_X(I) \notin B$, Y_J with $h_Y(J) \notin B$, and Z_K with $h_Z(K) \notin B$. Then all remaining block triples are contained in a bucket b for some $b \in B$.

For every bucket b, if it contains two level-$(\ell-1)$ triples $X_I Y_J Z_K$ and $X_I Y_{J'} Z_{K'}$ that share the same X-block, then we zero out X_I. We similarly handle Y-blocks. We repeatedly perform the previous zeroing-outs so that all

remaining triples do not share X- or Y-blocks. For every level-$(\ell-1)$ block X_I (or Y_J), we check whether the unique triple containing it is consistent with $\{\alpha_t\}_{t\in[s']}$; if not, we zero out X_I (or Y_J). We call the tensor after this step $\mathcal{T}_{\text{hash}}$.

The following claims, which are analogous to the claims in Section 5, still hold, and we omit their proofs to conciseness.

CLAIM 6.3 (Implicit in [13], see also [17]). *For a level-$(\ell-1)$ block triple $X_IY_JZ_K \in \mathcal{T}$, and for every $b \in \{0,\dots,M-1\}$,*

$$\Pr\Big[h_X(I) = h_Y(J) = h_Z(K) = b\Big] = \frac{1}{M^2}.$$

Furthermore, for two different block triples $X_IY_JZ_K, X_IY_{J'}Z_{K'} \in \mathcal{T}$ that share the same X-block, and for every $b \in \{0,\dots,M-1\}$,

$$\Pr\Big[h_X(I) = h_Y(J') = h_Z(K') = b \;\Big|\; h_X(I) = h_Y(J) = h_Z(K) = b\Big] = \frac{1}{M}.$$

This also holds analogously for different block triples that share the same Y-block or Z-block.

CLAIM 6.4. *For every $b \in B$ and for every level-$(\ell-1)$ block triple $X_IY_JZ_K \in \mathcal{T}$ that is consistent with $\{\alpha_t\}_{t\in[s']}$, the probability that $X_IY_JZ_K$ remains in $\mathcal{T}_{\text{hash}}$ conditioned on $h_X(I) = h_Y(J) = h_Z(K) = b$ is $\geq \frac{3}{4}$.*

CLAIM 6.5. *The expected number of level-$(\ell-1)$ block triples in $\mathcal{T}_{\text{hash}}$ is at least $N_\alpha \cdot M_0^{-1-o(1)}$.*

6.3 Compatibility zero-out I. Recall that for every $W \in \{X,Y,Z\}$ and $i'+j'+k' = 2^{\ell-1}$, $\beta_{W,t,i',j',k'}$ is a level-$(\ell-1)$ complete split distribution, and they satisfy

$$\beta_{W,t} = \sum_{i',j',k'} \alpha_t(i',j',k') \cdot (\beta_{W,t,i',j',k'} \times \beta_{W,t,i_t-i',j_t-j',k_t-k'}). \tag{6.20}$$

Let

$$S^{(I,J,K)}_{t,i',j',k'} := \{p \text{ is in the } t\text{-th term} \mid I_p = i', J_p = j', K_p = k'\},$$

and

$$S^{(K)}_{t,*,*,k'} := \{p \text{ is in the } t\text{-th term} \mid K_p = k'\}, \quad S_{t,*,*,*} := \{p \text{ is in the } t\text{-th term}\}.$$

If clear from the context, we will drop the superscript (I,J,K) or (K).

Recall that in $\mathcal{T}_{\text{hash}}$, every level-$(\ell-1)$ block X_I is in a unique block triple $X_IY_JZ_K$. For every level-1 block $X_{\hat{I}} \in X_I$, we will zero out $X_{\hat{I}}$ if $\mathsf{split}(\hat{I}, S_{t,i',j',k'}) \neq \beta_{X,t,i',j',k'}$ for any t,i',j',k'. Similarly, every level-ℓ block Y_J is in a unique block triple, and we zero out every $Y_{\hat{J}} \in Y_J$ where $\mathsf{split}(\hat{J}, S_{t,i',j',k'}) \neq \beta_{Y,t,i',j',k'}$ for any t,i',j',k'.

For every level-1 block $Z_{\hat{K}} \in Z_K$, we zero out $Z_{\hat{K}}$ if $\mathsf{split}(\hat{K}, S_{t,*,*,k'}) \neq \overline{\beta}_{Z,t,*,*,k'}$ for any t,k', where

$$\overline{\beta}_{Z,t,*,*,k'} := \frac{\sum_{i'+j'=2^{\ell-1}-k'} \big(\alpha(i',j',k') + \alpha(i_t-i', j_t-j', k_t-k')\big) \cdot \beta_{Z,i',j',k'}}{\sum_{i'+j'=2^{\ell-1}-k'} \big(\alpha(i',j',k') + \alpha(i_t-i', j_t-j', k_t-k')\big)}.$$

We call the tensor after the previous zeroing-outs $\mathcal{T}_{\text{comp}}$.

Next, we define the notion of compatibility.

DEFINITION 6.1 (Compatibility). *For some I,J,K, a level-1 block $Z_{\hat{K}} \in Z_K$ is compatible with a level-$(\ell-1)$ triple $X_IY_JZ_K$ if*

1. *For every t and every $(i',j',k') \in \mathbb{Z}^3_{\geq 0} \cap [0,i_t] \times [0,j_t] \times [0,k_t]$ with $i'+j'+k' = 2^{\ell-1}$, $i' = 0$ or $j' = 0$, there is $\mathsf{split}(\hat{K}, S_{t,i',j',k'}) = \beta_{Z,t,i',j',k'}$.*
2. *For every t and every index $k' \in \{0,1,\dots,\min\{2^{\ell-1}, k_t\}\}$, $\mathsf{split}(\hat{K}, S_{t,*,*,k'}) = \overline{\beta}_{Z,t,*,*,k'}$.*

CLAIM 6.6. *In $\mathcal{T}_{\text{comp}}$, for every remaining level-1 block triple $X_{\hat{I}}Y_{\hat{J}}Z_{\hat{K}}$ and the level-$(\ell-1)$ block triple $X_IY_JZ_K$ that contains it, $Z_{\hat{K}}$ is compatible with $X_IY_JZ_K$.*

The proof of this claim is the same as Claim 5.5.

6.4 Compatibility zero-out II: unique triple. In this step, we zero out all level-1 Z-block $Z_{\hat{K}}$ that are compatible with more than one level-$(\ell-1)$ triples and they become holes. After this step, each remaining level-1 Z-block $Z_{\hat{K}} \in Z_K$ is compatible with a unique level-$(\ell-1)$ triple $X_I Y_J Z_K$ containing it.

6.5 Usefulness zero-out. Next, we further zero out some level-1 Z-blocks using the following definition of *usefulness*.

DEFINITION 6.2 (Usefulness). *For a level-1 block $Z_{\hat{K}}$ and a level-$(\ell-1)$ triple $X_I Y_J Z_K$ containing it, if for all t, i', j', k' we have $\mathsf{split}(\hat{K}, S_{t,i',j',k'}) = \beta_{Z,t,i',j',k'}$, then we say that $Z_{\hat{K}}$ is* useful *for $X_I Y_J Z_K$.*

For each $Z_{\hat{K}}$, it appears in a unique triple $X_I Y_J Z_K$ by the previous zeroing out. Furthermore, if $Z_{\hat{K}}$ is not useful for this triple, we zero out $Z_{\hat{K}}$. We call the current tensor $\mathcal{T}_{\text{useful}}$.

Ideally, we want the subtensor of $\mathcal{T}_{\text{useful}}$ over each triple $X_I Y_J Z_K$ to be isomorphic to

$$\mathcal{T}^* = \bigotimes_{t\in[s']} \bigotimes_{i'+j'+k'=2^{\ell-1}} T_{i',j',k'}^{\otimes A_{t,1}\cdot(\alpha_t(i',j',k')+\alpha_t(i_t-i',j_t-j',k_t-k'))\cdot n_t}[\beta_{X,t,i',j',k'}, \beta_{Y,t,i',j',k'}, \beta_{Z,t,i',j',k'}].$$

However, there will be two types of holes. The first type of holes is caused by the fact that some level-1 subtensors are already missing in the input tensor because we enforced complete split distributions $\beta_{X,t}, \beta_{Y,t}, \beta_{Z,t}$ on it; the second type of holes is caused by zeroing out $Z_{\hat{K}}$ that are compatible with multiple level-$(\ell-1)$ triples. In the next section, we will analyze and fix these two types of holes.

6.6 Fixing holes. First, we analyze the fraction of holes that are caused by the complete split distributions enforced in the input. To do so, we focus on a fixed triple $X_I Y_J Z_K$ and the subtensor $\mathcal{T}^*$ we desire. Then we take a random level-1 block that is not zeroed out in $\mathcal{T}^*$, and upper bound the probability that this level-1 block is zeroed out in the input level-ℓ ε-interface tensor. By symmetry, it suffices to focus on X-blocks.

Fix any $(\hat{i}_1, \dots, \hat{i}_{2^{\ell-1}})$, let us analyze the fraction of its occurrences in a random level-1 X-block in $\mathcal{T}^*$. For every $t \in [s']$, and for every i', j', k', we first focus on the level-ℓ positions in the t-th term where (i_t, j_t, k_t) is split into (i', j', k') and $(i_t - i', j_t - j', k_t - k')$ (thus, there are $A_{t,1} \cdot \alpha_t(i', j', k') \cdot n_t$ such positions). Among these positions, we want to analyze the number of positions that correspond to the level-1 chunk $(\hat{i}_1, \dots, \hat{i}_{2^{\ell-1}})$. Therefore, the first half-chunk, which corresponds to (i', j', k'), should be $(\hat{i}_1, \dots, \hat{i}_{2^{\ell-2}})$, and the second half-chunk, which corresponds to $(i_t - i', j_t - j', k_t - k')$, should be $(\hat{i}_{2^{\ell-2}+1}, \dots, \hat{i}_{2^{\ell-1}})$.

There are $A_{t,1} \cdot (\alpha_t(i', j', k') + \alpha_t(i_t - i', j_t - j', k_t - k')) \cdot n_t$ level-$(\ell-1)$ positions corresponding to (i', j', k'), and among them, $A_{t,1} \cdot \alpha_t(i', j', k') \cdot n_t$ are in odd positions. By definition of $\mathcal{T}^*$, the fraction of $(\hat{i}_1, \dots, \hat{i}_{2^{\ell-2}})$ in these $A_{t,1} \cdot (\alpha_t(i', j', k') + \alpha_t(i_t - i', j_t - j', k_t - k')) \cdot n_t$ positions is $\beta_{X,t,i',j',k'}(\hat{i}_1, \dots, \hat{i}_{2^{\ell-2}})$, and if we take a random level-1 X-block in $\mathcal{T}^*$, the fraction of $(\hat{i}_1, \dots, \hat{i}_{2^{\ell-2}})$ among the odd positions corresponding to (i', j', k') is $\beta_{X,t,i',j',k'}(\hat{i}_1, \dots, \hat{i}_{2^{\ell-2}}) \pm o(1)$ with $1 - 1/\operatorname{poly}(n)$ probability, by concentration bounds. Furthermore, the subset of positions in these $A_{t,1} \cdot \alpha_t(i', j', k') \cdot n_t$ positions is also random. Similarly, with $1 - 1/\operatorname{poly}(n)$ probability, the fraction of $(\hat{i}_{2^{\ell-2}+1}, \dots, \hat{i}_{2^{\ell-1}})$ in the even positions corresponding to $(i_t - i', j_t - j', k_t - k')$ is $\beta_{X,t,i',j',k'}(\hat{i}_{2^{\ell-2}+1}, \dots, \hat{i}_{2^{\ell-1}}) \pm o(1)$, and the positions are also random. Applying concentration bounds again, we get that the fraction of level-ℓ positions corresponding to $(\hat{i}_1, \dots, \hat{i}_{2^{\ell-1}})$ among positions that split into (i', j', k') and $(i_t - i', j_t - j', k_t - k')$ is

$$\beta_{X,t,i',j',k'}(\hat{i}_1, \dots, \hat{i}_{2^{\ell-2}}) \cdot \beta_{X,t,i_t-i',j_t-j',k_t-k'}(\hat{i}_{2^{\ell-2}+1}, \dots, \hat{i}_{2^{\ell-1}}) \pm o(1).$$

Summing over all i', j', k', we get that with probability $1 - 1/\operatorname{poly}(n)$, the fraction of level-$\ell$ positions with $(\hat{i}_1, \dots, \hat{i}_{2^{\ell-1}})$ is

$$\sum_{i',j',k'} \alpha_t(i', j', k') \cdot \beta_{X,t,i',j',k'}(\hat{i}_1, \dots, \hat{i}_{2^{\ell-2}}) \cdot \beta_{X,t,i_t-i',j_t-j',k_t-k'}(\hat{i}_{2^{\ell-2}+1}, \dots, \hat{i}_{2^{\ell-1}}) \pm o(1)$$

(by Eq. (6.20)) $\quad = \beta_{X,t}(\hat{i}_1, \dots, \hat{i}_{2^{\ell-1}}) \pm o(1).$

The $o(1)$ term can become less than ε, and the $1-1/\operatorname{poly}(n)$ probability can be bounded by $1-1/n^2$ for sufficiently large n. Therefore, a random level-1 X-block appears in $\mathcal{T}$ with probability at least $1-1/n^2$. This means that the

fraction of holes caused by the complete split distributions enforced in the input is $1-1/n^2$ for the X-dimension. By symmetry, the same also holds for the Y- and Z-dimensions.

Next, we focus on holes caused by zeroing out $Z_{\hat{K}}$ that are compatible with multiple level-$(\ell-1)$ triples. The analysis will be similar to Section 5.6.

First, notice that for every level-1 Z-block $Z_{\hat{K}}$ that appears in the input of the constituent stage, its complete split distribution $\xi_{Z,t}$ in the t-th term must be within ε L_∞-distance to the given parameter $\beta_{Z,t}$. Then we define p_{comp} as follows:

DEFINITION 6.3 (p_{comp}). *For fixed $Z_{\hat{K}}$ and Z_K where $Z_{\hat{K}} \in Z_K$ and $\hat{K}$ has level-ℓ complete split distributions $\{\xi_{Z,t}\}_{t\in[s']}$, we define $p^*_{\mathrm{comp}}(\{\xi_{Z,t}\}_{t\in[s']})$ as the probability that a uniformly random block triple $X_IY_JZ_K$ consistent with $\{\alpha_t\}_{t\in[s']}$ is compatible with $Z_{\hat{K}}$. We further define $p_{\mathrm{comp}} \coloneqq \max_{\substack{\{\xi_{Z,t}\}_{t\in[s']}:\\ \|\xi_{Z,t}-\beta_{Z,t}\|_\infty\le\varepsilon\ \forall t}} p^*_{\mathrm{comp}}(\{\xi_{Z,t}\}_{t\in[s']})$.*

By symmetry between level-ℓ positions, this probability $p^*_{\mathrm{comp}}(\{\xi_{Z,t}\}_{t\in[s']})$ is the same for different $\hat{K}$ that have the same complete split distributions, so p^*_{comp} and p_{comp} is well-defined.

CLAIM 6.7. *The value of $p^*_{\mathrm{comp}}(\{\xi_{Z,t}\}_{t\in[s']})$ is at most*

$$2^{\sum_{t\in[s']}(\lambda_{Z,t}-H(\xi_{Z,t})+H(\gamma_{Z,t}))A_{t,1}\cdot n_t\pm o(n)},$$

where we recall that

$$\lambda_{Z,t} = \sum_{i',j',k' : i'=0 \text{ or } j'=0} \big(\alpha_t(i',j',k') + \alpha_t(i_t-i', j_t-j', k_t-k')\big)\cdot H(\beta_{Z,t,i',j',k'}) + \sum_{k'}\big(\alpha_t(+,+,k') + \alpha_t(<,<,k_t-k')\big)\cdot H(\overline{\beta}_{Z,+,+,k_t-k'}),$$

and

$$\alpha_t(+,+,k') = \sum_{i'>0,\, j'>0} \alpha_t(i',j',k'), \quad \alpha_t(<,<,k') = \sum_{i'<i_t,\, j'<j_t} \alpha_t(i',j',k').$$

Furthermore,

$$p_{\mathrm{comp}} \le 2^{\sum_{t\in[s']}\left(\lambda_{Z,t}-H(\beta_{Z,t})+H(\gamma_{Z,t})+o_{1/\varepsilon}(1)\right)A_{t,1}\cdot n_t+o(n)}. \tag{6.21}$$

Proof. Similar to before, it suffices to compute the following two quantities, and $p^*_{\mathrm{comp}}(\{\xi_{Z,t}\}_t)$ will be the ratio between them:

(1) the number of tuples $(I,J,K,\hat{K})$ where $X_IY_JZ_K$ is consistent with $\{\alpha_t\}_{t\in[s']}$, $\hat{K}\in K$, $\hat{K}$ has complete split distributions $\{\xi_{Z,t}\}_t$, and $Z_{\hat{K}}$ is compatible with $X_IY_JZ_K$;

(2) the number of $(I,J,K,\hat{K})$ where $X_IY_JZ_K$ is consistent with $\{\alpha_t\}_{t\in[s']}$, $\hat{K}\in K$, and $\hat{K}$ has complete split distributions $\{\xi_{Z,t}\}_t$.

We first compute the second quantity. First, the number of $Z_{\hat{K}}$ with the desired complete split distributions $\{\xi_{Z,t}\}_t$ is $2^{\sum_t H(\xi_{Z,t})\cdot A_{t,1}\cdot n_t\pm o(n)}$. Each of these $Z_{\hat{K}}$ uniquely determines a level-$(\ell-1)$ block Z_K. Also, for each Z_K, the number of block triples $X_IY_JZ_K$ consistent with $\{\alpha_t\}_{t\in[s']}$ is $\frac{N_\alpha}{N_{\mathrm{BZ}}} = 2^{\sum_t(H(\alpha_t)-H(\gamma_{Z,t}))\cdot A_{t,1}\cdot n_t\pm o(n)}$. Therefore, the second quantity is

$$2^{\sum_t(H(\xi_{Z,t})+H(\alpha_t)-H(\gamma_{Z,t}))\cdot A_{t,1}\cdot n_t\pm o(n)}. \tag{6.22}$$

Then, we compute the first quantity, which does not exceed the number of $(I,J,K,\hat{K})$ where $X_IY_JZ_K$ is consistent with $\{\alpha_t\}_{t\in[s']}$, $\hat{K}\in K$, and $Z_{\hat{K}}$ is compatible with $X_IY_JZ_K$. (We dropped the condition of having correct level-ℓ complete split distributions $\{\xi_{Z,t}\}_t$ and got an overestimation.)

First, the number of block triples $X_IY_JZ_K$ consistent with $\{\alpha_t\}_{t\in[s']}$ is N_α. Then, for each such block triple, we count the number of $Z_{\hat{K}}\in Z_K$ that is compatible with it. If we fix some $X_IY_JZ_K$, then we also have fixed the values of $S_{t,i,j,k}$ for all t,i,j,k. Then it is not difficult to see that the following condition is equivalent to the condition for $Z_{\hat{K}}$ being compatible with $X_IY_JZ_K$:

DEFINITION 6.4 (Compatibility'). *For level-$(\ell-1)$ triple $X_I Y_J Z_K$ consistent with $\{\alpha_t\}_{t\in[s']}$, a level-1 block $Z_{\hat{K}} \in Z_K$ is compatible with $X_I Y_J Z_K$ if*

- *For every t and every $(i',j',k') \in \mathbb{Z}^3_{\geq 0} \cap [0,i_t]\times[0,j_t]\times[0,k_t]$ with $i'+j'+k' = 2^{\ell-1}$, $i'=0$ or $j'=0$, there is $\mathsf{split}(\hat{K}, S_{t,i',j',k'}) = \beta_{Z,t,i',j',k'}$. (This is exactly Item 1 in Definition 6.1).*
- *For every t,k, let $S_{t,+,+,k} := \bigcup_{i>0,j>0} S_{t,i,j,k}$. Then $\mathsf{split}(\hat{K}, S_{t,+,+,k}) = \overline{\beta}_{Z,t,+,+,k}$.*

In this definition, there are constraints on the complete split distributions of $\hat{K}$ on some disjoint subsets of level-$(\ell-1)$ positions, i.e., subsets of $\left[2\sum_t A_{t,1} n_t\right]$. Therefore, we can count the number of valid subsequences of $\hat{K}$ for each of these subsets of indices, and multiply them together to get the number of valid $\hat{K}$. For every t and every $(i',j',k') \in \mathbb{Z}^3_{\geq 0} \cap [0,i_t]\times[0,j_t]\times[0,k_t]$ where $i'+j'+k' = 2^{\ell-1}$ with $i'=0$ or $j'=0$, we require that $\mathsf{split}(\hat{K}, S_{t,i',j',k'}) = \beta_{Z,t,i',j',k'}$, so the number of possibilities of $\hat{K}$ on the subset of indices $S_{t,i',j',k'}$ is

$$2^{H(\beta_{Z,t,i',j',k'})\cdot|S_{t,i',j',k'}|\pm o(n)} = 2^{H(\beta_{Z,t,i',j',k'})\cdot(\alpha_t(i',j',k')+\alpha_t(i_t-i',j_t-j',k_t-k'))\cdot A_{t,1}n_t \pm o(n)}.$$

For every t,k, we require that $\mathsf{split}(\hat{K}, S_{t,+,+,k'}) = \overline{\beta}_{Z,t,+,+,k'}$, so the number of possibilities of $\hat{K}$ on $S_{t,+,+,k'}$ is

$$2^{H(\overline{\beta}_{Z,t,+,+,k'})\cdot|S_{t,+,+,k'}|\pm o(n)} = 2^{H(\overline{\beta}_{Z,t,+,+,k'})\cdot(\alpha_t(+,+,k')+\alpha_t(<,<,k_t-k'))\cdot A_{t,1}n_t\pm o(n)}.$$

Overall, the number of possible compatible $\hat{K}$, multiplied by the number of block triples $X_I Y_J Z_K$, is

$$\begin{aligned} &N_\alpha \cdot \prod_{\substack{t,i',j',k' \\ i'=0 \text{ or } j'=0}} 2^{H(\overline{\beta}_{Z,t,i',j',k'})\cdot(\alpha_t(i',j',k')+\alpha_t(i_t-i',j_t-j',k_t-k'))\cdot A_{t,1}n_t\pm o(n)} \\ &\qquad \cdot \prod_{t,k'} 2^{H(\overline{\beta}_{Z,t,+,+,k'})\cdot(\alpha_t(+,+,k')+\alpha_t(<,<,k_t-k'))\cdot A_{t,1}n_t\pm o(n)} \\ &= 2^{\sum_t (H(\alpha_t)+\lambda_{Z,t})\cdot A_{t,1}n_t \pm o(n)}. \end{aligned} \tag{6.23}$$

Finally, as mentioned, $p^*_{\text{comp}}(\{\xi_{Z,t}\}_t)$ is the ratio between (6.23) and (6.22), so

$$p^*_{\text{comp}}(\{\xi_{Z,t}\}_{t\in[s']}) \leq 2^{\sum_t(\lambda_{Z,t}-H(\xi_{Z,t})+H(\gamma_{Z,t}))A_{t,r}\cdot n_t+o(n)}$$

as desired. The bound (6.21) on p_{comp} follows as the L_∞ distance between $\{\xi_{Z,t}\}_t$ and $\{\beta_{Z,t}\}_t$ is at most ε. ∎

The proof of the following claim is essentially the same as that of Claim 5.8.

CLAIM 6.8. *For every $b\in B$, every level-$(\ell-1)$ block triple $X_I Y_J Z_K$ consistent with $\{\alpha_t\}_{t\in[s']}$, and for each typical $Z_{\hat{K}} \in Z_K$, the probability that $Z_{\hat{K}}$ is compatible with multiple triples in $\mathcal{T}_{\text{comp}}$ is at most*

$$\frac{N_\alpha \cdot p_{\text{comp}}}{N_{\text{BZ}}\cdot M_0},$$

conditioned on $h_X(I) = h_Y(J) = h_Z(K) = b$.

Recall that we require M_0 to be at least $8\cdot\max\left\{\frac{N_{\alpha_X,\alpha_Y,\alpha_Z}}{N_{\text{BX}}}, \frac{N_{\alpha_X,\alpha_Y,\alpha_Z}}{N_{\text{BY}}}\right\}$. Now, we add another (and final) constraint: $M_0 \geq \frac{N_\alpha\cdot p_{\text{comp}}}{N_{\text{BZ}}}\cdot n^2$. That is, we will set M_0 to be

$$\begin{aligned} &\max\left\{\frac{8N_{\alpha_X,\alpha_Y,\alpha_Z}}{N_{\text{BX}}}, \frac{8N_{\alpha_X,\alpha_Y,\alpha_Z}}{N_{\text{BY}}}, \frac{N_\alpha\cdot p_{\text{comp}}}{N_{\text{BZ}}}\cdot n^2\right\} \\ &\leq 2^{\max\{\sum_t(H(\alpha_t)-P_{\alpha,t}-H(\gamma_{X,t}))A_{t,1}\cdot n_t,\ \sum_t(H(\alpha_t)-P_{\alpha,t}-H(\gamma_{Y,t}))A_{t,1}\cdot n_t,\ \sum_t(H(\alpha_t)+\lambda_{Z,t}-H(\beta_{Z,t}))A_{t,1}\cdot n_t\}+o(n)}. \end{aligned}$$

Similar to before, for every $b\in B$ and every level-$(\ell-1)$ block triple $X_I Y_J Z_K$ that is consistent with $\{\alpha_t\}_{t\in[s']}$ and $h_X(I) = h_Y(J) = h_Z(K) = b$, with constant probability, it remains in $\mathcal{T}_{\text{hash}}$ and the fraction of holes caused by enforcing that each $Z_{\hat{K}}$ is compatible with a unique triple is $1/n^2$. Additionally, as discussed earlier, the fraction of holes caused by the input complete split distribution constraints are also $1/n^2$. Overall, we expect to get $N_\alpha \cdot M^{-1-o(1)}$ copies of $\mathcal{T}^*$ whose fraction of holes is $O(1/n^2)$.

By Corollary 4.1, we can degenerate them into $N_\alpha\cdot M^{-1-o(1)}$ unbroken copies of $\mathcal{T}^*$ because $O(1/n^2) \leq \frac{1}{8N}$ for sufficiently large n.

6.7 Summary. In the analysis, we have degenerated $\bigotimes_{t=1}^{s'} T_{i_t,j_t,k_t}^{\otimes A_{t,1} n_t}\left[\beta_{X,t}^{(1)}, \beta_{Y,t}^{(1)}, \beta_{Z,t}^{(1)}, \varepsilon\right]$ into $\geq N_\alpha \cdot M_0^{-1-o(1)}$ copies of a level-$(\ell-1)$ interface tensor $\mathcal{T}^*$ with parameter list

$$\Big\{\Big(A_{t,1}\cdot n_t\cdot\big(\alpha_t^{(1)}(i',j',k')+\alpha_t^{(1)}(i_t-i',j_t-j',k_t-k')\big),\\ i',\ j',\ k',\ \beta_{X,t,i',j',k'}^{(1)},\ \beta_{Y,t,i',j',k'}^{(1)},\ \beta_{Z,t,i',j',k'}^{(1)}\Big)\Big\}_{t\in[s'],i'+j'+k'=2^{\ell-1}}.$$

By plugging in the bounds of N_α and M_0, we see that the number of copies we obtained (in the first region) is

$$2^{\min\left\{\sum_{t\in[s']} A_{t,1}\cdot n_t\cdot\left(H(\gamma_{X,t}^{(1)})-P_{\alpha,t}^{(1)}\right),\ \sum_{t\in[s']} A_{t,1}\cdot n_t\cdot\left(H(\gamma_{Y,t}^{(1)})-P_{\alpha,t}^{(1)}\right),\ \sum_{t\in[s']} A_{t,1}\cdot n_t\cdot\left(H(\beta_{Z,t}^{(1)})-\lambda_{Z,t}^{(1)}\right)\right\}-o_{1/\varepsilon}(n)-o(n)}.$$

We conclude the proof by applying the same method to the second and third region, where for the second region we perform asymmetric hashing that shares Y-variable blocks, and for the third region we perform asymmetric hashing that shares X-blocks, and taking the tensor product of these returned results.

7 Fixing Holes

In this section, we show (by generalizing a result by Duan [16]) that we can degenerate a direct sum of some broken copies of an interface tensor into an unbroken copy of the same tensor as long as we only have a small fraction of holes in the X-, Y-, Z-dimensions. Since our result of fixing holes in all X-, Y-, Z-variables might be of independent interest, we present our result in a more general setting.

Let us first describe the setup of this section. We consider a partitioned tensor T on variable sets $X=\{x_1,\dots,x_{N_X}\}$, $Y=\{y_1,\dots,y_{N_Y}\}$, $Z=\{z_1,\dots,z_{N_Z}\}$ of size $|X|=N_X$, $|Y|=N_Y$, $|Z|=N_Z$ with partitions $X=\bigsqcup_{i=1}^{M_X} X_i$, $Y=\bigsqcup_{j=1}^{M_Y} Y_k$, $Z=\bigsqcup_{k=1}^{M_Z} Z_k$ into equal-size parts $|X_i|=m_X$ for all $i\in[M_X]$, $|Y_j|=m_Y$ for all $j\in[M_Y]$, and $|Z_k|=m_Z$ for all $k\in[M_K]$. (We use the notation X_i to represent both the part itself and the set of elements in this part.) Let $\mathcal{P}_X=\{X_i \mid i\in[M_X]\}$ denote the set of parts in the partition of X, and similarly let $\mathcal{P}_Y,\mathcal{P}_Z$ denote the set of parts in the partition of Y and Z respectively. Note that by definition $N_X=M_X\cdot m_X$, $N_Y=M_Y\cdot m_Y$, $N_Z=M_Z\cdot m_Z$ and $|\mathcal{P}_X|=M_X$, $|\mathcal{P}_Y|=M_Y$, $|\mathcal{P}_Z|=M_Z$.

We consider the broken copies of T where some of the X-, Y- and Z-parts are missing which we call the *holes*. (Equivalently, the variables in a part are either all present or all missing.) More specifically, we say that T_{hole} is *a broken copy of T with holes $P_X^{(0)}\subseteq M_X$, $P_Y^{(0)}\subseteq M_Y$, $P_Z^{(0)}\subseteq M_Z$* when

$$T_{\text{hole}} \;=\; T\big|_{X\setminus\bigsqcup_{X_t\in P_X^{(0)}} X_t,\ Y\setminus\bigsqcup_{Y_t\in P_Y^{(0)}} Y_t,\ Z\setminus\bigsqcup_{Z_t\in P_Z^{(0)}} Z_t} \tag{7.24}$$

is obtained from T via zeroing out the variables in the parts $P_X^{(0)}\subseteq\mathcal{P}_X$, $P_Y^{(0)}\subseteq\mathcal{P}_Y$, $P_Z^{(0)}\subseteq\mathcal{P}_Z$. For simplicity, we define the notation

$$T\|_{P_X,\,P_Y,\,P_Z} \;\coloneqq\; T\big|_{\bigsqcup_{X_t\in P_X} X_t,\ \bigsqcup_{Y_t\in P_Y} Y_t,\ \bigsqcup_{Z_t\in P_Z} Z_t}$$

to represent the subtensor of T over the set of parts P_X,P_Y,P_Z. With this notation, Eq. (7.24) can be rewritten as

$$T_{\text{hole}} \;=\; T\big\|_{\mathcal{P}_X\setminus P_X^{(0)},\ \mathcal{P}_Y\setminus P_Y^{(0)},\ \mathcal{P}_Z\setminus P_Z^{(0)}}.$$

We call the ratios $\left|P_X^{(0)}\right|/M_X$, $\left|P_Y^{(0)}\right|/M_Y$, $\left|P_Z^{(0)}\right|/M_Z$ the *fraction of holes* in the X-, Y-, Z-dimension respectively.

We will show that we can degenerate sub-polynomially many broken copies of T with small fraction of holes in all three dimensions into an unbroken copy of T if T satisfies the following property.

PROPERTY 7.1. *There exists a subset $\mathcal{G}\subseteq\mathcal{S}_{N_X}\times\mathcal{S}_{N_Y}\times\mathcal{S}_{N_Z}$ of permutations over the variables of T where $\mathcal{S}_N$ denotes the symmetric group on $[N]$, such that $\mathcal{G}$ satisfies the following:*

1. *Every $(\pi_X,\pi_Y,\pi_Z)\in\mathcal{G}$* preserves the partitions. *Specifically, it permutes any part into some entire part, i.e., for every part $X_t\in\mathcal{P}_X$, there exists $X_{t'}\in\mathcal{P}_X$ such that $\pi_X(X_t)\coloneqq\{\pi_X(x)\mid x\in X_t\}=X_{t'}$. Similar conditions hold for Y- and Z-parts. Hence, π_X,π_Y,π_Z also induce permutations on $\mathcal{P}_X,\mathcal{P}_Y,\mathcal{P}_Z$, respectively.*

2. *Every* $(\pi_X, \pi_Y, \pi_Z) \in \mathcal{G}$ preserves the tensor structure of T. *Formally, the coefficient of* $x_i \cdot y_j \cdot z_k$ *in* T *equals the coefficient of* $\pi_X(x_i) \cdot \pi_Y(y_j) \cdot \pi_Z(z_k)$ *in* T, *for all variables* x_i, y_j, z_k.

3. *A uniformly random element* $(\pi_X, \pi_Y, \pi_Z) \in \mathcal{G}$ permutes any given part to a uniform random part. *Formally, for any fixed* $X_t, X_{t'} \in \mathcal{P}_X$, $Y_t, Y_{t'} \in \mathcal{P}_Y$, $Z_t, Z_{t'} \in \mathcal{P}_Z$, *and for a uniformly random* $(\pi_X, \pi_Y, \pi_Z) \in \mathcal{G}$, *we have*

$$\Pr[\pi_X(X_t) = X_{t'}] = \frac{1}{M_X},$$
$$\Pr[\pi_Y(Y_t) = Y_{t'}] = \frac{1}{M_Y},$$
$$\Pr[\pi_Z(Z_t) = Z_{t'}] = \frac{1}{M_Z}.$$

We show the following.

THEOREM 7.1. *Let* T *be a partitioned tensor defined above; let* $T_1, \dots, T_r$ *be broken copies of* T, *where in each* T_i *for* $i \in [r]$, *at most* $\frac{1}{4 \log M_X}$, $\frac{1}{4 \log M_Y}$, *and* $\frac{1}{4 \log M_Z}$ *fraction of* X*-,* Y*-, and* Z*-parts are holes, respectively. If* T *satisfies Property 7.1 with a set of permutations* $\mathcal{G}$, *then there exists a constant* C_0 *such that for* $r \geq C_0 \cdot M^{\frac{3}{\log \log N}}$ *where* $M = \max\{M_X, M_Y, M_Z\}$, *we have*

$$\bigoplus_{i=1}^{r} T_i \trianglerighteq T.$$

In particular, $M^{o(1)}$ *broken copies of* T *with fraction of holes* $O\big(\frac{1}{\log M}\big)$ *can degenerate into an unbroken copy of* T.

Before proving Theorem 7.1, we first show the following Lemma 7.1 that will explain why we need Item 3 in Property 7.1. The lemma essentially states that if T satisfies Property 7.1, then we can find a set of permutations π_X, π_Y, π_Z on the partitions of X-, Y-, Z-variables such that any set of parts can be permuted away from any set of positions that we specify. Specifically, one should think under the context of degenerating a broken copy of T with holes into some subtensor $T|_{X',Y',Z'}$, the lemma states that we can find a set of permutations preserving the tensor structure of T on the variable sets such that the holes are away from the terms in $T|_{X',Y',Z'}$. Then applying the permutation on the broken copy would give the subtensor $T|_{X',Y',Z'}$ without holes or with fewer amount of holes.

LEMMA 7.1. *Let* T *be a tensor satisfying the assumptions of Theorem 7.1 with* $\mathcal{G}$. *Then there exists* $(\pi_X, \pi_Y, \pi_Z) \in \mathcal{G}$ *such that for any sets of parts* $P_X, P'_X \subseteq \mathcal{P}_X$, $P_Y, P'_Y \subseteq \mathcal{P}_Y$, $P_Z, P'_Z \subseteq \mathcal{P}_Z$ *we have*

$$|P_X \cap \pi_X(P'_X)| \leq \frac{4|P_X| \cdot |P'_X|}{|\mathcal{P}_X|},$$
$$|P_Y \cap \pi_Y(P'_Y)| \leq \frac{4|P_Y| \cdot |P'_Y|}{|\mathcal{P}_Y|}, \tag{7.25}$$
$$|P_Z \cap \pi_Z(P'_Z)| \leq \frac{4|P_Z| \cdot |P'_Z|}{|\mathcal{P}_Z|}.$$

Proof. We prove the lemma using a probabilistic argument. Consider a uniformly random element $(\pi_X, \pi_Y, \pi_Z) \in \mathcal{G}$. By Item 3 in Property 7.1, for any $X_t \in \mathcal{P}_X$, we have

$$\Pr[\pi_X(X_t) \in P_X] = \frac{|P_X|}{|\mathcal{P}_X|}.$$

By linearity of expectation

$$\mathbb{E}[|P_X \cap \pi_X(P'_X)|] = \sum_{X_t \in P'_X} \Pr[\pi_X(X_t) \in P_X] = \frac{|P_X| \cdot |P'_X|}{|\mathcal{P}_X|}.$$

Thus by Markov's inequality, have

$$\Pr\left[|\pi_X(X_t) \in P_X| > \frac{4|P_X| \cdot |P'_X|}{\mathcal{P}_X}\right] \le \frac{1}{4}.$$

The argument works similarly for Y and Z, so by union bound over X, Y, Z, with probability $\ge \frac{1}{4}$, a random $(\pi_X, \pi_Y, \pi_Z) \in \mathcal{G}$ satisfies Eq. (7.25). Therefore, we can conclude that there exists such a $(\pi_X, \pi_Y, \pi_Z) \in \mathcal{G}$ satisfying Eq. (7.25). ∎

We now proceed to prove Theorem 7.1. The main idea is to first take a broken copy of T that covers most of the terms in the tensor, then write the missing terms as a sum of 7 smaller subtensors which we treat as 7 subproblems, and finally recurse on each of the subproblems with smaller sizes.

Proof of Theorem 7.1. Assume P_X, P_Y, P_Z are sets of h_X, h_Y, h_Z parts of X-, Y-, Z-dimension respectively, and assume that we need to produce $T\|_{P_X,P_Y,P_Z}$. The number of broken copies of T required for this purpose is denoted as $f(h_X, h_Y, h_Z)$. Clearly, $f(h_X, h_Y, h_Z) = 0$ when one of h_X, h_Y, h_Z equals zero (because $T\|_{P_X,P_Y,P_Z}$ would be an empty tensor), and we need to upper bound $f(M_X, M_Y, M_Z)$, which is the number of broken copies required to produce a complete copy of T.

Take a broken copy of T, namely $T_{\text{hole}} = T\|_{\mathcal{P}_X \setminus P_X^{(0)},\, \mathcal{P}_Y \setminus P_Y^{(0)},\, \mathcal{P}_Z \setminus P_Z^{(0)}}$, where $P_X^{(0)}$, $P_Y^{(0)}$, $P_Z^{(0)}$ are the set of holes. Then, applying Lemma 7.1 on $P_X, P_X^{(0)}, P_Y, P_Y^{(0)}, P_Z, P_Z^{(0)}$ gives $(\pi_X, \pi_Y, \pi_Z) \in \mathcal{S}_{N_X} \times \mathcal{S}_{N_Y} \times \mathcal{S}_{N_Z}$ such that

$$\begin{aligned}
\left|P_X^{(0)'}\right| &:= \left|P_X \cap \pi_X\left(P_X^{(0)}\right)\right| \le 4 \cdot \frac{|P_X| \cdot \left|P_X^{(0)}\right|}{M_X} \le \frac{1}{\log M_X} \cdot |P_X|, \\
\left|P_Y^{(0)'}\right| &:= \left|P_Y \cap \pi_Y\left(P_Y^{(0)}\right)\right| \le \frac{1}{\log M_Y} \cdot |P_Y|, \\
\left|P_Z^{(0)'}\right| &:= \left|P_Z \cap \pi_X\left(P_Z^{(0)}\right)\right| \le \frac{1}{\log M_Z} \cdot |P_Z|.
\end{aligned} \tag{7.26}$$

We relabel the variables in T_{hole} according to the permutations π_X, π_Y, π_Z, obtaining another broken copy of T with sets of holes $\pi_X\left(P_X^{(0)}\right)$, $\pi_Y\left(P_Y^{(0)}\right)$, $\pi_Z\left(P_Z^{(0)}\right)$. We then zero out all parts outside P_X, P_Y, P_Z. The obtained tensor, denoted by T'_{hole}, is a subtensor of the target tensor $T\|_{P_X,P_Y,P_Z}$:

$$T'_{\text{hole}} \;=\; T\|_{P_X \setminus P_X^{(0)'},\, P_Y \setminus P_Y^{(0)'},\, P_Z \setminus P_Z^{(0)'}} \;:=\; T\|_{P_X^{(1)'},\, P_Y^{(1)'},\, P_Z^{(1)'}},$$

where $P_X^{(0)'} := P_X \cap \pi_X\left(P_X^{(0)}\right)$ is the set of holes in X-parts, and $P_X^{(1)'} := P_X \setminus P_X^{(0)'}$; similar for Y- and Z-dimension.

Next, we write $T\|_{P_X,P_Y,P_Z}$ as a sum of 8 subtensors:

$$T\|_{P_X,P_Y,P_Z} \;=\; T\|_{P_X^{(1)'},\, P_Y^{(1)'},\, P_Z^{(1)'}} + \sum_{\substack{a,b,c \in \{0,1\} \\ 0 \in \{a,b,c\}}} T\|_{P_X^{(a)'},\, P_Y^{(b)'},\, P_Z^{(c)'}}.$$

Notice that the first term $T\|_{P_X^{(1)'},\, P_Y^{(1)'},\, P_Z^{(1)'}}$ equals T'_{hole} (which we already obtained by consuming one broken copy T_{hole}), and the other seven subtensors are significantly smaller than $T\|_{P_X,P_Y,P_Z}$, so we can obtain them recursively. The fact that $\left|P_X^{(1)'}\right| \le |P_X| = h_X$, $\left|P_Y^{(1)'}\right| \le |P_Y| = h_Y$, $\left|P_Z^{(1)'}\right| \le |P_Z| = h_Z$ together with Eq. (7.26) gives us the following recursion:

$$\begin{aligned}
f(h_X, h_Y, h_Z) \le 1 &+ f\left(\frac{h_X}{\log M_X}, h_Y, h_Z\right) + f\left(h_X, \frac{h_Y}{\log M_Y}, h_Z\right) + f\left(h_X, h_Y, \frac{h_Z}{\log M_Z}\right) \\
&+ f\left(\frac{h_X}{\log M_X}, \frac{h_Y}{\log M_Y}, h_Z\right) + f\left(\frac{h_X}{\log M_X}, h_Y, \frac{h_Z}{\log M_Z}\right) \\
&+ f\left(h_X, \frac{h_Y}{\log M_Y}, \frac{h_Z}{\log M_Z}\right) + f\left(\frac{h_X}{\log M_X}, \frac{h_Y}{\log M_Y}, \frac{h_Z}{\log M_Z}\right).
\end{aligned}$$

Since $|\mathcal{P}_X| = M_X$, $|\mathcal{P}_Y| = M_Y$, $|\mathcal{P}_Z| = M_Z$, we can solve the recursion for $f(M_X, M_Y, M_Z)$ and get

$$\begin{aligned} f(M_X, M_Y, M_Z) &\le 7^{1+\lceil \log_{\log M_X} M_X \rceil + \lceil \log_{\log M_Y} M_Y \rceil + \lceil \log_{\log M_Z} M_Z \rceil} \\ &\le C_0 \cdot M^{\frac{3}{\log \log M}} \end{aligned}$$

where C_0 is a sufficiently large constant and $M = \max\{M_X, M_Y, M_Z\}$ since the function $M^{1/\log \log M}$ is monotonic increasing for sufficiently large M. ∎

We remark that Theorem 7.1 also works for non-partitioned tensors satisfying Property 7.1 when considering on X-, Y-, Z-variables as partitioned into size-1 parts where each part consists of a single variable.

Now let us return our attention to the context of fast matrix multiplication and show that we can fix the holes in the interface tensors with holes obtained in our algorithm.

COROLLARY 7.1 (Restated). *Let T be a level-ℓ interface tensor with parameter list*

$$\{(n_t, i_t, j_t, k_t, \beta_{X,t}, \beta_{Y,t}, \beta_{Z,t})\}_{t\in[s]}.$$

Let $N = 2^{\ell-1} \cdot \sum_{t\in[s]} n_t$. Suppose $T_1, \dots, T_r$ are broken copies of T where $\le \frac{1}{8N}$ fraction of level-1 X-, Y- and Z-blocks are holes. If $r \ge 2^{C_1 N/\log N}$ for some large enough constant $C_1 > 0$, the direct sum $\bigoplus_{i=1}^{r} T_i$ can degenerate into an unbroken copy of T.

Proof. Consider the level-1 partition of the X-, Y-, Z-variables in T into level-1 blocks indexed by sequences in $\{0,1,2\}^N$ with length exactly $N = 2^{\ell-1} \cdot \sum_{t\in[s]} n_t$ as defined in the statement. By definition, the level-1 blocks remaining in T are consistent with the distributions $\beta_{X,t}, \beta_{Y,t}, \beta_{Z,t}$ over each term $t \in [s]$ in T, which means that every level-1 block $X_{\hat{I}}$ with index sequence $\hat{I} \in \{0,1,2\}^N$ has the same number of 0's, 1's, and 2's. This implies that each level-1 X-variable block contains the same number of variables and the number of level-1 blocks can be bounded by 3^N. Similarly, there are $\le 3^N$ level-1 Y- and Z-variable blocks and the partitions of Y- and Z-variables into level-1 blocks are partitions into equal-sized parts.

We let the partitions of X-, Y-, Z-variables into level-1 blocks be the partitions used for Theorem 7.1, and therefore the number of blocks $M_X, M_Y, M_Z \le 3^N$. Then suppose we can find an appropriate $\mathcal{G} \subseteq \mathcal{S}_{|X|} \times \mathcal{S}_{|Y|} \times \mathcal{S}_{|Z|}$ satisfying Property 7.1 for T, then by Theorem 7.1, as the fraction of holes in every broken copy T_i is at most $\frac{1}{8N} \le \frac{1}{4\log 3^N} \le \min\{\frac{1}{4\log M_X}, \frac{1}{4\log M_Y}, \frac{1}{4\log M_Z}\}$ in all three dimensions, a direct sum of $\left(3^N\right)^{\frac{3}{\log\log 3^N}} = 2^{C_1 \cdot \frac{N}{\log N}}$ broken copies (with sufficiently large constant $C_1 > 0$) of T can degenerate into an unbroken copy of T.

Thus it suffices to construct a set of permutations $\mathcal{G} \subseteq \mathcal{S}_{|X|} \times \mathcal{S}_{|Y|} \times \mathcal{S}_{|Z|}$ that together with T satisfies Property 7.1. Note that every X-, Y-, or Z-variable in T is indexed by a sequence in $\{0,1,\dots,q+1\}^N = \left(\{0,1,\dots,q+1\}^{2^{\ell-1}}\right)^n$, we call every $2^{\ell-1}$ consecutive indices a *chunk* and randomly permute chunks within the same term in T. Specifically, consider the set $\mathcal{H} = \mathcal{S}_{n_1} \times \cdots \times \mathcal{S}_{n_s}$. For each $\sigma = (\sigma_1, \dots, \sigma_s) \in \mathcal{H}$, consider that σ_t permutes the n_t length-$2^{\ell-1}$ chunks in the t-th term for $t \in [s]$. σ can be regarded as a permutation over $[n]$, indicating the destinations of all n chunks. It also induces a permutation $\sigma' \in \mathcal{S}_N$ over N level-1 indices. Formally, the j-th index in the i-th chunk is permuted to the j-th index in the $\sigma(i)$-th chunk, i.e., $\sigma'((i-1)\cdot 2^{\ell-1} + j) = (\sigma(i)-1)\cdot 2^{\ell-1} + j$ for all $i \in [n]$ and $j \in [2^{\ell-1}]$. Further, σ' induces a permutation π_X over all X-variables, given by

$$\pi_X\left(x_{(\hat{i}_1, \hat{i}_2, \dots, \hat{i}_N)}\right) \coloneqq x_{(\hat{i}_{\sigma'(1)}, \hat{i}_{\sigma'(2)}, \dots, \hat{i}_{\sigma'(N)})},$$

where $x_{(\hat{i}_1, \hat{i}_2, \dots, \hat{i}_N)}$ represents the X-variable indexed by $(\hat{i}_1, \hat{i}_2, \dots, \hat{i}_N) \in \{0,1,\dots,q+1\}^N$. The permutations π_Y, π_Z over Y- and Z-variables are defined similarly. Finally, $\mathcal{G}$ is defined as all permutations generated in the above way, i.e., $\mathcal{G} = \{(\pi_X, \pi_Y, \pi_Z) \text{ induced from } \sigma \in \mathcal{H}\}$.

Note that $\mathcal{G}$ is well-defined, since for any element $(\pi_X, \pi_Y, \pi_Z) \in \mathcal{G}$ and any level-1 index sequence $\hat{I}$ in T satisfying the complete split distributions $\{\beta_{X,t}\}_{t\in[s]}$, $\pi_X(X_{\hat{I}})$ must also satisfy the complete split distributions $\{\beta_{X,t}\}_{t\in[s]}$, because the permutation acts on each term individually. Now we check that $\mathcal{G}$ satisfies Property 7.1. It is easy to see by definition that the set $\mathcal{G}$ satisfies Item 1 and Item 2 since variables in one level-1 variable block all get permuted to the same level-1 variable block. Item 3 holds due to the symmetry of the chunks within the same term. ∎

8 Numerical Result

Let $\ell^* > 0$ be an integer and let $N = 2^{\ell^*-1} \cdot n$. Our upper bound of $\omega(1, \kappa, 1)$ is formed by successively applying Theorems 5.1, 6.1 and 6.2 to degenerate $2^{o(n)}$ independent copies of $\mathrm{CW}_q^{\otimes N} \equiv \big(\mathrm{CW}_q^{\otimes 2^{\ell^*-1}}\big)^{\otimes n}$ into independent matrix multiplication tensors of the form $\langle a, a^\kappa, a\rangle$, shown in Algorithm 1.

Algorithm 1: Procedure of Degeneration

Let $\varepsilon > 0$ be a fixed constant and $\ell^* > 0$ be an integer.

1. Degenerate $2^{o(n)}$ independent copies of $\big(\mathrm{CW}_q^{\otimes 2^{\ell^*-1}}\big)^{\otimes n}$ into V_{ℓ^*} (independent) copies of a level-ℓ^* $(\varepsilon \cdot 3^{\ell^*})$-interface tensor $\mathcal{T}_{\ell^*}$, where the number of copies V_{ℓ^*} and the parameter list of $\mathcal{T}_{\ell^*}$ are given in Theorem 5.1 and Proposition 5.1.
2. For each $\ell = \ell^*, \dots, 2$:
   - Degenerate *every* $2^{o(n)}$ copies of the level-ℓ $(\varepsilon \cdot 3^\ell)$-interface tensor $\mathcal{T}_\ell$ into $V_{\ell-1}$ independent copies of the tensor product of a level-$(\ell-1)$ $(\varepsilon \cdot 3^{\ell-1})$-interface tensor $\mathcal{T}_{\ell-1}$ and some matrix multiplication tensor $\langle a_\ell, b_\ell, c_\ell\rangle$. Here, the number of copies $V_{\ell-1}$, the parameter list of $\mathcal{T}_{\ell-1}$ and the matrix multiplication size $\langle a_\ell, b_\ell, c_\ell\rangle$ are all given in Theorem 6.2 and Proposition 6.1.
3. The level-1 3ε-interface tensor $\mathcal{T}_1$ can degenerate into a matrix multiplication tensor, written $\langle a_1, b_1, c_1\rangle$, according to Theorem 6.1.
4. So far, we have obtained $V \coloneqq \prod_{\ell=1}^{\ell^*} V_\ell$ copies of $\langle A, B, C\rangle \equiv \bigotimes_{\ell=1}^{\ell^*} \langle a_\ell, b_\ell, c_\ell\rangle$.

We first let $n \to \infty$ and apply Schönhage's asymptotic sum inequality (Theorem 3.2) on the above degeneration, obtaining a bound on $\omega(1, \kappa, 1)$ which might depend on ε; then, we let $\varepsilon \to 0$, obtaining the bound $\omega(1, \kappa, 1) \le \omega'$ as long as

$$\lim_{\varepsilon\to 0}\lim_{n\to\infty} V^{1/n} \cdot \min\big\{A, B^{1/\kappa}, C\big\}^{\omega'/n} \ge (q+2)^{2^{\ell^*-1}}. \tag{8.27}$$

Every degeneration step in Algorithm 1 requires a set of parameters, including the distribution α over constituent tensors, the proportions of tensor powers A_1, A_2, A_3 assigned to three regions, and others. If we are given an assignment to the parameters, we can precisely calculate

$$\lim_{\varepsilon\to 0}\lim_{n\to\infty} V_\ell^{1/n}, \quad \lim_{\varepsilon\to 0}\lim_{n\to\infty} a_\ell^{1/n}, \quad \lim_{\varepsilon\to 0}\lim_{n\to\infty} b_\ell^{1/n}, \quad \lim_{\varepsilon\to 0}\lim_{n\to\infty} c_\ell^{1/n}$$

according to Theorems 5.1, 6.1 and 6.2. Plugging them into (8.27) would verify the correctness of the claimed bound on $\omega(1, \kappa, 1)$.

Optimization strategy. Finding a set of parameters that lead to the best bound of $\omega(1, \kappa, 1)$ can be modeled as a constrained optimization problem:

$$\begin{aligned} &\text{minimize} && \omega' \\ &\text{subject to} && \text{all constraints in Theorems 5.1, 6.1 and 6.2} \\ &&& \lim_{\varepsilon\to 0}\lim_{n\to\infty} V^{1/n} \cdot \min\big\{A, B^{1/\kappa}, C\big\}^{\omega'/n} \ge (q+2)^{2^{\ell^*-1}}. \end{aligned} \tag{8.28}$$

We used *sequential quadratic programming (SQP)* to solve this optimization problem, which is a well-known iterative approach for solving nonlinear constrained optimization. The software package SNOPT [19] is used for performing SQP. Like all other optimization methods for nonlinear optimization, SQP does not guarantee finding the global optimum or a specific convergence rate; the quality of the solution and the time performance both rely on the *initial point* of the iterative process, which could be provided by the user.

For $\kappa = 1$, we take the parameters from [23] which Le Gall used to analyze $\mathrm{CW}_q^{\otimes 2^{\ell^*-1}}$ for square matrix multiplication, and transform it into a feasible solution to the optimization problem (8.28), which we set as the

initial point. Specifically, Le Gall's parameters consist of a distribution α over level-ℓ^* constituent tensors (for the global stage) together with a split distribution $\alpha_{i,j,k}$ for every constituent tensor $T_{i,j,k}$ (for the constituent stages). We specify our parameters as follows:

- For every constituent tensor $T_{i,j,k}$ that appears in our interface tensors, we directly set $\alpha_{i,j,k}$ as its split distribution in every region, and let $A_1 = A_2 = A_3 = 1/3$, which means that all three regions are symmetric to each other.
- The distribution used in our global stage is set to α as well. Other parameters are uniquely determined by these specified ones.
- For every constituent tensor $T_{i,j,k}$ that contains a zero, say $i = 0$, we choose its complete split distributions $\beta_X, \beta_Y, \beta_Z$ that maximizes its size as an inner product tensor, i.e., maximizes $H(\beta_Y)$.
- Other parameters are uniquely determined by the specified ones.

It is easy to see that these parameters form a feasible solution. Furthermore, these parameters actually lead to the same upper bound on ω as Le Gall's analysis. We start from this feasible solution and perform SQP to obtain an upper bound for $\omega = \omega(1,1,1)$.

For $\kappa \neq 1$, our strategy is to start with a solution for another κ nearby. For example, it is natural to believe that a good solution for $\omega(1, 0.95, 1)$ is similar to that for $\omega(1,1,1)$. Therefore, we use our parameters for $\omega(1,1,1)$ as the initial point for optimizing the bound of $\omega(1, 0.95, 1)$, and proceed with SQP to obtain the bound for $\omega(1, 0.95, 1)$. Then, we can further start with our parameters for $\omega(1, 0.95, 1)$ to obtain parameters for $\omega(1, 0.90, 1)$, and so on.

Lagrange multipliers. In Theorem 5.1, we need to calculate $P_\alpha = \max_{\alpha' \in D} H(\alpha') - H(\alpha)$ where D represents the set of distributions that share marginals with α. Although this definition of P_α is not a closed form in terms of α, we can let the max-entropy distribution $\alpha_{\max} \coloneqq \arg\max_{\alpha' \in D} H(\alpha')$ be an optimizable variable, and use the method of Lagrange multipliers to ensure that α' has the largest entropy among D.

Formally, we first add linear constraints to force $\alpha_{\max}$ and α to have the same marginals:

$$\sum_{j+k=2^{\ell^*}-i} \big(\alpha_{\max}(i,j,k) - \alpha(i,j,k)\big) = 0, \qquad \forall i = 0,1,\dots,2^{\ell^*}, \tag{8.29}$$

$$\sum_{i+k=2^{\ell^*}-j} \big(\alpha_{\max}(i,j,k) - \alpha(i,j,k)\big) = 0, \qquad \forall j = 0,1,\dots,2^{\ell^*}, \tag{8.30}$$

$$\sum_{i+j=2^{\ell^*}-k} \big(\alpha_{\max}(i,j,k) - \alpha(i,j,k)\big) = 0, \qquad \forall k = 0,1,\dots,2^{\ell^*}, \tag{8.31}$$

$$\sum_{i,j,k} \alpha_{\max}(i,j,k) = 1, \tag{8.32}$$

$$\alpha_{\max}(i,j,k) \geq 0, \qquad \forall i+j+k = 2^{\ell^*}. \tag{8.33}$$

Let $\lambda_X(i), \lambda_Y(j), \lambda_Z(k), \lambda_S$ $(0 \leq i,j,k \leq 2^{\ell^*})$ be Lagrange multipliers for (8.29), (8.30), (8.31), (8.32) respectively, which we also treat as optimizable variables. Then the first-order optimality of $H(\alpha_{\max})$ can be written as

$$\lambda_X(i) + \lambda_Y(j) + \lambda_Z(k) + \lambda_S = \ln \alpha_{\max}(i,j,k) + 1, \qquad \forall i+j+k = 2^{\ell^*}. \tag{8.34}$$

(Note that any $\alpha_{\max}$ satisfying (8.34) will also satisfy strict inequalities in (8.33), thus we do not need to create Lagrange multipliers for (8.33).) Since the entropy function $H(\cdot)$ is strictly concave, any $\alpha_{\max}$ satisfying these constraints is guaranteed to have maximum entropy. (Conversely, the true max-entropy distribution $\alpha_{\max}$ will satisfy all these requirements.) We include these Lagrange multiplier constraints (8.34) in our optimization problem (8.28).[5] Similarly, in Theorem 6.2, we also introduce Lagrange multiplier constraints when we need to ensure that some distribution has maximum entropy given its marginals.

[5] In the program, we use the exponential form of (8.28): $\exp(\lambda_X(i) + \lambda_Y(j) + \lambda_Z(k) + \lambda_S - 1) = \alpha_{\max}(i,j,k)$, in order to avoid numerical issues like $\ln 0$.

Smooth the landscape. In Theorems 5.1 and 6.2, the intermediate variables named E_1, E_2, E_3 are minimums of three terms. If we calculate them according to the definition, it would create a "spike" (non-differentiable point) in the landscape, which is unfriendly for many optimizable methods including SQP. (SQP requires all objective and constraint functions to be twice continuously differentiable.) To address this issue, we treat E_1, E_2, E_3 as optimizable variables and transform the minimum into linear inequality constraints:

$$E = \min(x, y, z) \qquad \Rightarrow \qquad E \le x,\ E \le y,\ E \le z.$$

Since E (any of E_1, E_2, E_3) is positively correlated with the number of matrix multiplication tensors we produce, we do not need to worry that E takes on a value smaller than $\min(x, y, z)$. The newly introduced constraints are linear and thus have smooth landscapes. We include these auxiliary optimizable parameters and constraints in the optimization problem (8.28). In practice, we also observe that SQP would not work well without this type of smoothing.

Numerical results. We wrote a MATLAB [26] program to solve the optimization problem (8.28), with the help of SNOPT [19], a software package for solving large-scale optimization problems. By running the program for different κ, we obtained various upper bounds of $\omega(1, \kappa, 1)$, as shown in Table 1. All bounds are obtained by analyzing the fourth power[6] of the CW tensor with $q = 5$. Specifically, we obtained the important bounds $\omega \le 2.371552$, $\alpha \ge 0.321334$, and $\mu \le 0.527661$. The code and parameters are available at `https://osf.io/7wgh2/?view_only=ce1a6a66d9fc432d8f6da39a6ea4b6e4`.

References

[1] J. Alman, *Limits on the universal method for matrix multiplication*, Theory Comput., 17 (2021), pp. 1–30.

[2] J. Alman and V. Vassilevska Williams, *Further limitations of the known approaches for matrix multiplication*, in Proceedings of the 9th Innovations in Theoretical Computer Science Conference (ITCS), 2018, pp. 25:1–25:15.

[3] ———, *Limits on all known (and some unknown) approaches to matrix multiplication*, in Proceedings of the 59th IEEE Annual Symposium on Foundations of Computer Science (FOCS), 2018, pp. 580–591.

[4] ———, *A refined laser method and faster matrix multiplication*, in Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA), 2021, pp. 522–539.

[5] N. Alon, A. Shpilka, and C. Umans, *On sunflowers and matrix multiplication*, Comput. Complex., 22 (2013), pp. 219–243.

[6] A. Ambainis, Y. Filmus, and F. Le Gall, *Fast matrix multiplication: limitations of the Coppersmith-Winograd method*, in Proceedings of the 47th Annual ACM on Symposium on Theory of Computing (STOC), 2015, pp. 585–593.

[7] F. A. Behrend, *On sets of integers which contain no three terms in arithmetical progression*, Proceedings of the National Academy of Sciences of the United States of America, 32 (1946), p. 331.

[8] J. Blasiak, T. Church, H. Cohn, J. A. Grochow, E. Naslund, W. F. Sawin, and C. Umans, *On cap sets and the group-theoretic approach to matrix multiplication*, Discret. Anal., 2017 (2017), pp. 1–27.

[9] J. Blasiak, T. Church, H. Cohn, J. A. Grochow, and C. Umans, *Which groups are amenable to proving exponent two for matrix multiplication?*, arXiv:1712.02302, (2017).

[10] M. Christandl, P. Vrana, and J. Zuiddam, *Barriers for fast matrix multiplication from irreversibility*, Theory Comput., 17 (2021), pp. 1–32.

[11] D. Coppersmith, *Rapid multiplication of rectangular matrices*, SIAM J. Comput., 11 (1982), pp. 467–471.

[12] ———, *Rectangular matrix multiplication revisited*, J. Complex., 13 (1997), pp. 42–49.

[13] D. Coppersmith and S. Winograd, *Matrix multiplication via arithmetic progressions*, J. Symb. Comput., 9 (1990), pp. 251–280.

[14] A. Czumaj, M. Kowaluk, and A. Lingas, *Faster algorithms for finding lowest common ancestors in directed acyclic graphs*, Theor. Comput. Sci., 380 (2007), pp. 37–46.

[15] A. M. Davie and A. J. Stothers, *Improved bound for complexity of matrix multiplication*, Proceedings of the Royal Society of Edinburgh: Section A Mathematics, 143 (2013), pp. 351–369.

[16] R. Duan. Personal communication, 2022.

[17] R. Duan, H. Wu, and R. Zhou, *Faster matrix multiplication via asymmetric hashing*, in Proceedings of the 64th IEEE Symposium on Foundations of Computer Science (FOCS), 2023.

[6] Our analysis also works for the eighth power, but it was too slow to solve the optimization problem due to the large number of parameters.

[18] F. Eisenbrand and F. Grandoni, *On the complexity of fixed parameter clique and dominating set*, Theor. Comput. Sci., 326 (2004), pp. 57–67.
[19] P. E. Gill, W. Murray, and M. A. Saunders, *SNOPT: An SQP algorithm for large-scale constrained optimization*, SIAM Rev., 47 (2005), pp. 99–131.
[20] X. Huang and V. Y. Pan, *Fast rectangular matrix multiplication and applications*, J. Complex., 14 (1998), pp. 257–299.
[21] S. Ke, B. Zeng, W. Han, and V. Y. Pan, *Fast rectangular matrix multiplication and some applications*, Science in China Series A: Mathematics, 51 (2008), pp. 389–406.
[22] F. Le Gall, *Faster algorithms for rectangular matrix multiplication*, in Proceedings of the 53rd Annual IEEE Symposium on Foundations of Computer Science (FOCS), 2012, pp. 514–523.
[23] ———, *Powers of tensors and fast matrix multiplication*, in Proceedings of the 39th International Symposium on Symbolic and Algebraic Computation (ISSAC), 2014, pp. 296–303.
[24] ———, *Faster rectangular matrix multiplication by combination loss analysis*, in Proceedings of the 2024 ACM-SIAM Symposium on Discrete Algorithms (SODA), 2024, p. to appear.
[25] F. Le Gall and F. Urrutia, *Improved rectangular matrix multiplication using powers of the Coppersmith-Winograd tensor*, in Proceedings of the 29th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), 2018, pp. 1029–1046.
[26] *MATLAB 9.12 (R2022a)*. The MathWorks Inc., Natick, Massachusetts, 2022.
[27] J. Nešetřil and S. Poljak, *On the complexity of the subgraph problem*, Comment. Math. Univ. Carol., 26 (1985), pp. 415–419.
[28] R. Salem and D. C. Spencer, *On sets of integers which contain no three terms in arithmetical progression*, Proceedings of the National Academy of Sciences, 28 (1942), pp. 561–563.
[29] A. Schönhage, *Partial and total matrix multiplication*, SIAM J. Comput., 10 (1981), pp. 434–455.
[30] A. Shapira, R. Yuster, and U. Zwick, *All-pairs bottleneck paths in vertex weighted graphs*, Algorithmica, 59 (2011), pp. 621–633.
[31] V. Strassen, *Gaussian elimination is not optimal*, Numer. Math., 13 (1969), pp. 354–356.
[32] ———, *The asymptotic spectrum of tensors and the exponent of matrix multiplication*, in Proceedings of the 27th Annual Symposium on Foundations of Computer Science (FOCS), 1986, pp. 49–54.
[33] V. Vassilevska Williams, *Multiplying matrices faster than Coppersmith-Winograd*, in Proceedings of the 44th Symposium on Theory of Computing Conference, (STOC), 2012, pp. 887–898.
[34] U. Zwick, *All pairs shortest paths using bridging sets and rectangular matrix multiplication*, J. ACM, 49 (2002), pp. 289–317.