Electronic Colloquium on Computational Complexity, Report No. 30 (2025)

Stronger Cell Probe Lower Bounds via Local PRGs

Oliver Korten* Toniann Pitassi† Russell Impagliazzo‡

Abstract

In this work we observe a tight connection between three topics: NC^0 cryptography, NC^0 range avoidance, and static data structure lower bounds. Using this connection, we leverage techniques from the cryptanalysis of NC^0 PRGs to prove state-of-the-art results in the latter two subjects. Our main result is a quadratic improvement to the best known static data structure lower bounds, breaking a barrier which has stood for several decades. Prior to our work, the best known lower bound for any explicit problem with M inputs and N queries was $S \geq N^{\frac{1}{t}}(\log M)^{1-\frac{1}{t}}$ for any setting of the word length w (where $S =$ space and $t =$ time) [Sie89]. We prove, for the same class of explicit problems considered in [Sie89], a quadratically stronger lower bound of the form $S \geq \tilde{\Omega}\left(N^{\frac{2}{t}} \cdot (\log M)^{1-\frac{2}{t}} \cdot 2^{-O(w)}\right)$ for all even $t > 0$. Second, for the restricted class of *nonadaptive bit probe* data structures, we improve on this lower bound polynomially: for all odd constants $t > 1$ we give an explicit problem with N queries and $M \leq N^{O(1)}$ inputs and prove a lower bound $S \geq \Omega(N^{\frac{2}{t}+\epsilon_t})$ for some constant $\epsilon_t > 0$. Our results build off of an exciting body of work on refuting *semi-random* CSPs (e.g., [AGK21, GKM22, HKM23]).

We then utilize our explicit cell probe lower bounds to obtain the best known unconditional algorithms for NC^0 range avoidance: we can solve any instance with stretch $n \mapsto m$ in polynomial time once $m >> n^{\frac{t}{2}}$ when t is even; with the aid of an NP oracle we can solve any instance with $m > n^{\frac{t}{2}-\epsilon_t}$ for $\epsilon_t > 0$ when t is odd. Finally, using our main correspondence we establish novel *barrier results* for obtaining significant improvements to our cell probe lower bounds: (i) near-optimal space lower bounds for an explicit problem with $t = 4, w = 1$ implies $\mathsf{EXP}^{\mathsf{NP}} \not\subseteq \mathsf{NC}^1$; (ii) under the widely-believed assumption that polynomial-stretch NC^0 PRGs exist, there is no *natural proof* of a lower bound of the form $S \geq N^{\Omega(1)}$ when $t = \omega(1)$, $w = 1$.

*Department of Computer Science, Columbia University. oliver.korten@columbia.edu

†Department of Computer Science, Columbia University. tonipitassi@gmail.com. Partially supported by NSF AF:Medium 2212136.

‡Department of Computer Science, University of San Diego. rimpagliazzo@ucsd.edu. Partially supported by NSF AF:Medium 2212136

ISSN 1433-8092

1 Introduction

We start by briefly introducing the three main subjects of the work: static data structure lower bounds, NC^0 pseudorandom generators, and range avoidance for NC^0 circuits.

Static Data Structure Lower Bounds: The cell probe model of data structures introduced in [EF75, Yao81] is a fundamental computational model in which to study tradeoffs between storage space and access time for information retrieval problems. In the *static* cell probe model, a data structure problem $F : [N] \times [M] \to \{0,1\}$ is fixed, where $[M]$ is an abstract set of "datapoints," $[N]$ is an abstract set of "queries" one might wish to make about a data point, and $F(x,y)$ is the correct answer to the x^{th} query about the datapoint y. The goal is to preprocess any given datapoint $y \in [M]$ into a data structure $\mathcal{E}(y) \in \Sigma^S$ consisting of S "cells," each cell holding a character from some finite alphabet Σ, so that later we may answer different queries $x \in [N]$ about y by reading at most t cells from the data structure $\mathcal{E}(y)$. S is called the "space complexity," t the "time complexity," and $w = \log|\Sigma|$ the "word-length" (number of bits to store the contents of a single cell). As a prototypical example (which our main lower bounds will apply to), let $d \leq n$ be given, $N = 2^n$, $M = 2^{\binom{n}{\leq d}}$. Identifying $[N]$ with $\mathbb{F}_2^n$, and $[M]$ with the set of degree $\leq d$ multilinear polynomials over $\mathbb{F}_2^n$, we can define the *polynomial evaluation problem* $\mathbb{F}_2\text{-Eval}_n^d : [N] \times [M] \to \{0,1\}$ with $\mathbb{F}_2\text{-Eval}_n^d(x,p) = p(x)$; in this case to solve the data structure problem, we want a method for preprocessing polynomials p into small space data structures $\mathcal{E}(p)$, so that upon any query $x \in \mathbb{F}_2^n$ we may evaluate the polynomial p at the point x by querying only a few cells of the data structure.

The basic problem in static data structure lower bounds is to prove, for a particular problem F, that any data structure solution for F must use a large amount of either space, time, or word length (word length is generally of secondary importance). For *any explicit problem* $F : [N] \times [M] \to \{0,1\}$, the best lower bound provable by techniques known prior to this work, for any $t > 2$ and any setting of the word length w, is at best:

$$S \geq N^{\frac{1}{t}} \cdot (\log M)^{1-\frac{1}{t}} \tag{1}$$

Lower bounds of this kind were first achieved by [Sie89] using a method that was rediscovered in later work and which is now known as *cell sampling*. A significant body of work on static data structure lower bounds has formed since, some of which we discuss later in Section 1.4. None of these works are able to prove a lower bound superior to (1).

NC^0 Pseudorandom Generators: A prominent line of work initiated in [Gol00, CM01] has studied the existence of cryptographic PRGs in NC^0. An NC^0_t-PRG is function $G : \{0,1\}^n \to \{0,1\}^m$, $m > n$, so that: (1) each output depends on only t inputs, and (2) G is a cryptographic pseudorandom generator: no polynomial time algorithm can distinguish a random output of G from a truly random m bit string. We say that G is an NC^0 generator if it is NC^0_t for some $t = O(1)$. The existence of such PRGs has been the subject of a great deal of work over the past two decades; we refer the reader to a comprehensive survey [App16] and the references therein.

An important parameter is the *stretch* of the generator, which denotes the relation between n, m. We say that G has nontrivial stretch if $m > n$, and polynomial stretch if $m \geq n^{1+\Omega(1)}$. NC^0 generators with nontrivial stretch can be shown to exist under standard number theoretic cryptographic assumptions [AIK06], however their utility has so far been limited. On the other hand, NC^0 generators with polynomial stretch are not known to exist under any more standard hardness assumptions, but have been found to have a wide range of advanced cryptographic applications [IKOS08], most notably in the construction of Indistinguishability Obfuscation [JLS21, JLS22]. For this reason, a

great deal of work has gone into studying the existence of polynomial stretch NC^0 PRGs as a foundational assumption in its own right, and various algorithmic methods for distinguishing such PRGs have been studied. The current state of the art shows that an NC^0_t generator $G : \{0,1\}^n \to \{0,1\}^m$ can be broken in polynomial time once $m >> n^{\frac{t}{2}}$ ([MST03, AOW15, AGK21]).

NC^0 Range Avoidance: We finally turn to our third main subject of study, NC^0 range avoidance. The object in question here is the same as above: an NC^0_t generator $G : \{0,1\}^n \to \{0,1\}^m$ with $m > n$ and $t = O(1)$. However, instead of trying the distinguish the output of G from random, our goal now is to solve the *range avoidance problem* for G: output a string $y \in \{0,1\}^m$ such that $y \notin \text{range}(G)$. The range avoidance problem (for general $G : \{0,1\}^n \to \{0,1\}^m$ computed by polynomial size circuits) was introduced in [KKMP21], and shown in [Kor21] to be intimately connected to derandomization and to the problem of proving circuit lower bounds for exponential time classes. In all cases, it is already of great interest to develop algorithms for this problem that use an NP oracle. [RSW22] was the first to investigate the range avoidance problem for NC^0 circuits, and showed that a P^{NP} algorithm for this problem would imply the breakthrough circuit lower bound $\mathsf{EXP}^{\mathsf{NP}} \not\subseteq \mathsf{NC}^1$. The same question of *stretch* arises here: the aforementioned result of [RSW22] requires us to solve instances with barely nontrivial stretch. On the other hand unconditional algorithms have been given once the stretch is sufficiently large, in particular once $m >> \frac{n^{t-1}}{\log n}$ ([GLW22, GGNS23]). An important question posed in [RSW22] is whether unconditional P^{NP} algorithms for NC^0 range avoidance can be found in the general polynomial stretch regime.

The Main Correspondence: The starting point of our work is the observation that the 3 problems discussed above are tightly related to one another. The relation between NC^0 PRGs and NC^0 range avoidance is rather obvious (although still insufficiently explored prior to this work), and so the key point is the relation we observe between these two topics pertaining to NC^0 circuits on the one hand, and the project of proving cell probe lower bounds on the other. This relationship follows directly from the definitions of the objects in question, and is more appropriately understood as a *perspective shift* rather than a new result.

To state the connection in its simplest form, we focus our attention for now on a special setting of the cell probe model: the word length is 1 (i.e. $\Sigma = \{0,1\}$, often called the *bit probe* model), and the data structures in question are *nonadpative*. Here, nonadpative means that the data structure will decide which cells to probe as a function only of the query, in contrast to a general *adaptive* data structure which may decide on the next cell to probe based on the outcomes of previous probes. In this setting we have:

Observation. *Let $F : [N] \times [M] \to \{0,1\}$ be a data structure problem, which we interpret as a matrix $F \in \{0,1\}^{N \times M}$ with N rows and M columns below. The following are equivalent:*

1. *F has* does not have *nonadaptive bit probe data structures with space complexity S and time complexity t.*
2. *The columns of F are a set of M strings $\mathcal{F} \subseteq \{0,1\}^N$, $|\mathcal{F}| = M$, such that for any NC^0_t generator $G : \{0,1\}^S \to \{0,1\}^N$, there exists a string $f \in \mathcal{F}$ with $f \notin \text{range}(G)$.*

Proof. Say that F has nonadaptive bit probe data structures of space S and time t. Then, to every $y \in [M]$ we may associate an encoding $E_y \in \{0,1\}^S$, so that for every $x \in [N]$ we may determine $F(x,y)$ by probing (nonadaptively) t bits of E_y. Considering the y^{th} column of F as a string $f_y \in \{0,1\}^N$, we see that there is a NC^0_t function $G : \{0,1\}^S \to \{0,1\}^N$ so that $G(E_y) = f_y$ for every $y \in [M]$; in particular, the x^{th} output of G runs the data structure's nonadaptive query procedure corresponding to the query x. The same reasoning applies in the reverse direction. □

In other words, finding an explicit data structure problem $F : [N] \times [M] \to \{0,1\}$ which requires large space for time t in the nonadaptive bit probe model is *exactly equivalent* to constructing explicit *list-solutions* for NC^0_t range avoidance. By a list-solution, we mean a list of candidate strings, so that for *any instance* of NC^0_t range avoidance G (with appropriate parameters), one of the strings in this list must be a range avoidance solution for G. We emphasize that an analogous connection exists for the general cell probe model (adaptive and with arbitrary word length), which is explained formally in Section 2.2. The problem of finding list-solutions for NC^0 range avoidance was considered already in [GLW22]. Indeed, via the above correspondence the results of [GLW22] recover a nonadaptive bit-probe lower bound of [GGS23] in a certain range of parameters.

The connection between NC^0-PRGs and cell probe lower bounds is more difficult to state formally, but even more central to our main results. In one direction, we will show how to use the best known methods for distinguishing NC^0 PRGs to achieve state of the art lower bounds in the cell probe model; this is the most significant contribution in this work, and after stating our main results below we will describe this methodology in more detail in Section 1.2 ("Our Techniques"). In the other direction we will use the conjectured security of various NC^0 PRGs to exhibit a *natural properties barrier* for data structure lower bounds, in the spirit of Razborov-Rudich [RR97].

We note that connections of a similar flavor have been utilized in the literature on sampling lower bounds initiated by Viola [Vio12]. Connections between restricted variants of range avoidance and data structures were first observed in [GLW22], who used a cell probe data structure from [Pat08] to develop more efficient reductions to range avoidance; this proof was our original inspiration to explore more deeply the connections between NC^0 range avoidance and the cell probe model.

1.1 Our Results

State of the Art Cell Probe Lower Bounds: Using the perspective shift introduced above, we are able to break a decades-old barrier in static data structure complexity, and prove cell probe lower bounds for a explicit problems which are quadratically better (for space as a function of time) than all previous methods. Our first lower bound holds any data structure problem supporting a k-wise independent distribution; this is the class same problems originally considered by [Sie89] and in various follow up works (including [PTW10, Lar12]), and we are able to obtain tight lower bounds across essentially the entire range of values $k \leq N$, provided the time complexity t is an even number:

Theorem 1. *Let $F : [N] \times [M] \to \{0,1\}$ be a data structure problem such that there exists a k-wise independent distribution supported on its columns. Let $S, t, w \in \mathbb{N}$ be given with t an even number and assume $k > tw + 1$. Then, for any space S, time t, word length w cell probe data structure solving F we must have:*

1. $S \geq \left(\frac{N}{\log N}\right)^{\frac{2}{t}} \cdot \left(\frac{k}{\log N}\right)^{1-\frac{2}{t}} \cdot t^{-1} 2^{-O(w)}$ *for all* $t \log N \leq k \leq N 2^{-O(tw)} t^{-\frac{t}{2}}$
2. $S \geq \left(\frac{N}{\log N}\right)^{\frac{1}{t(\frac{1}{2}+\frac{2}{k})}} \cdot t^{-1} 2^{-O(w)}$ *for all* $tw + 1 \leq k \leq t \log N$

Remember that we often interpret $F : [N] \times [M] \to \{0,1\}$ as a boolean matrix with N rows and M columns, which explains our use of the phrase "the columns of F." If t is an odd number, the above lower bound holds with $t + 1$ in place of t; hence this lower bound gives a polynomial improvement over previous methods for all $t > 3$, gives a near-quadratic improvement for all even t, and approaches a quadratic improvement for large odd values of t. Using any of the various explicit

examples of k-wise independent distributions with support size $N^{O(k)}$ (e.g. univariate polynomials over $\mathbb{F}_{2^n}$), we obtain explicit problems $F : [N] \times [M] \to \{0,1\}$ satisfying the lower bound

$$S \geq \tilde{\Omega}\Big(N^{\frac{2}{t}} \cdot \big(\log M\big)^{1-\frac{2}{t}} \cdot 2^{-O(w)}\Big)$$

across essentially the entire spectrum of nontrivial values $M \in \{1, \ldots, 2^N\}$. We thus achieve a near-quadratic improvement in the state of the art explicit cell probe lower bounds across this entire spectrum of relations between M, N. We emphasize that, in addition to giving a quadratic improvement over the *cell sampling* bound (the first improvement of any kind), our lower bound is also the first to achieve a polynomial improvement over the much simpler *communication bound* $S \gtrsim N^{\frac{1}{t}}$ for a problem with N queries and $M \leq \exp(N^{o(1)})$ datapoints (by setting $k = \log N$ in the above); this is discussed further in the Section 1.4.

By known constructions [Sie89, Sie04] (see also [LPS97]), the lower bound in Theorem 1 is tight up to $\log N$ factors and the factors depending only on w provided t is not too large. Hence, for all small enough even values of t and small enough w, we essentially completely resolve the cell probe complexity of representing k-wise independent distributions. The almost-matching upper bounds are nonconstructive, relying on unbalanced bipartite graphs with very strong unique-expansion properties; they yield data structures which are nonadaptive and have word length $w = 1$, so the tightness of Theorem 1 holds even in this restricted setting of nonadaptive bit probe data structures.

Our lower bound for k-wise independent problems is actually a bit stronger than stated above, as we are able to establish the same bound for problems which only exhibit γ-*almost* k-wise independence, for a nontrivially large value of γ. The stronger form (which is the form which will appear as Theorem 1 in Section 3.2) is as follows:

Theorem (Theorem 1, Strengthened To Almost Independence)**.** *Let $F : [N] \times [M] \to \{0,1\}$ be a data structure problem such that there exists a γ-almost k-wise independent distribution supported on its columns. Then the same lower bound as in the first presentation of Theorem 1 (above) holds for F provided $\gamma \leq N^t 2^{-O(ktw)}$.*

An important consequence of this improvement is that, using known constructions of almost independent distributions [NN90], in the case $t, w = O(1)$ we can set $k = O(\log N), \gamma = N^{-O(1)}$ and obtain an explicit problem $F : [N] \times [M] \to \{0,1\}$ with $M \leq \text{poly}(N)$ and satisfying the lower bound $S \geq \tilde{\Omega}(N^{\frac{2}{t}})$. This improvement will be necessary later on for obtaining polynomial time range avoidance algorithms with minimum possible stretch.

We next establish a lower bound that improves on the previous theorems in the special setting of *nonadaptive bit probe* data structures; remember that "bit probe" denotes the word length setting $w = 1$. For this lower bound, the hardness property of our data structure problems is *small bias* [NN90]: a distribution over $\{0,1\}^N$ is γ-biased if it fools all parity functions with error $\leq \frac{\gamma}{2}$.

Theorem 2. *Let $t > 1$, $t = O(1)$ be a fixed odd number. Let $F : [N] \times [M] \to \{0,1\}$ be a data structure problem such that there exists a N^{-c}-biased distribution supported on its columns, for $c = O(1)$ a sufficiently large constant (depending on t). Then any nonadaptive bit probe data structure for F with space S and time t must satisfy:*

$$S \geq \tilde{\Omega}(N^{\frac{1}{\frac{t}{2} - \frac{t-2}{2(t+2)}}}) \geq N^{\frac{2}{t} + \epsilon_t}$$

where $\epsilon_t > 0$ is a positive constant depending only on t.

Again, using explicit constructions of small bias probability spaces in [NN90], we obtain explicit problems $F : [N] \times [\text{poly}(N)] \to \{0,1\}$ which satisfy the lower bound $S \geq N^{\frac{2}{t}+\epsilon_t}$ for every odd constant t, where $\epsilon_t > 0$ depends only on t. We emphasize that, because of the optimality of the bound in Theorem 1 (there is an upper bound almost matching the bound in Theorem 1 even in the nonadaptive bit probe setting and even for odd values of t), the above Theorem 2 is provably unattainable using *only limited independence*. More formally, to construct any data structure problem $F : [N] \times [M] \to \{0,1\}$ satisfying a lower bound $S \geq N^{\frac{2}{t}+\epsilon}$ for any constants $t > 1$, $\epsilon > 0$, and $M \leq \exp(N^{o(1)})$, we provably cannot use k-wise independence as our only hardness property for the problem F. This result is therefore of special interest, as it is the first cell probe lower bound that is provably stronger than the best bound provable using limited independence alone: the previous best lower bound methods [Sie89, PTW10, Lar12, GGS23] can be carried out using limited independence as the sole hardness condition.

NC^0 Range Avoidance Algorithms: In Section 5 we give the best known unconditional algorithms for NC^0 range avoidance. Using our primary observation connecting range avoidance to data structure lower bounds together with known constructions of limited independence and small bias spaces [NN90], we can immediately obtain state of the art P^{NP} algorithms for NC^0 range avoidance using our new cell probe lower bounds as a black box. In the case of Theorem 2 we get:

Theorem 3. *Let $t = O(1)$ be a fixed odd number. There is a polynomial time algorithm which outputs a list of strings in $\{0,1\}^m$, such that for any NC^0_t generator $G : \{0,1\}^n \to \{0,1\}^m$ with*

$$m \geq \tilde{O}(n^{\frac{t}{2} - \frac{t-2}{2(t+2)}})$$

one of the strings in the list must be a range-avoidance solution for G. In particular, there is a polynomial time NP-oracle algorithm for range avoidance on instances of the above form.

For our lower bound in Theorem 1, we are able to analyse it more effectively (rather than applying it as a black box) and remove the use of an NP oracle in the associated range avoidance algorithm:

Theorem 4. *Let t, n be given with t even. There is a $m^{O(t^2)}$-time algorithm which solves range avoidance given any NC^0_t generator $G : \{0,1\}^n \to \{0,1\}^m$ with*

$$m \geq n^{\frac{t}{2}} \log n \cdot t^{\frac{t}{2}+o(t)}$$

More generally, the algorithm works if each output of G is computed by a depth t decision tree.

We emphasize the last point; not only do we decrease the required stretch quadratically for NC^0_t functions compared to previous results [GLW22, GGNS23], but our algorithm also works in the more general setting that every output of G is a depth t decision tree over the inputs. The full statement of Theorem 4 is more general in two other ways which we didn't highlight above. First, it solves the harder *remote point* problem for G: for any constant $\epsilon > 0$, our algorithm will output a string which is $(\frac{1}{2} - \epsilon)$-far in relative hamming distance from every string in the range of G. Second, it works if G is of the form $G : \Sigma^n \to \{0,1\}^m$ for some nonboolean alphabet $|\Sigma| > 2$, and each output is computable by an adaptive decision tree querying at most t input symbols.

Small Bias Generators in NC^0: In [MST03] it was shown that, if $G : \{0,1\}^n \to \{0,1\}^m$ is an NC^0_t generator such that $G(\boldsymbol{x})$ is $\text{poly}(n)^{-1}$ biased when $\boldsymbol{x}$ is uniformly distributed on $\{0,1\}^n$, then we must have $m \leq O(2^t n^{\lceil \frac{t}{2} \rceil})$; in other words, once $m >> n^{\lceil \frac{t}{2} \rceil}$ there is an $\mathbb{F}_2$-linear test that can distinguish $G(\boldsymbol{x})$ from random. Our Theorem 2 improves this result for all odd values of t; in particular, as an immediate corollary we obtain:

Corollary 1. *For every odd constant t there is $\epsilon_t > 0$, so that every NC^0_t generator $G : \{0,1\}^n \to \{0,1\}^m$ is distinguished with $\mathrm{poly}(n)^{-1}$ advantage by an $\mathbb{F}_2$-linear test once $m \geq n^{\frac{t}{2}-\epsilon_t}$.*

We note that Theorem 2 in fact implies something much stronger, that $G(\boldsymbol{x})$ can be distinguished from uniform by linear tests for *any* distribution of $\boldsymbol{x} \in \{0,1\}^n$. The original result in [MST03] is specially tailored to the case where the random "seed" $\boldsymbol{x}$ is uniformly distributed on $\{0,1\}^n$, and hence cannot be used to give a lower bound in the bit probe model. An interesting feature of this is the following: because $\mathbb{F}_2$-linear tests can be performed by polynomial-size circuits, our result implies that every NC^0_t PRG can be distinguished efficiently by *nonuniform* algorithms in the stretch regime $n \mapsto n^{\frac{t}{2}-\epsilon_t}$. However, in the typical approach to NC^0 PRG constructions pioneered by Goldreich [Gol00], the generator G is sampled at random from a distribution of large ($\approx m$) entropy; in particular, such a cryptographic primitive does not consist of a single NC^0 PRG, but rather a *sampleable ensemble* of NC^0 PRGs. Since our Corollary 1 gives no indication as to how to *find* the linear distinguisher for G given the description of G, we seemingly cannot use it to attack Goldreich's generator in the stretch regime $n \mapsto n^{\frac{t}{2}-\epsilon_t}$, even if we allow nonuniformity in our attacker.

Complexity-Theoretic Barriers to Higher Cell Probe Lower Bounds: Next, in Section 6 we use our main correspondence between cell probe lower bounds and local generators to establish *barrier results* for improving state of the art cell probe lower bounds. Our focus here is on the bit probe model (word length $w = 1$), and on nonadaptive data structures with time complexity $t = O(1)$. We show the following two barrier-type results:

Theorem 5 (Consequences of Explicit Data Structure Lower Bounds)**.**

1. *If there is an explicit data structure problem $F : [N] \times [M] \to \{0,1\}$ with $M \leq \exp(N^{o(1)})$ which requires space $S \geq N - N^{o(1)}$ for data structures making 4 nonadaptive bit probes, then $\mathsf{EXP}^{\mathsf{NP}} \not\subseteq \mathsf{NC}^1$.*

2. *If there is a universal constant $\epsilon > 0$ and an explicit data structure problem $F : [N] \times [M] \to \{0,1\}$ with $M = \mathrm{poly}(N)$ requiring space $S \geq N^{\epsilon}$ for nonadaptive data structures making $O(1)$ bit probes, then there is a polynomial time NP oracle algorithm for NC^0 range avoidance with polynomial stretch.*

The conclusion in (1) above would be a monumental breakthrough in complexity theory and is considered extremely hard. The conclusion in (2) above was directly posed in [RSW22] as an important direction in unconditional range avoidance algorithms. We note that (1) of Theorem 5 has a similar flavor to results in [Vio19, DGW19]. [Vio19] shows that improved data structure lower bounds imply breakthrough lower bounds against certain classes of low-depth circuits over a complete basis of large fan-in gates. On the other hand, [DGW19] shows that improved data structure lower bounds (of a different kind) imply breakthrough constructions of rigid matrices. Neither of these results ([Vio19] or [DGW19]) seem formally comparable ours.

Next, we turn toward a different and more novel kind of barrier in the realm of data structure lower bounds, based on the *natural proofs* paradigm of Razborov and Rudich [RR97]. Following Razborov-Rudich, we define a *natural data structure lower bound* to be some efficiently testable property of functions $F : [N] \times [M] \to \{0,1\}$, so that random functions have the property, and any function with the property satisfies a corresponding cell probe lower bound (see Section 6.2 for a formal definition). We argue in Section 6.2 that the strongest lower bound methods in the cell probe model known prior to this work can be made *natural* in this sense. This continues to hold for our new lower bound in Theorem 1. On the other hand, we show that under two widely

believed cryptographic assumptions, certain lower bounds in the bit probe model are unachievable using proofs of this nature:

Theorem 6 (Natural Proofs Barriers for Data Structure Lower Bounds)**.**

1. *Assume the existence of* NC^1 *PRGs. Then there is no natural proof of a space lower bound* $S \geq N - N^{0.99}$ *for data structures making 4 nonadaptive bit probes solving a problem* $F : [N] \times [\mathrm{poly}(N)] \to \{0,1\}$*.*
2. *Assume existence of polynomial stretch* NC^0 *PRGs. Then for every* $\epsilon > 0$ *there is* $t \in \mathbb{N}$ *so that no natural proof can establish a lower bound* $S \geq N^{\epsilon}$ *for data structures making* t *nonadaptive bit probes solving a problem* $F : [N] \times [\mathrm{poly}(N)] \to \{0,1\}$*.*

As mentioned above our lower bound in Theorem 1 can be made natural, as well as all lower bounds proven prior using the cell sampling method or communication arguments. However, interestingly, we *do not* know if the lower bound for small bias distributions in Theorem 2 can. It is therefore unclear at the moment whether or not the "natural properties barrier" for data structures, according to the definitions used above, has already been broken by the new results in this paper. Regardless, we argue in Section 6.2 that the lower bound in Theorem 2 is still "natural" in a weaker and less formal sense, and that significantly stronger lower bounds will probably have to differ from it in this regard assuming the nonuniform security of NC^0 PRGs. We discuss further the question of the "naturalness" of Theorem 2 in Section 5.

Approaches to Stronger Lower Bounds via Communication Complexity: Finally, in Section 6.3 we discuss possible approaches to proving lower bounds in the nonadaptive bit probe model of the form $S \geq N^{\epsilon}$ for time $t = O(1)$, where $\epsilon > 0$ is a fixed universal constant not depending on t; this corresponds precisely to the regime in which the existence of cryptographically secure polynomial-stretch NC^0 PRGs yields a natural proofs barrier. We formulate cell probe lower bounds in this regime in terms of an equivalent communication model, reminiscent of PIR schemes, in which a single party Alice holding an input $x \in X$ communicates with a council of t Bobs holding a shared input $y \in Y$. We discuss an approach to analyzing this model based on "lifting" a query complexity measure investigated by [HR15, LRT22], and relate this model to other *high end* communication complexity classes such as $\mathsf{PH}^{\mathsf{cc}}$.

1.2 Our Techniques: Derandomizing Semirandom CSP Refutations

We now discuss at a high level our approach to proving our main cell probe lower bound for k-wise independent problems (Theorem 1). We will narrow down our discussion in three respects: (1) we focus on the case $t = O(1)$ and ignore multiplicative factors depending on t in our space lower bound, (2) we restrict attention to *nonadaptive bit probe* data structures, and (3) we focus on achieving a lower bound of the form $S \geq \tilde{\Omega}(N^{\frac{2}{t}})$ for k-wise independent problems in the special case $k = O(\log N)$.

In this setting, using our main observation connecting NC^0 generators and data structures, we need to show that any $O(\log N)$-wise independent random variable $\boldsymbol{f}^{\mathrm{pseud}} \in \{0,1\}^N$ has the following property: for any NC^0_t generator $G : \{0,1\}^S \to \{0,1\}^N$ with $S = \tilde{o}(N^{\frac{2}{t}})$, with nonzero probability $\boldsymbol{f}^{\mathrm{pseud}} \notin \mathrm{range}(G)$. Say for a moment that we view a candidate generator G with the above parameters as a *cryptographic PRG*. In this case, a body of work dedicated to the cryptanalysis of NC^0 PRGs has shown that any such G can be distinguished by polynomial time algorithms; more strongly, there is a polynomial time algorithm which can generate certificates of

the fact $\boldsymbol{f}^{\text{unif}} \notin \text{range}(G)$ with very high probability when $\boldsymbol{f}^{\text{unif}} \in \{0,1\}^N$ is a uniformly random vector[1]. Our goal is to look at the certificates generated by these algorithms, and argue that they will still be produced with high probability when we replace the truly random vector $\boldsymbol{f}^{\text{unif}}$ with the *pseudorandom* vector $\boldsymbol{f}^{\text{pseud}}$.

The task of certifying $\boldsymbol{f}^{\text{unif}} \notin \text{range}(G)$ for a worst case NC^0_t generator G is a special case of *semirandom CSP refutation*. In particular, given an NC^0_t generator $G : \{0,1\}^S \to \{0,1\}^N$ and any $f \in \{0,1\}^N$, we may define a t-CSP instance with S variables and N constraints, which is satisfiable if and only if $f \in \text{range}(G)$: for each $x \in [N]$ with G_x depending on input cells $i^1_x, \ldots, i^t_x \in [S]$, we add the constraint

$$G_x(E(i^1_x), \ldots, E(i^t_x)) = f_x$$

where $E(1), \ldots, E(S)$ are boolean-valued variables. If G is fixed and $\boldsymbol{f}^{\text{unif}}$ is uniformly random, we end up with a random distribution of t-CSP instances, where the *left hand side* of each constraint is fixed (worst-case), while the *right hand side* is sampled uniformly at random (average case). If we can give an efficient *refutation algorithm* which, with high probability over this distribution, outputs a certificate that the given CSP is *unsatisfiable*, then we can use it to distinguish the generator G: for a uniformly random right hand side the refutation algorithm will typically output "UNSAT," while for $f \in \text{range}(G)$ the algorithm can never output "UNSAT."

This hybrid worst-case/average-case model is known as the *semirandom CSP* model[2]. This model was first introduced in [Fei07] and has been the subject of many followups including [Fei07, AOW15, AGK21, GKM22] (see the theses of Witmer and Manohar [Wit17, Man19] for a more in depth literature review). State of the art semirandom refutation algorithms ([AGK21, GKM22, HKM23]) can certify $\boldsymbol{f}^{\text{unif}} \notin \text{range}(G)$ for an NC^0_t generator $G : \{0,1\}^n \to \{0,1\}^m$ whenever $m \geq \tilde{\omega}(n^{t/2})$. Taking $m = N$, $n = S$ this is precisely the parameter regime $S \leq \tilde{o}(N^{\frac{2}{t}})$ we aim to investigate in our cell probe lower bound for $O(\log N)$-wise independent problems. It is considerably easier to carry out the arguments in [AGK21, GKM22, HKM23] in the case t is even; we will only prove lower bounds for even values of t in this work and hence avoid the technical difficulties that arise for odd t.

In the case t is even, the refutation procedures in [AGK21, GKM22, HKM23] can be divided into roughly three steps:

1. Generate from G a second NC^0_t generator $G' : \{0,1\}^{O(S)} \to \{0,1\}^N$ which is $\mathbb{F}_2$-*linear* (each output of G' XORs together t inputs); reduce the task of certifying $f \notin \text{range}(G)$ for an arbitrary NC^0_t PRG G, to certifying that f is $(\frac{1}{2} - \epsilon)$ far in Hamming distance from $\text{range}(G)$[3] for some small constant $\epsilon > 0$. This is referred to as *strong XOR refutation*.

2. Generate from G' a sequence of real matrices $A_1, \ldots, A_N \in \mathbb{R}^{m \times m}$ for some $m = \text{poly}(N)$; reduce the the task of certifying that f is far from $\text{range}(G')$ (strong XOR refutation) to certifying that the $\infty \to 1$ norm of the matrix

$$A[f] = \sum_{x \in [N]} (2f_x - 1)A_x$$

[1]Some NC^0 cryptanalysis techniques, e.g. Theorem 6 in [MST03], do not yield certificates $f \notin \text{range}(G)$, and only give statistical distinguishers which rely crucially on the fact that the *seed* of G is generated uniformly; such techniques seem to have little relavance to cell probe lower bounds.

[2]The semirandom CSP model is more general than the "random right hand side" variant introduced here, see e.g. [AGK21] for a general definition and comparison to the special case presented here.

[3]We actually need both f and $\neg f$ to be far in Hamming distance from the range.

is small. The reader unfamiliar with the $\infty \to 1$ norm should think of it for now as close relative of the spectral norm which can be approximated in polynomial time (a formal definition is in Section 2.1).

3. Show that the $\infty \to 1$ norm of $A[\boldsymbol{f}^{\text{unif}}]$ is small with high probability where $\boldsymbol{f}^{\text{unif}}$ is uniformly distributed on $\{0,1\}^N$.

To carry out our lower bound, we need to modify the final point, and prove that it continues to hold with high probability when we replace the truly random $\boldsymbol{f}^{\text{unif}}$ with our pseudorandom $\boldsymbol{f}^{\text{pseud}}$. This is argued in Section 3.3 and consists of two steps. The first is a reweighting trick from [GKM22], which replaces the $A_1, \ldots, A_N$ with a second sequence $B_1, \ldots, B_N$ so that for any $f \in \{0,1\}^N$, whenever $A[f]$ has a large $\infty \to 1$ norm, $B[f]$ has a large spectral norm ($B[f]$ is defined analogously to $A[f]$). We then need to prove that $B[\boldsymbol{f}^{\text{pseud}}]$ has small spectral norm with high probability. For this we use the fact that, for any matrix D in m dimensions and any even number k, $\|D\|$ is approximated by the trace-moment polynomial $\text{tr}((D^\top D)^{k/2})^{1/k}$ within a factor $m^{1/k}$. Since in our case $B[f]$ lies in $\mathbb{R}^{m \times m}$ with $m \leq \text{poly}(N)$ and $k = O(\log N)$, we have that $\text{tr}((B[f]^\top B[f])^{k/2})^{1/k}$ approximates the spectral norm of $B[f]$ to within a constant factor. Since $\text{tr}((B[\cdot]^\top B[\cdot])^{k/2})^{1/k}$ is (the k^{th} root of) a degree k polynomial, we can argue that its distribution with respect to $\boldsymbol{f}^{\text{pseud}}$ and $\boldsymbol{f}^{\text{unif}}$ will be similar which will complete the proof.

There are several technical difficulties that we are leaving out of our discussion. First, to make our lower bound work in the setting of a general locally computable generator G (with larger word length and adaptive decision trees at its outputs) we need a more involved argument in place of (1) above. Second, we need to exercise a bit more care in our analysis of the value of the trace moment polynomial $\text{tr}((B[f]^\top B[f])^{k/2})^{1/k}$ to achieve the same bounds in the case that our distribution on f is only $N^{-O(1)}$-wise $O(\log N)$-wise independent, rather than perfectly $O(\log N)$-wise independent. Finally, to deal with the case $k >> \log N$ we need to generalize (2) above, using another construction from semirandom CSP refutation called the "Kikuchi Matrix" [WEAM19, GKM22].

We will not sketch in any detail our proof of Theorem 2 which gives a lower bound $N^{\frac{t}{2}+\Omega(1)}$ for odd constants t in the nonadaptive bit probe model, however we note that it follows the same overall form: we take a semirandom refutation procedure (in this case, the refutations in Section 9 of [GKM22] based on earlier work of [FKO06]), and show that we can still find the associated certificates in the pseudorandom case, for some appropriate notion of pseudorandomness. In this case we rely instead on small bias (fooling parity functions) as our pseudorandom property in place of k-wise independence. As mentioned earlier in the introduction, it is provably necessary to move beyond k-wise independence to achieve a lower bound of this form.

1.3 Open Problems

We present here three potentially tractable open problems we believe are of particular interest. We start with the (seemingly) most accessible problem:

Problem 1. *Extend the bound in Theorem 1 to hold for odd values of t, and the bound in Theorem 2 to hold for even values of t.*

The natural approach here is to utilize the CSP refutation procedures for odd order CSPs developed in [AGK21, GKM22] and analyze them in the "pseudorandom case" as we have done for even order CSPs. It seems likely that a solution to the first part of Problem 1 would automatically solve the second part (see the discussion at the end of Section 4).

Our second problem relates to Theorem 2. While our lower bounds for k-wise independent data structure problems are known to be optimal up to lower order terms (for even t), our nonadaptive

bit probe lower bounds for ϵ-biased distributions have no matching upper bound. We highlight closing this gap as an important open problem:

Problem 2. *For $t \in \mathbb{N}$, determine the largest constant $\delta_t > 0$ so that, in the nonadaptive bit probe model, we have the lower bound $S \geq N^{\delta_t - o(1)}$ for any data structure problem $F : [N] \times [M] \to \{0,1\}$ which supports an $\exp(-N^{o(1)})$-biased distribution on its columns.*

Our Theorem 2 yields the lower bound $\delta_t \geq \left(\frac{t}{2} - \frac{t-2}{2(t+2)}\right)^{-1}$, in particular $\delta_t > \frac{2}{t}$, for every odd value $t \geq 3$. Theorem 1 (the second form described above) also implies $\delta_t \geq \frac{2}{t}$ for even values of t since ϵ-almost k-wise independence is a special case of ϵ-biasedness; this latter lower bound holds in the more general adaptive setting and with larger word lengths. Both of these lower bounds hold even if we only aim to achieve bias $\mathrm{poly}(N)^{-1}$ (rather than $\exp(-N^{o(1)})$ as stated above). On the other hand a construction of [MST03] implies that $\lim_{t\to\infty} \delta_t = 0$, in particular $\delta_t \leq O(\sqrt{\frac{1}{t}})$. Closing this gap in either direction remains an intriguing open problem.

Finally, we highlight two closely related problems: removing the NP oracle in Theorem 3 and removing the nonuniformity in Corollary 1:

Problem 3. *For NC^0_t generators $G : \{0,1\}^n \to \{0,1\}^m$:*

(a) for any $t \geq 3, \epsilon > 0$, give a subexponential time deterministic algorithm that solves range avoidance for G whenever $m \geq n^{\frac{t}{2}-\epsilon}$

(b) for any $t \geq 5, \epsilon > 0$, give a subexponential time uniform algorithm that distinguishes G with subexponential advantage whenever $m \geq n^{\frac{t}{2}-\epsilon}$

One approach to this problem is to prove an *algorithmically constructive* variant of the hypergraph Moore bound of [GKM22, HKM23]. Such an algorithm would also improve the best known polynomial time refutation algorithm for random t-CSPs, and has been an open problem (in the case $t = 3$) since the original work of [FKO06]; there appears to be some significant possibility that no such algorithm exists. A different approach, tailored specifically to (b), is to show that the lower bound in Theorem 2 can be made *natural* in the sense of Razborov-Rudich. More specifically, it would actually be sufficient to develop an efficient algorithm with the following behavior: for some absolute constant $c \in \mathbb{N}$, the algorithm takes as input a uniformly random set $A \subseteq \mathbb{F}_2^n$, $|A| = n^c$, and with nonnegligble probability must output a certificate that A is a $\frac{1}{4}$-biased probability space. See Section 5 for more details.

1.4 Prior Work on Data Structure Lower Bounds

We give here a slightly more involved historical account of lower bound techniques that can prove the strongest lower bounds in the static cell probe model for some explicit problem[4]. As mentioned earlier, the best known lower bound for static data structures prior to our work was $S \geq N^{\frac{1}{t}}(\log M)^{1-\frac{1}{t}}$, first achieved by [Sie89] for k-wise independent problems ($M \approx N^k$). Lower bounds of the form $S \gtrsim N^{\frac{1}{t}}$ can be established by a direct reduction to deterministic two party communication complexity. A more precise communication complexity analysis (keeping track of the number of rounds and the number of bits sent by each party in each round) developed in [Mil94, MNSW98] gives a similar but more precise lower bound in various settings. To achieve a lower bound which includes the extra factor $\approx \log M$ requires techniques more specially adapted

[4] A great deal of work in the field is dedicated to adapting these techniques to prove the best possible lower bounds for particular *natural* problems of interest, which we will not cover here.

to the cell probe model. This can be achieved either by expansion-based arguments known as *cell sampling* ([Sie89, Sie04, PTW10, Lar12]), or in some cases by a more involved communication complexity argument using direct sum theorems [PT06]. In recent work of [GGS23], a lower bound slightly improving the above in the case of *nonadaptive bit probe* data structures is proven; in this restricted setting they obtain a lower bound of the form $\approx N^{\frac{1}{t-1}}(\log M)$. Their method is again cell sampling, and they prove that the given lower bound is essentially the best that this method can achieve. The results in [GGS23] can be applied to larger word lengths provided we transition to data structure problems with nonboolean output, i.e. $F : [N] \times [M] \to Z$ for some $|Z| > 2$.

Observe that, in the setting where M is not much larger than N, e.g. $M \leq \exp(\text{poly} \log N)$, the factor $(\log M)^{1-\frac{1}{t}}$ has negligible effect on the bound $N^{\frac{1}{t}}(\log M)^{1-\frac{1}{t}}$. In this setting, no previously known method (even cell sampling) gives a lower bound significantly stronger than the direct communication reduction, which yields $S \gtrsim N^{\frac{1}{t}}$. We explain this reduction here since it is so simple. Say F has a time t space S data structure with word length w. Consider a protocol in which Alice holds x, Bob holds y; Bob constructs the data structure encoding $\mathcal{E}(y)$, Alice simulates the query procedure on x, requesting cells of $\mathcal{E}(y)$ from Bob who responds with their contents. Alice sends $\log S$ bits in each round to request a cell, Bob sends w bits to reveal the cell's contents, and communication lasts t rounds, so the overall communication is $t \log S + tw$. For any problem F with maximal two party communication complexity, e.g. $F : [N] \times [N] \to \{0, 1\}$ given by $F(x, y) = \mathbb{1}\{x = y\}$, we then have $t \log S + tw \geq \log N$, i.e. $S \geq N^{\frac{1}{t}} \cdot 2^{-w}$.

Finally it is important to keep in mind that for *non-explicit* problems, counting arguments show that most functions $F : [N] \times [N] \to \{0, 1\}$ require space very close to N or time very close to $\log N$ [Mil93]. The situation here is similar to many other areas in complexity; a nonconstructive counting argument implies that random functions exhibit an extremely high lower bound, and the heart of the subject is to prove similar bounds for explicit (and ideally natural) problems.

2 Preliminaries

We introduce some basic notation and definitions which will be used throughout. We strongly encourage the reader not to skip Section 2.2 which defines our notation for locally computable generators and cell probe data structures, and establishes the connection between them which will be used in the rest of the paper.

2.1 Basics

For a natural number N use $[N] = \{1, \ldots, N\}$. For a finite set V and integer k we use $\binom{V}{k}$ to denote the set of subsets of V of size k. If X is a finite set, $\phi : X \to \mathbb{R}$ we write $\mathbb{E}_{x \in X} \phi(x)$ as shorthand for $|X|^{-1} \sum_{x \in X} \phi(x)$. We will use boldface variables e.g. $\boldsymbol{v}$ exclusively for random variables. When $\boldsymbol{v}$ is a random variable and A is a set, we write $\boldsymbol{v} \in A$ to indicate that the support of $\boldsymbol{v}$ is contained in A. When an expression references only a single random variable $\boldsymbol{v}$ we will use $\mathbb{E}\, \phi(\boldsymbol{v})$ for some $\phi : \text{supp}(\boldsymbol{v}) \to \mathbb{R}$ to mean that the expectation is taken with respect to $\boldsymbol{v}$; this is in contrast to the notation $\mathbb{E}_{x \sim X} \phi(x)$ above, where we always include the subscript. At one point we will use the set theoretic operator Δ denoting symmetric difference $a \Delta b = (a \setminus b) \cup (b \setminus a)$. For a collection of sets $(a_j)_{j \in J}$ we define $\bigtriangleup_{j \in J} a_j$ in the natural way using the commutativity/associativity of Δ as a binary operation.

In Section 3.2 we will deal with various vector and matrix norms, whose (standard) definitions we review here:

Definition 1 (Vector and Matrix Norms)**.** *For a vector $u \in \mathbb{R}^m$ we use $\|u\|$ to denote the euclidean (L_2) norm of u, and $\|u\|_\infty$ to denote its L_∞ norm. For a matrix $A \in \mathbb{R}^{n\times n}$, we will use $\|A\|$ to denote the spectral norm given by*

$$\|A\| = \max_{\|u\|,\|v\|\leq 1} |u^\top A v|$$

and use $\|A\|_{\infty\to 1}$ to denote the $\infty \to 1$ operator norm given by

$$\|A\|_{\infty\to 1} = \max_{\|u\|_\infty,\|v\|_\infty\leq 1} |u^\top A v| = \max_{u,v\in\{\pm 1\}^m} |u^\top A v|$$

We now define the notions of limited independence and small-bias probability spaces which we will use extensively. In our main lower bounds we will work over the signed hypercube $\{\pm 1\}^N$, and we present the following definitions in this setting; each definition has an equivalent form over $\{0,1\}^N$ via the natural correspondence $\{0,1\} \leftrightarrow \{\pm 1\}$.

Definition 2. *Let $\boldsymbol{f} \in \{\pm 1\}^N$ be a random variable, $\gamma \in (0,1)$, $k \leq N$. We say that $\boldsymbol{f}$ is γ-almost k-wise independent if, for all $X \subseteq [N]$, $|X| = k$ and all $g \in \{\pm 1\}^X$ we have*

$$\Pr[\boldsymbol{f}|_X = g] \in [2^{-k} - \gamma, 2^{-k} + \gamma]$$

We say that $\boldsymbol{f}$ is k-wise independent if it is 0-almost k-wise independent.

In other words, a distribution over $\{\pm 1\}^N$ is γ-almost k-wise independent if its projection to any k coordinates is γ-close to the uniform distribution on $\{\pm 1\}^k$ in L_∞ distance; the statement applies without change in the $\{0,1\}^N$ basis. Turning our attention to small bias distributions:

Definition 3 ([NN90])**.** *Let $\boldsymbol{f} \in \{\pm 1\}^N$ be a random variable, $\gamma \in (0,1), k \leq N$. We say that $\boldsymbol{f}$ (γ, k)-biased if, for all $X \subseteq [N]$, $|X| \leq k$ we have*

$$\mathbb{E} \prod_{x\in X} \boldsymbol{f}_x \in [-\gamma, \gamma]$$

We say that $\boldsymbol{f}$ is γ-biased if it is (γ, N)-biased.

In the $\{0,1\}$ basis, the property of being γ-biased has the following interpretation: for any *parity function* $\bigoplus_X : \{0,1\}^N \to \{0,1\}$ given by $\bigoplus_X(f) = \sum_{x\in X} f_x \mod 2$ for some $\emptyset \neq X \subseteq [N]$, we have $|\Pr[\bigoplus_X(\boldsymbol{f}) = b] - \frac{1}{2}| \leq \frac{\gamma}{2}$ for each $b \in \{0,1\}$. In other words a γ-biased distribution over $\{0,1\}^N$ fools all parity functions to within error $\frac{\gamma}{2}$. We will use the following well-known relationship:

Lemma 1 ([NN90])**.** *If $\boldsymbol{f}$ is (γ, k)-biased than it is γ-almost k-wise independent.*

Finally, we rely on the following construction of explicit small-bias probability spaces:

Theorem ([NN90])**.** *For every γ, k, N there exists an explicit distribution on $\{\pm 1\}^N$ with support size $(\frac{k \log N}{\gamma})^{O(1)}$.*

We elaborate on our notion of explicit: for the purposes of our range avoidance algorithms in Section 5, we can take explicit to mean, there is a uniform algorithm which, given k, N, ϵ will run in time $\mathrm{poly}(N, k, \frac{1}{\epsilon})$ and output a list of strings in $\{\pm 1\}^N$, such that the uniform distribution on these strings is (γ, k)-biased. The distributions in [NN90] are also explicit in the more standard/informal sense, being supplied by a direct construction.

2.2 Generators and Data Structures

We now define the two main objects of study in our work: the cell probe model of static data structures on the one hand, and locally-computable generators on the other. We then state the tight connection between them, giving some more detail than was supplied in the introduction. Broader implications of this connection will be explored in more depth in Section 6.

Definition 4 (Cell Probe Model)**.** *Let $F : [N] \times [M] \to \{0,1\}$, which we refer to as a "data structure problem." Elements $x \in [N]$ are referred to as "queries," $y \in [M]$ as "datapoints." We say that F has cell probe data structures with space complexity S, time complexity t, and word length w if there exist procedures of the following form:*

1. *(Encoding:) A computationally-unbounded encoding procedure which, given $y \in [M]$, preprocesses it into a data structure $\mathcal{E}(y) \in \Sigma^S$, $|\Sigma| \leq 2^w$*
2. *(Query Procedure:) A computationally-unbounded t-query algorithm $\mathcal{Q}$ which, given $x \in [N]$ and some encoding $E \in \Sigma^S$, queries t indices ("cells") of E and outputs a bit $\mathcal{Q}(x, E) \in \{0,1\}$*

The data structure correctly solves the problem F if, for all $(x, y) \in [N] \times [M]$, $\mathcal{Q}(x, \mathcal{E}(y)) = F(x, y)$. We say the data structure is nonadaptive *if the query procedure decides ahead of time, as a function only of x, on the set of t cells which it will query.*

As mentioned in the introduction, for a problem $F : [N] \times [M] \to \{0,1\}$, we will interchangeably view it as a matrix $F \in \{0,1\}^{N \times M}$ and refer to its set of *columns*; the columns of F consist of the vectors of the form $F_y = (F(x, y))_{x \in [N]} \in \{0,1\}^N$ for each $y \in [M]$.

Next we turn to the formal definition of locally-computable generators:

Definition 5 (Locally-Computable Generators)**.** *We say that $G : \Sigma^S \to \{0,1\}^N$ is a locally-computable generator with time complexity t and space complexity S, provided that each output $(G_x)_{x \in [N]}$ is computed by a depth t decision tree T_x over the alphabet Σ. The word length of the generator is $\lceil \log |\Sigma| \rceil$. We will refer to strings $E \in \Sigma^S$ as "encodings" which are decoded to $G(E)$ by the generator G. Each index $i \in [S]$ is referred to as a "cell," and for an encoding E the contents of the i^{th} cell is $E(i)$. If the decision trees $(T_x)_{x \in X}$ are nonadaptive, we refer to G as a non-adaptive generator. In the case $\Sigma = \{0,1\}$ and G is nonadaptive, we also refer to G as an NC^0_t generator.*

We emphasize that in the literature on cryptographic NC^0 generators, the word "local generator" typically refers to the special case of nonadaptive generators. We use the term "locally-computable" to refer to the more general notion based on decision tree depth, and always use the qualifier "nonadaptive" to refer to the special case of nonadaptive decision trees, or else NC^0_t in the nonadaptive case when $\Sigma = \{0,1\}$.

We now arrive at the observation at the heart of our investigations, whose proof is a matter of definitional manipulation:

Observation 1. *Let $F : [N] \times [M] \to \{0,1\}$ be a data structure problem, interpreted also as a matrix $F \in \{0,1\}^{N \times M}$ below. The following are equivalent:*

1. *F has time t, space S, word length w cell probe data structures (resp. nonadaptive)*
2. *There is a (resp. nonadaptive) locally-computable generator $G : \Sigma^S \to \{0,1\}^N$ with $|\Sigma| \leq 2^w$ and time complexity t, such that every column of F lies in* $\mathrm{range}(G)$

The proof is essentially identical to that of the special case discussed in the introduction.

Proof. Say that F has nonadaptive bit probe data structures of space S and time t. Then, to every $y \in [M]$ we may associate an encoding $E_y \in \{0,1\}^S$, so that for every $x \in [N]$ we may determine $F(x,y)$ by probing (resp. nonadaptively probing) t bits of E_y. Considering the y^{th} column of F as a string $f_y \in \{0,1\}^N$, we see that there is a (resp. nonadaptive) locally-computable function $G : \{0,1\}^S \to \{0,1\}^N$ with time complexity t, so that $G(E_y) = f_y$ for every $y \in [M]$. In particular, the x^{th} output of G runs the data structure's query procedure corresponding to the query x. Every step in this argument in fact utilized a direct equivalence between premise and conclusion; reading it in reverse yields a proof for the other direction. □

3 Cell Probe Lower Bounds via CSP-Pseudorandomness

Using our main observation from the previous section (Observation 1), proving cell probe lower bounds for a data structure problem $F : [N] \times [M] \to \{0,1\}$ is equivalent to showing that, for the set $\text{cols}(F) = \{(F(x,y))_{x \in [N]} \mid y \in [M]\} \subseteq \{0,1\}^N$ consisting of the columns of F, there does not exist a locally-computable generator $G : \Sigma^S \to \{0,1\}^N$ whose range contains every vector $f \in \text{cols}(F)$. If we instead interpret a given locally-computable generator $G : \Sigma^S \to \{0,1\}^N$ as a candidate *pseudorandom generator*, a series of works on semirandom CSP refutation [Fei07, AOW15, AGK21, GKM22, HKM23] have developed techniques which can efficiently generate *certificates* $f \notin \text{range}(G)$ for randomly chosen $f \sim \{0,1\}^N$. In this section, we show how to apply these techniques to *pseudorandom* vectors $f \sim \mathcal{F} \subseteq \{0,1\}^N$ with $\mathcal{F}$ being a distribution with suitable *limited independence* properties. This will imply that if $\text{cols}(F)$ supports such a pseudorandom distribution, then we can argue $\text{cols}(F) \not\subseteq \text{range}(G)$ *for any* generator G with appropriate parameters, which will yield a cell probe lower bound.

For the remainder of the section it will be more convenient to work over $\{\pm 1\}$ rather than $\{0,1\}$; hence our data structure problems will be of the form $F : [N] \times [M] \to \{\pm 1\}$ and our generators of the form $G : \Sigma^S \to \{\pm 1\}^N$.

3.1 From Locally-Computable Generators to XOR Schemes

The first step in the semirandom refutation procedures of [Fei07, AGK21, GKM22] is a reduction from refuting arbitrary t-CSPs to refuting max-t-XOR CSPs. We prove a kind of analogue of this argument for locally-computable generators. For any time t, space S locally-computable generator G and any $f \in \{\pm 1\}^N$, we will associate a system of N different t-XOR equations over $\approx S$ variables. This system of equations will have one part with depends only on G, and another which depends only on f. We will argue that, in the case $f \in \text{range}(G)$, the system of equations will be "moderately satisfiable," in the sense that there is some assignment which satisfies signficantly more than half of the equations, or else falsifies significantly more than half. In subsequent sections, we will show that if f is sampled from a distribution with suitable limited independence properties, any such CSP arising from G, f will not be moderately satisfiable in this sense, which will give us our lower bound.

The above discussion leaves out some complications of the argument; for a given G, f we will actually construct a small collection of t-XOR CSPs, and can only guarantee that *one of them* has a large value in the case $f \in \text{range}(G)$. To obtain good parameters in our main reduction (Theorem 7) we also need some additional arguments more specially tailored to the cell probe model which do not seem to have a direct analogue in the CSP-refutation literature.

Definition 6. *We define a "t-XOR scheme" $\mathcal{C} = (X, V, c)$ to consist of the following components:*

1. *A finite "index set" X and a finite "variable set" V*

2. $c : X \to \binom{V}{t}$ *identifies each* $x \in X$ *with a set of variables ("constraint")* $c_x \subseteq V$, $|c_x| = t$.

We say that $\mathcal{C} = (X, V, c)$ *is "indexed by* X*." For a sign-pattern* $f : X \to \{\pm 1\}$ *and assignment* $R \in \{\pm 1\}^V$ *to the underlying variables, we define*

$$\mathrm{Val}(\mathcal{C}, f, R) := \left| \mathop{\mathbb{E}}_{x \in X} f_x \prod_{i \in c_x} R(i) \right|$$

In some cases, we will have a ground set $[N]$*, and a subset* $X \subseteq [N]$ *with an XOR scheme* $\mathcal{C}$ *indexed by* X*; in this case, for* $f \in \{\pm 1\}^N$ *and an assignment* R*, we define* $\mathrm{Val}(\mathcal{C}, f, R)$ *to equal* $\mathrm{Val}(\mathcal{C}, f', R)$*, where* f' *is the restriction of* f *to the coordinates in* X*.*

Finally, for an XOR scheme $\mathcal{C}$ *and sign pattern* f*, we define*

$$\mathrm{OPT}(\mathcal{C}, f) := \max_{R \in \{\pm 1\}^V} \mathrm{Val}(\mathcal{C}, f, R)$$

As foreshadowed in the discussion preceding this definition, the name "XOR scheme" stems from an interpretation in terms MAX-t-XOR CSPs. For a scheme $\mathcal{C} = ([N], V, c)$ and some $f \in \{\pm 1\}^N$, we may think of the *pair* $(\mathcal{C}, f)$ as representing a t-XOR CSP in the following way. Under the natural identification $(\{\pm 1\}, \times) \leftrightarrow (\mathbb{F}_2, +)$, consider the system of N linear equations over $\mathbb{F}_2$-valued variables $\{v_i \mid i \in V\}$, given by

$$(\sum_{i \in c_x} v_i = f_x \mod 2) \quad \text{for each } x \in [N]$$

We then see that $\mathrm{OPT}(\mathcal{C}, f)$ is the maximum, over all assignments of the variables to values in $\mathbb{F}_2$, of the difference between the fraction of satisfied and unsatisfied equations. In particular, if some assignment can satisfy (resp. falsify) every equation the value is 1; if the value is small, then every assignment will have close to the same number of equations being satisfied/unsatisfied. If we fix $\mathcal{C}$, then we are fixing the left-hand sides of the above system of equations, and as we range over f, we range over different possible right hand sides which may cause the system to be more or less satisfiable. This is the source of the word *scheme*: a single scheme $\mathcal{C}$ defines a family of 2^N different t-XOR CSP instances as we range over the 2^N possible values of the "right hand side" $f \in \{\pm 1\}^N$.

Returning to the $\{\pm 1\}$ viewpoint, observe that we may extend the definition of $\mathrm{Val}(\mathcal{C}, f, R)$ to any $R \in \mathbb{R}^V$. In this view, $\mathrm{Val}(\mathcal{C}, f)$ defines a multilinear polynomial in $|V|$ many real variables, such that the absolute value of the polynomial on any assignment $R \in \mathbb{R}^V$ is equal to $\mathrm{Val}(\mathcal{C}, f, R)$. This leads to the following;

Observation 2. *For any* $\mathcal{C}, f$*,*

$$\max_{R \in \{-1,1\}^V} \mathrm{Val}(\mathcal{C}, f, R) = \max_{R \in [-1,1]^V} \mathrm{Val}(\mathcal{C}, f, R) \geq \max_{R \in \{0,1\}^V} \mathrm{Val}(\mathcal{C}, f, R)$$

Proof. The second inequality is trivial ($\{0, 1\} \subseteq [-1, 1]$). The first is a standard consequence of multilinearity. To confirm it, we may start with any assignment $R \in [-1, 1]^V$ achieving a value $\delta \in [0, 1]$, and replace it by an assignment in $\{\pm 1\}^V$ achieving value $\geq \delta$ by iterating over $i \in V$ and greedily replacing each $R(i)$ by an element in $\{\pm 1\}$ which does not decrease the value. At a given step, we have

$$\mathrm{Val}(\mathcal{C}, f, R) = \Big| R(i) \cdot P((R(j))_{j \neq i}) + Q((R(j))_{j \neq i}) \Big|$$

for some polynomials P, Q depending only on $(R(j))_{j \neq i}$. This is the absolute value of a linear function in $R(i)$ whose maximum is achieved at the boundary of $[-1, 1]$. □

The right-hand side in Observation 2 may be seen as a kind of hypergraph discrepancy parameter. Viewing $(c_x)_{x\in X}$ as a t-uniform (multi-)hypergraph over vertex set V with hyperedges indexed by X, we may think of $f \in \{\pm 1\}^X$ as assigning a sign to each hyperedge. Then $\max_{R\in\{0,1\}^V} \mathrm{Val}(\mathcal{C}, f, R)$ is the maximum, over all subsets of vertices $U \subseteq V$, of the sum of the values f_x with x ranging over hyperedges in the sub-hypergraph induced by U. The smallness of $\max_{R\in\{0,1\}^V} \mathrm{Val}(\mathcal{C}, f, R)$ indicates that every large induced subhypergraph has its edges colored by f in an approximately balanced way.

We now show that, for any locally-computable generator over output space $\{\pm 1\}^N$, we can associate a small collection of XOR schemes indexed by $[N]$ (or by large subsets $X \subseteq [N]$), so that for any $f \in \mathrm{range}(G)$, we have that $\mathrm{OPT}(\mathcal{C}, f)$ is large for one of the schemes $\mathcal{C}$ in our collection.

Theorem 7. *Let $G : \Sigma^S \to \{\pm 1\}^N$ be a locally-computable generator with time complexity t, word length $w = \log|\Sigma|$. There exists a collection of sets $(X_j \subseteq [N])_{j\le 2^{tw+1}}$, $|X_j| \ge 2^{-tw-1}N$, and a collection of t-XOR schemes $(\mathcal{C}_j)_{j\le 2^{tw+1}}$, with $\mathcal{C}_j$ a t-XOR scheme indexed by X_j and having $\le tS$ variables, so that for every $f \in \mathrm{range}(G)$ there exists $j \le 2^{tw+1}$ with $\mathrm{Val}(\mathcal{C}_j, f) \ge 2^{-tw-1}$. In particular, if $\boldsymbol{f} \in \{\pm 1\}^N$ is a random variable supported on $\mathrm{range}(G)$, then there exists an XOR scheme $\mathcal{C}$ over some $X \subseteq [N]$, $|X| \ge 2^{-tw-1}N$, so that $\Pr[\mathrm{OPT}(\mathcal{C}, \boldsymbol{f}) \ge 2^{-tw-1}] \ge 2^{-tw-1}$.*

Proof. By definition, for every $x \in [N]$, $G_x : \Sigma^S \to \{\pm 1\}$ is computed by a decision tree T_x of depth t over the alphabet Σ. We develop here some more detailed notation to describe the trees $(T_x)_{x\in[N]}$. Let $\mathcal{L}(T_x)$ denote the leaves of T_x and $\mathcal{I}(T_x)$ denote the internal nodes. Each internal node $v \in \mathcal{I}(T_x)$ is associated with a cell $p_x^v \in [S]$ which it probes, i.e. if T_x reaches the internal node v at some point in its computation on the input $E \in \Sigma^S$, then it will query $E(p_x^v)$ next and then take a corresponding step down the tree T_x. Each leaf $\ell \in \mathcal{L}(T_x)$ corresponds uniquely to an element $\pi \in \Sigma^t$, where π is the label of the unique path from the root to ℓ: the first cell queried by T_x on input $E \in \Sigma^S$ has value π_1, the second has value π_2, and so on. Thus overall, $T_x(E)$ will follow the unique path π that is consistent with E to the leaf ℓ. Finally, each $\pi \in \Sigma^t$ is labeled by a value $\phi_x(\pi) \in \{\pm 1\}$, so that if T_x reaches the leaf $\ell \in \mathcal{L}(T_x)$ associated with π on input $E \in \Sigma^S$, then $G_x(E) = \phi_x(\pi)$. For a given value $\pi \in \Sigma^t$, define the sequence $q_x^{(\pi,1)}, \dots, q_x^{(\pi,t)} \in [S]$ as follows: let $v_1, \dots, v_t \in \mathcal{I}(T_x)$ be the internal nodes found on the root to leaf path in T_x with leaf corresponding to π, and set $q_x^{(\pi,j)} = p_x^{v_j}$.

Using the structure of the generator G we now define the collection of XOR schemes guaranteed in the Theorem statement. Let $V^1, \dots, V^t$ be sets of size S, each associated canonically with $[S]$, and $V = V^1 \cup \cdots \cup V^t$. For each $x \in [N]$, $\pi \in \Sigma^t$, let $c_x^\pi = \{q^{(\pi,1)}, \dots, q_x^{(\pi,t)}\}$, where $q_x^{(\pi,j)}$ is considered to live inside V^j. For each $\pi \in \Sigma^t$ and each $b \in \{\pm 1\}$ let $X^{(\pi,b)} = \{x \mid \phi_x(\pi) = b\}$. For each π, b, we define the XOR scheme $\mathcal{C}^{(\pi,b)} = (X^{(\pi,b)}, V, c^\pi)$; we will ignore any $\mathcal{C}^{(\pi,b)}$ with $|X^{(\pi,b)}| < 2^{-tw-1}|X|$. We emphasize that the variables V in our XOR schemes are $\{\pm 1\}$-valued (as per the definition of XOR schemes) rather than Σ-valued, despite being derived from some correspondence to the domain of G, whose cells are Σ-valued.

Now, let $f \in \mathrm{range}(G)$, and choose some canonical preimage E_f with $G(E_f) = f$. Then there exists $\pi^f \in \Sigma^t$ so that, for the set $\tilde{X}^{(\pi^f,f)}$ consisting of all $x \in [N]$ such that $T_x(E_f)$ reaches the leaf corresponding to π^f, we have $|\tilde{X}^{(\pi^f,f)}| \ge 2^{-tw}N$. There must then exist a value $b^f \in \{\pm 1\}$ such that $|\tilde{X}^{(\pi^f,f)} \cap X^{(\pi^f,b^f)}| \ge 2^{-tw-1}N$; clearly $|X^{(\pi^f,b^f)}| \ge 2^{-tw-1}N$ as well.

At this point, we have

$$\sum_{x\in X^{(\pi^f,b^f)}} \prod_{j\le t} \mathbb{1}\{E_f(q_x^{(\pi,j)}) = \pi_j^f\} = |\tilde{X}^{(\pi^f,f)} \cap X^{(\pi^f,b^f)}| \ge 2^{-tw-1}N \ge 2^{-tw-1}|X^{(\pi^f,b^f)}|$$

Now define the boolean assignment $R_f \in \{0,1\}^V$, where for $i \in V^j \subseteq V$, we set $R_f(i) = \mathbb{1}\{E_f(i) = \pi_j\}$. In this new notation we have $\prod_{i \in c_x^{(\pi^f)}} R_f(i) = \prod_{j \leq t} \mathbb{1}\{E_f(q_x^{(\pi,j)}) = \pi_j\}$ for every x. Recall also that $G_x(E_f) = b^f$, and hence $f_x = b^f$, whenever $\prod_{j \leq t} \mathbb{1}\{E_f(q_x^{(\pi,j)}) = \pi_j\} = 1$, $x \in X^{(\pi^f, b^f)}$. So overall we conclude

$$\begin{aligned}\mathrm{Val}(\mathcal{C}^{(\pi^f,b^f)}, f, R_f) =& |X^{(\pi^f,b^f)}|^{-1} \cdot \Big| \sum_{x \in X^{(\pi^f,b^f)}} f_x \prod_{i \in c_x^{(\pi^f)}} R_f(i) \Big| \\ =& |X^{(\pi^f,b^f)}|^{-1} \cdot \Big| b^f \sum_{x \in X^{(\pi^f,b^f)}} \prod_{j \leq t} \mathbb{1}\{E_f(q_x^{(\pi,j)}) = \pi_j^f\} \Big| \geq 2^{-tw-1}\end{aligned}$$

Using Observation 2, we may replace the boolean assignment $R_f \in \{0,1\}^V$ by a $\{\pm 1\}$-assignment $\tilde{R}_f \in \{\pm 1\}^V$ achieving a value at least as large, so overall we conclude that $\mathrm{OPT}(\mathcal{C}^{(\pi^f,b^f)}, f) \geq 2^{-tw-1}$. This concludes the proof of the initial claim in the theorem. The last sentence of the Theorem ("in particular...") follows from the first part by taking a union bound over $j \leq 2^{tw+1}$. □

We focus our attention on the "in particular" clause at the end of the statement of Theorem 7. For a data structure problem $F : [N] \times [M] \to \{\pm 1\}$, we have reduced the problem of proving cell probe lower bounds for F to the following: find a random vector $\boldsymbol{f} \in \{\pm 1\}^N$ supported on the columns of F, so that for *any* t-XOR scheme $\mathcal{C}$ defined over a large subset of $[N]$, we have that $\mathrm{OPT}(\mathcal{C}, \boldsymbol{f})$ is very small with very high probability. This may be seen as a specialized notion of pseudorandomness; in particular we may define:

Definition 7 (CSP-Pseudorandomness)**.** *Let $\boldsymbol{f} \in \{\pm 1\}^N$ be a random variable. We say that $\boldsymbol{f}$ is "(N, K, δ, ϵ)-CSP-Pseudorandom" if, for any XOR scheme $\mathcal{C}$ indexed by $[N]$ with K variables, we have*

$$\Pr[\mathrm{OPT}(\mathcal{C}, \boldsymbol{f}) \geq \delta] < \epsilon$$

We will not use this extra notation in our formal arguments later on, and will instead refer directly to Theorem 7, but we isolate the definition here for the purpose of discussion. We may think of each $\mathcal{C}$ indexed by $[N]$ with K variables and each δ as defining a "statistical test" $\mathcal{T} : \{\pm 1\}^N \to \{0,1\}$, with $\mathcal{T}(f) = \mathbb{1}\{\mathrm{OPT}(\mathcal{C}, f) < \delta\}$. Provided $K << N$ and δ is not too small, we can easily argue that a *truly random* f will pass the test with overwhelming probability (chernoff bound + union bound over $\{\pm 1\}^K$); our goal is then to define a pseudorandom distribution which is similar to the uniform distribution with respect to passing this class of tests. This is a standard way to define a notion of pseudorandomness in terms of a class of statistical tests which it "fools." Many examples of such distributions have been constructed in the case that the family of tests in question are of very low computational complexity. However, in this case the statistical test defining $\mathcal{T}$ is NP hard to compute in general: it references the optimal value of a CSP defined by f. For this reason, analyzing this class of tests directly and constructing a suitable pseudorandom distribution appears extremely difficult. This is precisely where *efficient CSP refutation algorithms* come in to play: if we can come up with a more tractable test $\mathcal{T}'$, so that $\mathcal{T}'(f) \to \mathcal{T}(f)$ and still $\mathcal{T}'$ happens with high probability for a random f, we may reduce the task at hand to fooling the simpler test $\mathcal{T}'$. Technically speaking the notions of "tractability" needed here vs. in the setting CSP refutation are quite different. For us, it is neither necessary nor sufficient that $\mathcal{T}'$ be computable in polynomial time; rather we would like the proof of the statement "a random f satisfies $\mathcal{T}'(f) = 1$" to hold over some pseudorandom distribution with significantly smaller entropy.

We make a final note on the discrepancy between Definition 7 and Theorem 7: in Theorem 7 we may have to pass to some large subset $X \subseteq [N]$ and define an XOR scheme indexed only by X, while

in Definition 7 we refer only to XOR schemes over the original set of indices $[N]$. This distinction will be essentially irrelevant to us, since we will deal with notions of pseudorandomness for vectors $\boldsymbol{f} \in \{\pm 1\}^N$ which are preserved under passing to subsets of indices $X \subseteq [N]$. Hence, for most of the following we will state results in terms of XOR schemes indexed by $[N]$ and random vectors in $\{\pm 1\}^N$, with the understanding that we may have started with some original $[N']$, $\boldsymbol{f'} \in \{\pm 1\}^{N'}$, passed to $X \subseteq [N']$ of size N, and then identified X with $[N]$ in some canonical way. We will then remember that this "passing to a subset" operation has occured only when we get to our ultimate proof of the lower bound in Theorem 1, where the density of the subset we passed to will have some mild effect on the parameters.

3.2 CSP-Pseudorandomness via Limited Independence

In this section we prove that if $\boldsymbol{f} \in \{\pm 1\}^N$ satisfies suitable *limited independence* properties, then for any t-XOR scheme $\mathcal{C}$ with suitable parameters, $\mathrm{OPT}(\mathcal{C}, \boldsymbol{f})$ will be very small with very high probability. In the terminology of Definition 7, we aim to prove that limited independence implies CSP-pseudorandomness (within a certain regime of parameters).

We accomplish this in the following way: starting with any t-XOR scheme $\mathcal{C}$ indexed by $[N]$, we construct a certain ensemble of matrices indexed by $[N]$, which has the property that whenever we take a *signed sum* of these matrices using f as a signing, we have that the $\infty \to 1$ norm of the resulting matrix is lower bounded by $\mathrm{OPT}(\mathcal{C}, f)$. We then prove that, when $\boldsymbol{f}$ is chosen from a pseudorandom distribution with suitable limited-independence properties, the operator norm of a random matrix sum signed by $\boldsymbol{f}$ will be small with high probability. The proof of this latter claim is delayed to the next subsection (Section 3.3). We start with the definition of a matrix ensemble:

Definition 8. *A matrix ensemble $\mathcal{A} = (A_x \in \mathbb{R}^{m \times m})_{x \in [N]}$ is a collection of matrices of the same dimension; we say that $\mathcal{A}$ lives in dimension m and is indexed by the set $[N]$. For $f \in \{\pm 1\}^N$, we define the matrix*

$$\mathcal{A}[f] = \sum_{x \in [N]} f_x A_x$$

If each matrix A_x is a subpermutation matrix (boolean with at most one nonzero entry per row and column), we say that $\mathcal{A}$ is a permutation ensemble, and we define $d(\mathcal{A}) = m^{-1} \sum_{x,i,j} A_x(i,j)$ to be the "average degree" of the ensemble.

We will construct matrix ensembles from XOR schemes via the "Kikuchi-Matrix Method" pioneered by [WEAM19] and developed further in [GKM22]. We describe first a special case of the construction, which will suffice for our lower bound on k-wise independent problems in the important case $k \approx \log N$. The idea is simple: we create a new "meta variable" for each $t/2$-tuple of original variables, and each original t-ary constraint $c_x = \{i_1, \ldots, i_t\}$ becomes a 2-ary constraint $c'_x = \{(i_1, \ldots, i_{t/2}), (i_{t/2+1} \ldots, i_t)\}$ over the new variables. This reduces things to the case of a 2-XOR scheme, while blowing up the number of variables by an exponent of $t/2$. For 2-ary XOR schemes, the optimization problem associated with $\mathrm{OPT}(\mathcal{C}, f)$ can be interpreted as maximizing the quadratic form of some matrix over the signed hypercube; this optimization problem is precisely captured by the $\infty \to 1$ norm. In the proof below we will cut out the intermediate step of defining a 2-CSP, and construct the relevant matrices directly from the original t-CSP.

Lemma 2. *Let $\mathcal{C} = ([N], V, c)$ be a t-XOR scheme with $|V| = K$ and t even. There exists a matrix ensemble $\mathcal{A} = (A_x)_{x \in [N]}$ so that:*

1. *Each A_x lives in $\mathbb{R}^{m \times m}$ for $m = \binom{K}{t/2}$, and has exactly 1 nonzero entry with value 1*

2. $d(\mathcal{A}) \geq K^{-\frac{t}{2}} \cdot N$

3. *For any* $f \in \{\pm 1\}^N$, *we have* $\|\mathcal{A}[f]\|_{\infty\to 1} \geq d(\mathcal{A})m \cdot \mathrm{OPT}(\mathcal{C}, f)$

Proof. Associate $[m]$ with $\binom{V}{t/2}$, so $m \leq K^{\frac{t}{2}}$. For $x \in [N]$, write its associated constraint in $\mathcal{C}$ as $c_x = \{i_x^1, \ldots, i_x^t\} \subseteq V$ with its variables ordered in some canonical way. We define our matrix ensemble by

$$A_x(a,b) = \mathbb{1}\{a = \{i_x^1, \ldots, i_x^{t/2}\} \text{ and } b = \{i_x^{t/2+1}, \ldots, i_x^t\}\}$$

Observe that each A_x has exactly one nonzero entry with value one, and hence $d(\mathcal{A}) = m^{-1}N \geq K^{-\frac{t}{2}} \cdot N$. Now, let $f \in \{\pm 1\}^N$ be given, and let $R \in \{\pm 1\}^V$ be such that $\mathrm{OPT}(\mathcal{C}, f) = \mathrm{Val}(\mathcal{C}, f, R)$. Consider the vectors $\tilde{R} \in \{\pm 1\}^m$, given by $\tilde{R}(a) = \prod_{i \in a} R(i)$. Then

$$\|\mathcal{A}[f]\|_{\infty\to 1} \geq \tilde{R}^\top \mathcal{A}[f]\tilde{R} = \sum_{a,b \in \binom{V}{t/2}} \tilde{R}(a)\tilde{R}(b)\mathcal{A}[f](a,b) = \sum_{x \in [N]} f_x \sum_{a,b} A_x(a,b) \prod_{i \in a} R(i) \prod_{j \in b} R(j)$$

$$= \sum_{x \in [N]} f_x \prod_{i \in c_x} R(i) = N \cdot \mathrm{Val}(\mathcal{C}, f, R) = N \cdot \mathrm{OPT}(\mathcal{C}, f)$$

In other words $\|\mathcal{A}[f]\|_{\infty\to 1} \geq d(\mathcal{A})m \cdot \mathrm{OPT}(\mathcal{C}, f)$. □

Now, say that we started with a time-t generator $G : \Sigma^S \to \{0,1\}^N$ with $t = O(1)$, $S = o(\frac{N}{\log N})^{\frac{2}{t}}$, and had some random vector $\boldsymbol{f} \in \{\pm 1\}^N$ supported on the range of G which is $O(\log N)$-wise independent. Applying Theorem 7 followed by Lemma 2 above, we would obtain a matrix ensemble $\mathcal{A}$ indexed by some large subset $X \subseteq [N]$, $|X| \geq \Omega(N)$ with the following properties: the ensemble lives in $\mathbb{R}^{m \times m}$ for $m = N^{O(1)}$, has average degree $\omega(\log N) = \omega(\log m)$, and with high probability $\|\mathcal{A}[\boldsymbol{f}]\|_{\infty\to 1} \geq \Omega(N)$. In the following we will argue that for any $\boldsymbol{f}$ which is $O(\log N)$-wise independent, we will have $\|\mathcal{A}[\boldsymbol{f}]\| = o(N)$ for any $\mathcal{A}$ in dimension $m = N^{O(1)}$ and with $d(\mathcal{A}) = \omega(\log N)$; we would then reach a contradiction with our initial assumption $S \leq o(\frac{N}{\log N})^{\frac{2}{t}}$, which would give us the desired cell probe lower bound $S \geq \tilde{\Omega}(N^{\frac{2}{t}})$ for $O(\log N)$-wise independent problems.

In the case $k >> \log N$ we need a generalization of the construction in Lemma 2 from [WEAM19, GKM22]. We will reproduce the proof (in the appendix) since our notation is somewhat different. The parameter "L" occurring in this Theorem is called the "Kikuchi level" of the construction, and we will set it to be $\approx k$ in our subsequent arguments when proving lower bounds for k-wise independent problems for larger values of k:

Theorem 8 ([WEAM19, GKM22])**.** *Let* $\mathcal{C} = ([N], V, c)$ *be a* t*-XOR scheme,* $|V| = K$*, and* t *even. For any* $t \leq L \leq \frac{K}{8}$*, we may associate with* $\mathcal{C}$ *a permutation ensemble* $\mathcal{A}$ *indexed by* $[N]$ *with the following properties:*

1. $\mathcal{A}$ *lives in* $\mathbb{R}^{m \times m}$ *for* $m = \binom{K}{L}$

2. $d(\mathcal{A}) \geq \left(\frac{L}{K}\right)^{t/2} N$

3. *For any* $f \in \{\pm 1\}^N$, *we have* $\|\mathcal{A}[f]\|_{\infty\to 1} \geq d(\mathcal{A})m \cdot \mathrm{OPT}(\mathcal{C}, f)$

The construction is a straightforward generalization of that in Lemma 2: identifying $[m]$ with $\binom{V}{L}$, we set $A_x(a,b) = \mathbb{1}\{c_x = a \Delta b\}$ where Δ denotes symmetric difference. The proof is also essentially identical to that of the special case Lemma 2 and so we relegate it to the appendix

(Section A). From now on we will refer only to the more general Theorem 8 and leave behind the special case Lemma 2.

It remains to show that if $\mathcal{A}$ is a permutation ensemble with sufficiently small dimension and large average degree, then for any random vector $\boldsymbol{f}$ exhibiting suitable *limited independence* properties, we will have $\|\mathcal{A}[\boldsymbol{f}]\|_{\infty\to 1}$ small with high probability. We defer the proof to the next subsection (Section 3.3), but state the bounds here:

Theorem 9. *Let $\mathcal{A}$ be a permutation ensemble in $\mathbb{R}^{m\times m}$ indexed by $[N]$, let $\boldsymbol{f} \in \{\pm 1\}^N$ be a random vector which is γ-almost k-wise independent for k even, and let $\delta > 0$ be sufficiently small. Assume the following relations are satisfied:*

1. *$d(\mathcal{A}) \geq C\frac{1}{\delta^2}\log(\frac{1}{\delta})m^{2/k}(k+\log m)$ for a sufficiently large universal constant C*
2. *$\gamma \leq 2^{-2k-1}\delta^k m^{-1}$*

Then we have the tail bound:

$$\Pr[\|\mathcal{A}[\boldsymbol{f}]\|_{\infty\to 1} \geq \delta md] \leq 2^{-k}$$

Using Theorem 8 we may directly translate this to a statement about XOR schemes:

Theorem 10. *Let $\mathcal{C} = ([N], V, c)$ be a t-XOR scheme, $|V| = K$, t even, $L \in \mathbb{N}$ with $t \leq L \leq \frac{K}{8}$, and say that $\boldsymbol{f} \in \{\pm 1\}^N$ is γ-almost k-wise independent, $k \leq L\log N$. Then for any $\delta > 0$ sufficiently small we have*

$$\Pr[\mathrm{OPT}(\mathcal{C}, \boldsymbol{f}) \geq \delta] \leq 2^{-k}$$

provided that $\frac{N}{\log N} \geq C\frac{1}{\delta^2}\log\frac{1}{\delta}K^{\frac{t}{2}+\frac{2L}{k}}\cdot L^{1-\frac{t}{2}}$ where C is a universal constant, and $\gamma \leq 2^{-2k-1}\delta^k m^{-1}$.

Finally applying the above Theorem together with Theorem 7, we obtain our main cell probe lower bound:

Theorem (Theorem 1 restated)**.** *Let $F : [N]\times[M] \to \{\pm 1\}$ be a data structure problem such that there exists a γ-almost k-wise independent random vector $\boldsymbol{f} \in \{\pm 1\}^N$ supported on the columns of F. Let $S, t, w \in \mathbb{N}$ be given with t an even number and assume $k > tw + 1$. Then, for any space S, time t, word length w cell probe data structure solving F, we must have:*

1. *$S \geq \left(\frac{N}{\log N}\right)^{\frac{2}{t}} \cdot \left(\frac{k}{\log N}\right)^{1-\frac{2}{t}} \cdot t^{-1}2^{-(6-o(1))w}$ for all $t\log N \leq k \leq N2^{-(3-o(1))tw}t^{-\frac{t}{2}}$*
2. *$S \geq \left(\frac{N}{\log N}\right)^{\frac{1}{t(\frac{1}{2}+\frac{2}{k})}} \cdot t^{-1}2^{-(6-o(1))w}$ for all $t \leq k \leq t\log N$,*

provided $\gamma \leq N^{-t}2^{-O(ktw)}$ in each case.

Proof. Set $\delta = 2^{-tw-1}$. Say that F has a time t, space S data structure with t even. Fix some $L \in \mathbb{N}$ satisfying $t \leq L \leq \frac{tS}{8}$ and $k \leq L\log N$ to be specified later, and assume that $\gamma \leq (tS)^{-L}2^{-2k-1}\delta^{-k}$. Applying Theorem 7 we may find $X \subseteq [N]$ of size $|X| \geq \delta N$ and an XOR scheme $\mathcal{C} = (X, V, c)$, $|V| = K := tS$ indexed by X so that, with probability at least δ over $\boldsymbol{f}$, $\mathrm{OPT}(\mathcal{C}, \boldsymbol{f}) \geq \delta$. Applying Theorem 10 and using the fact that $2^{-k} < \delta$ (since $k > tw + 1$), this is only possible provided

$$\delta\frac{N}{\log N} \leq \frac{|X|}{\log|X|} < C\frac{1}{\delta^2}\log\frac{1}{\delta}K^{\frac{t}{2}+\frac{2L}{k}}\cdot L^{1-\frac{t}{2}}$$

We have $C\frac{1}{\delta^3}\log\frac{1}{\delta} \le 2^{3tw+o(tw)}$. We also may safely assume $S \le \frac{N}{t}$ and $\log K \le \log N$. Plugging in these estimates and tS in place of K we obtain the bound:

$$S > \Big(\frac{N}{\log N}\Big)^{\frac{1}{\frac{t}{2}+\frac{2L}{k}}} \cdot L^{\frac{\frac{t}{2}-1}{\frac{t}{2}+\frac{2L}{k}}} \cdot t^{-1}2^{-(6-o(1))w} \tag{2}$$

To finish the proof, we set $L = \frac{k}{\log N}$ in case (1) and $L = t$ in case (2); in both cases the condition $t \le L \le \frac{tS}{8}$ will be satisfied, and $\gamma \le (tS)^{-L}2^{-2k-1}\delta^{-k}$ will hold provided $\gamma \le N^{-t}2^{-Ckw}$ for a sufficiently large constant C.

□

3.3 Proof of Theorem 9

We now prove Theorem 9 which is reproduced below for convenience:

Theorem (Theorem 9, restated)**.** *Let $\mathcal{A}$ be a permutation ensemble in $\mathbb{R}^{m\times m}$ indexed by $[N]$, let $\boldsymbol{f} \in \{\pm 1\}^N$ be a random vector which is γ-almost k-wise independent for k even, and let $\delta > 0$ be sufficiently small. Assume the following relations are satisfied:*

1. *$d(\mathcal{A}) \ge C\frac{1}{\delta^2}\log(\frac{1}{\delta})m^{2/k}(k+\log m)$ for a sufficiently large universal constant C*
2. *$\gamma \le 2^{-2k-1}\delta^k m^{-1}$*

Then we have the tail bound:

$$\Pr[\|\mathcal{A}[\boldsymbol{f}]\|_{\infty\to 1} \ge \delta md] \le 2^{-k}$$

The first half of our proof will follow a strategy of [HKM23]: we will pass from $\mathcal{A}$ to a second matrix ensemble $\mathcal{B}$ over the same index set $[N]$ so that whenever $\|\mathcal{A}[f]\|_{\infty\to 1}$ is large, $\|\mathcal{B}[f]\|$ is large, where $\|\cdot\|$ denotes the spectral norm. We then prove that limited independence fools the spectral norm. For this we use a standard technique, which approximates the spectral norm of a matrix by the trace of a suitable matrix power; the trace of this power will be low degree polynomial in the underlying variables $\{f_x \mid x \in [N]\}$, and hence takes on a similar value when $\boldsymbol{f}$ is chosen truly uniformly, or chosen from distribution with limited independence. We remark that a similar argument is made in passing in the introduction of [dT22] in the context of refuting fully-random CSPs. Finally, it remains to prove that the spectral norm of $\mathcal{B}[\boldsymbol{f}]$ is low when $\boldsymbol{f}$ is truly uniform; for this we rely on matrix concentration inequalities as in various prior works on semirandom refutation [AGK21, GKM22].

We start with some additional definitions and inequalities necessary for the proof:

Definition 9. *For a matrix A and even integer k, define $A^{\bowtie k} := (A^\top A)^{k/2}$. For a matrix ensemble $\mathcal{B} = (B_x)_{x\in[N]}$, let*

$$\nu_{\mathrm{row}}(\mathcal{B}) = \|\sum_x B_x^\top B_x\|, \quad \nu_{\mathrm{col}}(\mathcal{B}) = \|\sum_x B_x B_x^\top\|, \quad \nu(\mathcal{B}) = \max\{\sigma_{\mathrm{row}}(\mathcal{B}), \sigma_{\mathrm{col}}(\mathcal{B})\}$$

and let $\chi(\mathcal{B}) = \max_{x\in[N]}\|B_x\|$.

We need the following well-known fact:

Lemma 3 (Trace/Norm Inequality)**.** *For $A \in \mathbb{R}^{m\times m}$ and any even k,*

$$m^{-1/k}\cdot \mathrm{tr}(A^{\bowtie k})^{1/k} \le \|A\| \le \mathrm{tr}(A^{\bowtie k})^{1/k}$$

Proof. For the PSD matrix $A^{\bowtie 2}$, we have $\sqrt{\lambda_{\max}(A^{\bowtie 2})} = \|A\|$, and $\lambda_{\max}((A^{\bowtie 2})^{k/2}) = (\lambda_{\max}(A^{\bowtie 2}))^{k/2}$, hence overall we have $\|A\| = \lambda_{\max}((A^{\bowtie 2})^{k/2})^{1/k} = \lambda_{\max}(A^{\bowtie k})^{1/k}$. On the other hand for any PSD matrix $D \in \mathbb{R}^{m\times m}$ we have $m^{-1}\mathrm{tr}(D) \le \lambda_{\max}(D) \le \mathrm{tr}(D)$; apply this to $D = A^{\bowtie k}$ and take k^{th} roots. □

As described above, the first step in our proof is to pass from a permutation ensemble $\mathcal{A}$ to another ensemble $\mathcal{B}$, so that the spectral norm of $\mathcal{B}[f]$ controls the $\infty \to 1$ norm of $\mathcal{A}[f]$. For this we use a strategy laid out in the introduction of [HKM23], which multiplies each $A_x(i,j)$ by some global scaling factors $w_i \cdot w_j$ so as to normalize the total entry weight lying in each row and column.

Lemma 4. *Let $\mathcal{A} = (A_x)_{x\in[N]}$ be a permutation ensemble in $\mathbb{R}^{m\times m}$. Then there exists an ensemble $\mathcal{B} = (B_x)_{x\in[N]}$ of the same dimensions, so that*

1. *For every $f \in \{\pm 1\}^N$, we have $\frac{1}{2md(A)}\|\mathcal{A}[f]\|_{\infty\to 1} \le \|\mathcal{B}[f]\|$*
2. *$\nu(\mathcal{B}), \chi(\mathcal{B}) \le d(\mathcal{A})^{-1}$*
3. *$\mathrm{tr}(B^{\bowtie k}) \le m$ for all even integers k, where $B = \sum_{x\in[N]} B_x$*
4. *$\|\mathcal{B}[f]\| \le 1$ for all $f \in \{\pm 1\}^N$*

Proof. Let $d = d(\mathcal{A})$. For each $i \in [m]$ let $d_{\mathrm{row}}(i) = \sum_{x,j} A_x(i,j)$ and let $\Gamma_i = d + d_{\mathrm{row}}(i)$; we define $d_{\mathrm{col}}(j)$, Λ_j symmetrically w.r.t. rows and columns. Define the ensemble $\mathcal{B} = (B_x)_{x\in[N]}$ by

$$B_x(i,j) = \Gamma_i^{-1/2}\Lambda_j^{-1/2}A_x(i,j)$$

Let $u, v \in [-1,1]^m$ be such that $\|\mathcal{A}[f]\|_{\infty\to 1} = |u^\top \mathcal{A}[f]v|$. Define

$$\hat{u} = (u(i)\cdot\Gamma_i^{1/2})_{i\in[m]}, \quad \hat{v} = (v(j)\cdot\Lambda_j^{1/2})_{j\in[m]}$$

Then we have $|\hat{u}^\top\mathcal{B}[f]\hat{v}| = |u^\top\mathcal{A}[f]v| = \|\mathcal{A}[f]\|_{\infty\to 1}$, and hence

$$\|\mathcal{B}[f]\| \ge \frac{|\tilde{u}^\top\mathcal{B}[f]\tilde{v}|}{\|\hat{u}\|\cdot\|\hat{v}\|} \ge \|\mathcal{A}[f]\|_{\infty\to 1}\cdot\Big(\sum_{i\le m}(d_{\mathrm{row}}(i)+d)\Big)^{-1/2}\Big(\sum_{j\le m}(d_{\mathrm{col}}(j)+d)\Big)^{-1/2} = \frac{\|\mathcal{A}[f]\|_{\infty\to 1}}{2md}$$

which gives (1). For (2), we have $\chi(\mathcal{B}) \le \max_{i,j}|\Gamma_i^{-1/2}\Lambda_j^{-1/2}| \le d^{-1}$ since the spectral norm of each B_x, being a reweighted subpermutation matrix, is bounded above by the magnitude of its largest entry. For $\nu(\mathcal{B})$, note that $B_x^\top B_x$ is a diagonal matrix with $(i,i)^{th}$ entry equal to $\sum_j B_x(i,j)^2$. Hence $\sum_x B_x^\top B_x$ is diagonal as well, with spectral norm equal to the magnitude of its largest entry, in particular:

$$\nu_{\mathrm{row}}(\mathcal{B}) = \max_i \sum_{x,j} B_x(i,j)^2 = \max_i \sum_{x,j}\Gamma_i\Lambda_j A_x(i,j) \tag{3}$$

$$\le \max_i \Gamma_i^{-1}\sum_{x,j}\frac{A_x(i,j)}{d} = \frac{1}{d}\max_i\frac{d_{\mathrm{row}}(i)}{\Gamma_i} \le \frac{1}{d} \tag{4}$$

The same argument applies to columns, and we conclude $\nu(\mathcal{B}) \le d^{-1}$. Next we verify (3). Let $2k$ be an even integer and let $B = \sum_x B_x$, $A = \sum_x A_x$. We argue here exactly as in the introduction

of [HKM23]. Expanding the definition of $\mathrm{tr}(B^{\bowtie 2k})$:

$$\begin{aligned}\mathrm{tr}(B^{\bowtie 2k}) &= \sum_{i_1}\sum_{j_1} B(i_1,j_1)\sum_{i_2} B(i_2,j_1)\cdots\sum_{j_k} B(i_k,j_k)\sum_{i_{k+1}} B(i_{k+1},j_k)\mathbb{1}\{i_{k+1}=i_1\}\\ &= \sum_{i_1}\sum_{j_1} A(i_1,j_1)\Gamma_{i_1}^{-1/2}\Lambda_{j_1}^{-1/2}\sum_{i_2} A(i_2,j_1)\Gamma_{i_2}^{-1/2}\Lambda_{j_1}^{-1/2}\cdots\sum_{i_{k+1}} A(i_{k+1},j_k)\Gamma_{i_1}^{-1/2}\Lambda_{j_k}^{-1/2}\mathbb{1}\{i_{k+1}=i_1\}\\ &= \sum_{i_1}\Gamma_{i_1}^{-1}\sum_{j_1} A(i_1,j_1)\Lambda_{j_1}^{-1}\sum_{i_2} A(i_2,j_1)\Gamma_{i_2}^{-1}\cdots\Lambda_{j_k}^{-1}\sum_{i_{k+1}} A(i_{k+1},j_k)\mathbb{1}\{i_{k+1}=i_1\}\\ &\le \sum_{i_1} 1 \le m\end{aligned}$$

where the domain of each summand i,j is $[m]$, and in the second to last inequality we are using the fact that for any i we have $\Gamma_i^{-1}\sum_j A(i,j)\le 1$ (and symmetrically for columns, Λ_j). Finally for (4), observe that since $\mathrm{tr}(\mathcal{B}[f]^{\bowtie 2k})$ is a polynomial in the variables $\{f_x \mid x\in[N]\}$ with nonnegative coefficients, (3) implies that $\mathrm{tr}(\mathcal{B}[f]^{\bowtie 2k}) \le m$ for all f and all k. Applying the Trace/Norm inequality (Lemma 3) we have that $\|\mathcal{B}[f]\|^{2k}\le m$ for arbitrarily large k, hence $\|\mathcal{B}[f]\|\le 1$. □

We now wish to argue that, for the kind of matrix ensemble $\mathcal{B}$ obtained in the last lemma, the upper tail of $\|\mathcal{B}[\hat{\boldsymbol{f}}]\|$ is controlled by that of $\|\mathcal{B}[\boldsymbol{f}]\|$, where $\hat{\boldsymbol{f}}$ and $\boldsymbol{f}$ are k-wise independent and truly uniform vectors respectively. This follows directly from the trace norm inequality and the fact that $\mathrm{tr}(\mathcal{B}[\cdot]^{\bowtie k})$ is a degree k polynomial. If we assume only that $\hat{\boldsymbol{f}}$ is γ-almost k-wise independent, for $\gamma>0$, some error term involving γ and the total monomial weight of the associated polynomial will be introduced. In the end we get:

Lemma 5. *Let $\mathcal{B}=(B_x)_{x\in[N]}$ be a matrix ensemble in $\mathbb{R}^{m\times m}$ such that $\|\mathcal{B}[f]\|\le 1$ for all $f\in\{\pm1\}^N$ and let $k>3$ be an even integer. Say that $\boldsymbol{f},\hat{\boldsymbol{f}}\in\{\pm1\}^N$ are random variables, with $\boldsymbol{f}$ uniform on $\{\pm1\}^N$ and $\hat{\boldsymbol{f}}$ sampled from a γ-almost k-wise independent distribution. Then for any $\delta>0$ with $\gamma\cdot\mathrm{tr}((\sum_x B_x)^{\bowtie k})\le 2^{-2k-1}\delta^k$ we have:*

$$\Pr[\|\mathcal{B}[\hat{\boldsymbol{f}}]\|\ge\delta]\ge 2^{-k}\longrightarrow \Pr[\|\mathcal{B}[\boldsymbol{f}]\|\ge\delta(8m)^{-1/k}]\ge 2^{-k-2}\delta^k m^{-1}$$

Proof. Set $\epsilon:=2^{-k}$, $r:=\mathrm{tr}((\sum_x B_x)^{\bowtie k})$. Say that $\Pr[\|\mathcal{B}[\hat{\boldsymbol{f}}]\|\ge\delta]\ge\epsilon$, hence $\Pr[\|\mathcal{B}[\hat{\boldsymbol{f}}]\|^k\ge\delta^k]\ge\epsilon$ and therefore $\mathbb{E}\,\|\mathcal{B}[\hat{\boldsymbol{f}}]\|^k\ge\epsilon\delta^k$. By the trace/norm inequality, we know that for all f, $\mathrm{tr}(\mathcal{B}[f]^{\bowtie k})\ge\|\mathcal{B}[f]\|^k$, hence we have $\mathbb{E}\,\mathrm{tr}(\mathcal{B}[\hat{\boldsymbol{f}}]^{\bowtie k})\ge\epsilon\delta^k$. Since $\mathrm{tr}(B[\cdot]^{\bowtie k})$ is a degree k polynomial with total monomial weight $\le r$, we have that $|\,\mathbb{E}\,\mathrm{tr}(\mathcal{B}[\hat{\boldsymbol{f}}]^{\bowtie k})-\mathbb{E}\,\mathrm{tr}(\mathcal{B}[\boldsymbol{f}]^{\bowtie k})|\le\gamma r 2^k$, hence $\mathbb{E}\,\mathrm{tr}(\mathcal{B}[\boldsymbol{f}]^{\bowtie k})\ge\epsilon\delta^k-\gamma r2^k\ge\frac{1}{2}\epsilon\delta^k$ and therefore, applying the other side of the trace/norm inequality, $\mathbb{E}\|\mathcal{B}[\boldsymbol{f}]\|^k\ge\frac{\epsilon\delta^k}{2m}$. Thus $\Pr[\|\mathcal{B}[\boldsymbol{f}]\|^k\ge\frac{\epsilon\delta^k}{4m}]\ge\frac{\epsilon\delta^k}{4m}$ (we use here that $\|\mathcal{B}[\boldsymbol{f}]\|\le 1$ probability 1), and so $\Pr[\|\mathcal{B}[\boldsymbol{f}]\|>\delta(\frac{\epsilon}{4m})^{1/k}]\ge\frac{\epsilon\delta^k}{4m}$. Recall that $\epsilon=2^{-k}$, so $\delta(\frac{\epsilon}{4m})^{1/k}=\delta(8m)^{1/k}$. □

It then remains only to argue that, for the ensemble $\mathcal{B}$ arising in our argument, $\|\mathcal{B}[\boldsymbol{f}]\|$ has a suitably decaying upper tail when $\boldsymbol{f}$ is a uniformly random signing. For this we rely on the Matrix Bernstein inequality.

Theorem 11 (Matrix Bernstein [Tro12])**.** *Let $\mathcal{B}=(B_x)_{x\in X}$ be an ensemble in $\mathbb{R}^{m\times m}$ and $\boldsymbol{f}\in\{\pm1\}^X$ a uniformly random sign vector. For any $\alpha>0$ we have*

$$\Pr[\|\mathcal{B}[\boldsymbol{f}]\|\ge\alpha]\le 2m\exp\left(-\frac{1}{6}\cdot\frac{\alpha^2}{\nu(\mathcal{B})+\alpha\chi(\mathcal{B})}\right)$$

We are now ready to prove Theorem 9:

Proof of Theorem 9. Let $\mathcal{A} = (A_x)_{x\in[N]}$ be a permutation ensemble in $\mathbb{R}^{m\times m}$ with average degree $d = d(\mathcal{A})$, let $\delta > 0$ be sufficiently small. Let $\boldsymbol{f}, \hat{\boldsymbol{f}} \in \{\pm 1\}^X$ be random vectors, with $\boldsymbol{f}$ uniform and $\hat{\boldsymbol{f}}$ γ-almost k-wise independent for some even k, and assume $\gamma \leq 2^{-2k-1}\delta^k m^{-1}$. Assume that $\Pr[\|\mathcal{A}[\hat{\boldsymbol{f}}]\|_{\infty\to 1} \geq \delta md] > 2^{-k}$; we then wish to establish that $d \leq C\frac{1}{\delta^2}\log(\frac{1}{\delta})m^{2/k}\log m$ for a suitably large constant C. Applying Lemma 4, we construct from $\mathcal{A}$ an ensemble $\mathcal{B}$, so that $\nu(\mathcal{B}), \chi(\mathcal{B}) \leq d^{-1}$, $\mathrm{tr}((\sum_x B_x)^{\bowtie k}) \leq m$, $\|\mathcal{B}[f]\| \leq 1$ for all f, and $\|\mathcal{B}[f]\| \geq (2md)^{-1}\|\mathcal{A}[f]\|_{\infty\to 1}$ for all f. In particular this implies

$$\Pr[\|\mathcal{B}[\hat{\boldsymbol{f}}]\| \geq \frac{\delta}{2}] \geq \Pr[\|\mathcal{A}[\hat{\boldsymbol{f}}]\|_{\infty\to 1} \geq \delta md] > 2^{-k}$$

We then apply Lemma 5 and conclude that:

$$\Pr[\|\mathcal{B}[\boldsymbol{f}]\| \geq \frac{\delta}{2}(8m)^{-1/k}] \geq 2^{-2k-2}\delta^k m^{-1}$$

Now setting $\alpha := \frac{\delta}{2}(8m)^{-1/k}$, from the Bernstein inequality we know that

$$\Pr[\|\mathcal{B}[\boldsymbol{f}]\| \geq \alpha] \leq 2m\exp(-\frac{1}{6}\cdot\frac{\alpha^2}{\frac{1}{d}+\frac{\alpha}{d}}) \leq 2m\exp(-\frac{\alpha^2 d}{12}) \leq 2^{\log m - \frac{\alpha^2 d}{12}}$$

We then have $2^{\log m - \frac{\alpha^2 d}{12}} \geq 2^{-2k-1}\delta^k m^{-1}$, which implies $d \leq C\frac{1}{\delta^2}\log\frac{1}{\delta}m^{2/k}(k+\log m)$ for C a suitable absolute constant and δ sufficiently small. □

3.4 Lower Bounds for Concrete Problems

We may apply the lower bound in Theorem 1 to any problem $F : [N]\times[M] \to \{0,1\}$ exhibiting suitable limited independence properties. If our goal is simply to minimize M as a function of N, subject to proving the best quantitative bound and subject to F being *explicit*, our best course of action is to use any explicit construction of γ-almost k-wise independent distributions with minimum possible support size (i.e. minimum seed length), for example [NN90] or any of its subsequent improvements. However, it is also of interest to prove lower bounds for *natural* problems which have some more concrete interpretation as an information retrieval task. We highlight here two such natural problems to which we can apply the lower bound in Theorem 1. Both pertain to evaluating low degree polynomials of some kind over a finite field.

3.4.1 Evaluating Univariate Polynomials over a large field $\mathbb{F}$

The first problem we discuss is standard in data structure complexity (e.g. [Lar12, GGS23]): the evaluation of low degree univariate polynomials over a large finite field. The k-wise independence properties of this problem are well known and this lower bound does not require much elaboration. We define the polynomial evaluation problem formally below.

Definition 10. *For any finite field $\mathbb{F}$ and $d \in \mathbb{N}$ we define the data structure problem $\mathbb{F}$-Eval^d. Datapoints are univariate polynomials p of degree $\leq d$ over $\mathbb{F}$; queries are elements $x \in \mathbb{F}$; the correct answer is the evaluation of p on x, i.e. $\mathbb{F}$-$\mathsf{Eval}^d(x,p) = p(x)$. For any function $\chi : \mathbb{F} \to \{0,1\}$, define the restricted problem χ-$\mathbb{F}$-Eval^d, which has the same datapoints and queries, but the required answer is only the bit $\chi(p(x))$.*

Note that the problem $\mathbb{F}$-Eval^d does not have a boolean output, however it is clearly at least has hard as the boolean-valued problem χ-$\mathbb{F}$-Eval^d for any χ. A typical choice of χ is as follows: say that $\mathbb{F} = \mathbb{F}_{2^n}$ is identified with $\{0,1\}^n$ under some natural encoding (e.g. $x \in \mathbb{F}$ is interpreted as a degree $\leq n-1$ univariate polynomial over $\mathbb{F}_2$ modulo some fixed irreducible and is coded by its vector of coefficients), and set $\chi(x)$ to be the first bit of x's encoding. We then have, as an immediate consequence of Theorem 1 and the well known d-wise independent properties of degree d polynomials:

Theorem 12. *Let $\mathbb{F}$ be a finite field of characteristic 2 and let $\chi : \mathbb{F} \to \{0,1\}$ be any function which is balanced, i.e. $|\chi^{-1}(0)| = |\chi^{-1}(1)|$. Let $d, S, t, w \in \mathbb{N}$ be given with t an even number and assume $d > tw + 1$. Then, for any space S, time t, word length w cell probe data structure solving χ-$\mathbb{F}$-Eval^d we must have:*

1. $S \geq \left(\frac{|\mathbb{F}|}{\log|\mathbb{F}|}\right)^{\frac{2}{t}} \cdot \left(\frac{d}{\log|\mathbb{F}|}\right)^{1-\frac{2}{t}} \cdot t^{-1}2^{-(6-o(1))w}$ *for all* $t\log|\mathbb{F}| \leq d \leq |\mathbb{F}|2^{-(3-o(1))tw}t^{-\frac{t}{2}}$
2. $S \geq \left(\frac{|\mathbb{F}|}{\log|\mathbb{F}|}\right)^{\frac{1}{t(\frac{1}{2}+\frac{2}{k})}} \cdot t^{-1}2^{-(6-o(1))w}$ *for all* $t \leq d \leq t\log|\mathbb{F}|$,

3.4.2 Evaluating Multivariate Polynomials over $\mathbb{F}_2$

The second and perhaps less known problem we highlight pertains to evaluating low degree multivariate polynomials $p : \mathbb{F}_2^n \to \mathbb{F}_2$, defined formally below:

Definition 11. *For any $d \leq n$ we define the data structure problem $\mathbb{F}_2$-Eval^d_n. Datapoints are multilinear polynomials $p : \mathbb{F}_2^n \to \mathbb{F}_2$ of degree $\leq d$; queries are elements $x \in \mathbb{F}_2^n$; the correct answer is the evaluation of p on x, i.e. $\mathbb{F}_2$-$\mathsf{Eval}^d_n(x,p) = p(x)$.*

We use the following result of [Raz88], improved quantitatively by [Sav95]:

Lemma 6 ([Raz88, Sav95])**.** *Let $k \leq n$. There exists a random function variable $\boldsymbol{f} : \{0,1\}^n \to \{0,1\}$ which is k-wise independent, so that with probability 1 $\boldsymbol{f}$ is computed by a degree $\leq \log k + 1$ polynomial $p : \mathbb{F}_2^n \to \mathbb{F}_2$ (under the natural correspondence $\mathbb{F}_2 \leftrightarrow \{0,1\}$).*

The above statement does not occur explicitly in [Raz88, Sav95]; rather, [Sav95] shows that for every $k, \gamma > 0$, there exists a random function $\boldsymbol{f} : \{0,1\}^n \to \{0,1\}$ which is γ-almost k-wise independent and computed (with probability 1) by a formula of the form:

$$\phi(x) = \bigoplus_{i \leq m} \bigwedge_{j \leq \log k + 1} \bigoplus_{u \leq 2n+2} \ell_{i,j,u} \tag{5}$$

for $m = 2\log(\frac{1}{\gamma})\log k$, where each $\ell_{i,j,u}$ is a literal in $\{x_1, \ldots, x_n, \neg x_1, \ldots, \neg x_n, 0, 1\}$. Clearly such a formula is representable by a polynomial of degree at most $\log k + 1$ regardless of the value of γ; taking γ suitably small (below the granularity of the implied distribution, which is bounded away from zero by a function of only n, k) we obtain the desired result. Combining Lemma 6 and Theorem 1 we arrive at the following cell probe lower bound for $\mathbb{F}_2$-Eval^d_n:

Theorem 13. *Let n, d, t, w be given, and assume that d lies in the range:*

$$\log n + \log t + \log w + O(1) \leq d \leq n - O(wt\log t)$$

Then the problem $\mathbb{F}_2$-Eval^d_n satisfies the space/time/word length tradeoff

$$S \geq \left(2^{n-\log n}\right)^{\frac{2}{t}} \cdot \left(2^{d-\log n}\right)^{1-\frac{2}{t}} \cdot t^{-1}2^{-(6-o(1))w}$$

We make some observations about this lower bound. Consider the setting $w, t = O(1)$ and $d = \log n + O(1)$ and let F be the data structure problem $F = \mathbb{F}_2\text{-}\mathsf{Eval}^d_n$. Then F is of the form $F : [N] \times [M] \to \{0,1\}$ for $N = 2^n$, $M = 2^{n^{\log n + O(1)}}$, i.e. $M = 2^{2^{O(\log\log N)^2}}$ which is close to polynomial in N (we may say that it is quasi-quasi-polynomial). This problem F satisfies the (new as of this paper) strongest known cell probe lower bound for any problem with this relation between M, N: we require space $S \geq \tilde{\Omega}(N^{\frac{2}{t}})$ for time t, word length w data structures. On the other hand, we can observe that F has rank at most $2^{O(\log\log N)^2}$ over $\mathbb{F}_2$, which is very close to *logarithmic* in the number of queries N; this is because the columns of F are spanned by the $n^{\log n + O(1)}$ monomials of degree $\leq d$. In the language of communication complexity, this places F in the communication class $\oplus\mathsf{P}^{\mathsf{cc}}$. This is of some interest in relation to our later discussions in Section 6.3, where we relate cell probe complexity to other "high-end" communication complexity classes such as $\mathsf{PH}^{\mathsf{cc}}$ and $(\mathsf{PH}[\oplus_{\mathsf{p}}])^{\mathsf{cc}}$

Second, we note that we could have defined directly an "evaluation problem" for depth 3 $\mathsf{AC}^0[\oplus]$ formula of the form occuring in (5) above. For an appropriate setting of γ and using our bounds for almost k-wise independence we would obtain a problem $F : \{0,1\}^n \times \{0,1\}^{\mathrm{poly}(n)} \to \{0,1\}$ satisfying the new state of the art time space tradeoff, which in addition has the property that $F(x,y)$ is computable by a $\mathrm{poly}(n)$-size $\mathsf{AC}^0[\oplus]$ formula.

4 Beating the Limited Independence Bound

As discussed in the introduction, the lower bounds in the previous section are essentially tight for k-wise independent data structure problems, even in the special case of nonadaptive bit probe data structures and even when t is odd[5]. In particular for any constant $t > 1$, if we wish to obtain a data structure problem $F : [N] \times [M] \to \{\pm 1\}$ with $M \leq \exp(N^{o(1)})$ and satisfying a space/time tradeoff $S > N^{\frac{2}{t}+\epsilon}$ for some absolute constant $\epsilon > 0$ (again, even for nonadaptive bit probe data structures), we *provable cannot* rely solely on k-wise independence as our hardness property for the problem F. In this section we achieve lower bounds of this kind for every $t \in \mathbb{N}$ which is an odd number using a method from t-CSP refutation developed by [FKO06] in the case $t = 3$ and generalized to all t by [GKM22]. To establish the lower bounds here, we will have to pass from k-wise independent distributions to small-bias probability spaces.

For starters, we will need a variant of Theorem 7 which is specialized to nonadaptive generators with word length 1; remember from Section 2.2 that such generators, with time complexity t, are referred to as NC^0_t generators. The proof of this theorem more or less follows arguments in Section 9 of [GKM22], adapted to the setting of NC^0_t generators (in fact the argument becomes simpler in this setting):

Theorem 14. *Let $G : \{\pm 1\}^S \to \{\pm 1\}^N$ be an NC^0_t generator. Then there is $X \subseteq [N]$, $|X| \geq 2^{-2^t} N$, and a collection of 2^t different XOR schemes indexed by X, $(\mathcal{C}_u)_{u \subseteq [t]}$ so that:*

1. *Each XOR scheme has $\leq tS$ variables, and $\mathcal{C}_u$ is a $|u|$-XOR scheme.*
2. *For every $f \in \mathrm{range}(G)$ we have $\max_{u \subseteq [t], |u| < t} \mathrm{Val}(\mathcal{C}_u, f) \geq (1 - \mathrm{Val}(\mathcal{C}_{[t]}, f))2^{-t}$*

Proof. For each $x \in [N]$, $G_x : \{\pm 1\}^S \to \{\pm 1\}$ depends on t input cells which we list in some fixed order $p^1_x, \ldots, p^t_x$. For each x there is a predicate $h_x : \{\pm 1\}^t \to \{\pm 1\}$ so that if, on input $E \in \{\pm 1\}^S$, $G_x(E)$ sees the query outcomes $\xi_1 = E(p^1_x), \ldots, \xi_t = E(p^t_x)$, then $G_x(E) = h_x(\xi)$. Hence there

[5] We cannot really say "tight" in the case t is odd since we have not proven the lower bound in this case; what we mean is that there is an upper bound matching the form of Theorem 1 even for odd values of t.

exists $h : \{\pm 1\}^t \to \{\pm 1\}$ so that, for the set $X = \{x \in [N] \mid h_x = h\}$ we have $|X| \geq 2^{2^{-t}} N$. Fix this choice X, h for the remainder of the proof.

Let $V^1, \ldots, V^t$ be copies of $[S]$. For each $u \subseteq [t]$, we define the XOR scheme $\mathcal{C}_u = (X, V^u, c^u)$ indexed by X with variable set $V^u = \cup_{j \in u} V^j$ and constraints $c^u_x = \{p^j_x \mid j \in u\}$ where p^j_x is considered to live inside V^j; note that this is well-defined even for $u = \emptyset$, in which case we get a 0-XOR scheme $\mathcal{C}_\emptyset$ over 0 variables with $\mathrm{OPT}(\mathcal{C}_\emptyset, f) = \mathbb{E}_{x \in X} f_x$. The function h may be expressed uniquely as $h(\xi) = \sum_{u \subseteq [t]} \hat{h}(u) \cdot \prod_{j \in u} \xi_j$, where $(\hat{h}(u))_{u \subseteq [t]}$ are its Fourier coefficients. Now let $f \in \mathrm{range}(G)$ with encoding E_f, we have:

$$\begin{aligned} 1 &= \mathop{\mathbb{E}}_{x \in X} f_x \cdot h(E_f(p_1(x)), \ldots, E_f(p_t(x))) = \mathop{\mathbb{E}}_{x \in X} f_x \sum_{u \subseteq [t]} \hat{h}(u) \prod_{j \in u} E_f(p_j(x)) \\ &\leq \sum_{u \subseteq [t]} |\hat{h}(u)| \cdot |\mathop{\mathbb{E}}_{x \in X} f_x \prod_{j \in u} E_f(p_j(x))| = \sum_{u \subseteq [t]} |\hat{h}(u)| \cdot |\mathop{\mathbb{E}}_{x \in X} f_x \prod_{i \in c^u_x} E^u_f(i)| \\ &\leq \sum_{u \subseteq [t]} |\hat{h}(u)| \cdot \mathrm{Val}(\mathcal{C}_u, f) \leq \sum_{u \subseteq [t]} \mathrm{Val}(\mathcal{C}_u, f) \end{aligned}$$

where, in the second to last line, we use E^u_f to denote the vector in $\{0,1\}^{V_u}$ obtained from E_f by setting $E^u_f(i) = E_f(i)$ for every $i \in V^j \subseteq V_u$ ($E^\emptyset_f$ is the empty vector $\emptyset \mapsto \{0,1\}$). □

Using Theorem 14, we obtain from any NC^0_t generator a sequence of r-XOR schemes with various orders $r \leq t$. We will deal with the schemes $(\mathcal{C}_u)_{|u|<t}$ of order $< t$ using limited independence as in Section 3.2; for these we can hope to obtain upper bounds on $\mathrm{Val}(\mathcal{C}_u, \boldsymbol{f})$ which approach 0 at a reasonably fast rate, provided $S << N^{\frac{2}{(t-1)}}$. For the highest order scheme $\mathcal{C}_{[t]}$, we will only be able to obtain an upper bound on $\mathrm{Val}(\mathcal{C}_{[t]}, \boldsymbol{f})$ that is somewhat bounded away from one; however, if these two upper bounds match up correctly, we still have room to complete the lower bound according to Theorem 14.

The highest order term is dealt with using a technique developed in [FKO06, GKM22]. The main technical ingredient is the following theorem:

Theorem 15 ([GKM22])**.** *Let $t > 1$, $t = O(1)$ be fixed. Let $e : [N] \to \binom{[K]}{t}$ be a t-uniform multi-hypergraph with N edges and K vertices, and let $L \in \{t, \ldots, \frac{K}{8}\}$, such that $N \geq C \cdot K\left(\frac{K}{L}\right)^{\frac{t}{2}-1} \log^{4t+1} K$ for some constant C (depending on t). Then there exist a collection of disjoint nonempty sets $E_1, \ldots, E_r \subseteq [N]$, $r \geq \Omega(L^{-1}\left(\frac{K}{L}\right)^{\frac{t}{2}} \log^{4t} K)$, so that for every $j \leq r$ we have $\mathop{\Delta}_{x \in E_j} e_x = \emptyset$, where Δ denotes the symmetric difference operator.*

In [FKO06] as somewhat similar bound is obtained in the special case $t = 3$. We now use Theorem 15 to give a nontrivial upper bound on $\mathrm{OPT}(\mathcal{C}, \boldsymbol{f})$ when $\boldsymbol{f}$ has small bias:

Theorem 16. *Fix $t > 1$, $t = O(1)$. Let $\mathcal{C} = ([N], V, c)$ be a t-XOR scheme, $|V| = K$, and L be any integer with $t \leq L \leq \frac{K}{8}$. Assume that $N \geq C \cdot K(\frac{K}{L})^{\frac{t}{2}-1} \log^{4t+1} K$ for a constant C and that K is sufficiently large. Let $\boldsymbol{f} \in \{\pm 1\}^N$ be a random vector which is $\frac{1}{4}$-biased. Then*

$$\Pr[\mathrm{OPT}(\mathcal{C}, \boldsymbol{f}) \leq 1 - \Omega(\frac{K^{\frac{t}{2}} \log^{4t} K}{N L^{\frac{t}{2}+1}})] \geq \frac{1}{5}$$

Proof. For $R \in \{\pm 1\}^V$, let $\mathrm{Sat}(\mathcal{C}, f, R) = \Pr_{x \sim [N]}[f_x = \prod_{i \in c(x)} R(i)]$, $\quad \mathrm{Sat}(\mathcal{C}, f) = \max_R \mathrm{Sat}(\mathcal{C}, f, R)$. Then we have $\mathrm{Val}(\mathcal{C}, f, R) = |\mathrm{Sat}(\mathcal{C}, f, R) - \mathrm{Sat}(\mathcal{C}, -f, R)|$ for every R. We will show that $\Pr[\mathrm{Sat}(\mathcal{C}, \boldsymbol{f}) >$

$1-\delta] \le \frac{2}{5}$ for $\delta = \epsilon \cdot \frac{K^{\frac{t}{2}} \log^{4t} K}{10NL^{\frac{t}{2}+1}}$ for a sufficiently small constant $\epsilon > 0$ depending on t. Since the property of being γ-biased is invariant under negation, the same bound will hold for $-\boldsymbol{f}$ and so the theorem will follow by a union bound. By Theorem 15, there exist a collection of disjoint nonempty sets $E_1, \ldots, E_r \subseteq [N]$, $r \ge \epsilon \cdot L^{-1}(\frac{K}{L})^{\frac{t}{2}} \log^{4t} K$, such that for each $j \le r$ we have $\underset{x \in E_j}{\Delta} c_x = \emptyset$ where Δ denotes the symmetric difference operator. We now observe, as in [FKO06, GKM22], the following: for any f, if $\prod_{x \in E_j} f_x = -1$, then *for every* $R \in \{\pm 1\}^V$, there exists $x \in E_x$ so that $f_x \neq \prod_{i \in c_x} R(i)$. This follows since

$$\prod_{x \in E_j} \prod_{i \in c_x} R(i) = \prod_{i \in (\underset{x \in E_j}{\Delta} c_x)} R(i) = \prod_{i \in \emptyset} R(i) = 1$$

Hence, if we define $v(f) = \sum_{j \le r} \mathbb{1}\{\prod_{x \in E_j} f_x = -1\}$, then for every f we have $\mathrm{Sat}(\mathcal{C}, f) \le 1 - \frac{v(f)}{N}$. We will show that with probability at least $\frac{2}{5}$, we have $v(\boldsymbol{f}) \ge \frac{r}{10}$, in which case we have $\mathrm{Sat}(\mathcal{C}, f) \le 1 - \frac{r}{10N}$ and the theorem is proven. Now let $\gamma = \frac{1}{4}$ and define the random vector $\boldsymbol{g} \in \{\pm 1\}^r$, $\boldsymbol{g}_j = \prod_{x \in E_j} \boldsymbol{f}_x$. Observe that $\boldsymbol{g}$ is γ-biased inside $\{\pm 1\}^r$: for any $\emptyset \neq u \subseteq [r]$ we have

$$\mathbb{E} \prod_{j \in u} \boldsymbol{g}_j = \mathbb{E} \prod_{j \in u} \prod_{x \in E_j} \boldsymbol{f}_x = \mathbb{E} \prod_{x \in \cup_{j \in u} E_j} \boldsymbol{f}_x \in [-\gamma, \gamma]$$

using in the last step that $\boldsymbol{f}$ is γ-biased. So it suffices to show that, for γ-biased $\boldsymbol{g} \in \{\pm 1\}^r$, we have $|\sum_{j \in r} \boldsymbol{g}_j| \le \frac{8r}{10}$ with high probability. This is standard and can be verified by inspecting the second moment:

$$\Pr[|\sum_j \boldsymbol{g}_j| > \frac{8r}{10}] \le (\frac{8r}{10})^{-2} \mathbb{E}(\sum_j \boldsymbol{g}_j)^2 = (\frac{8r}{10})^{-2} \sum_{j,j'} \mathbb{E}\, \boldsymbol{g}_j \boldsymbol{g}_{j'} \le \frac{100}{64r^2}(\gamma r^2 + r) = \frac{100}{256} + 9r^{-1} \le \frac{2}{5}$$

Where the final inequality follows provided r is sufficiently large, which in turn follows provided K is sufficiently large. □

We also need the following easy fact:

Observation 3. *Let $\mathcal{C} = ([N], V, c)$ be a t-XOR scheme. Then for any $t' \ge t$ there is a t'-XOR scheme $\mathcal{C}' = ([N], V', c')$ with $|V'| \le |V| + (t' - t)$ so that $\mathrm{OPT}(\mathcal{C}, f) \le \mathrm{OPT}(\mathcal{C}', f)$ for every $f \in \{\pm 1\}^X$.*

Proof. Let $r = t' - t$ and set $V' = V \cup \{1, \ldots, r\}$, $c'_x = c_x \cup \{1, \ldots, r\}$. □

Observe that the above argument is valid even in the degenerate case $t = 0$. Finally we obtain:

Theorem (Theorem 2, restated)**.** *Let $t > 2$, $t = O(1)$ be a fixed odd number. Let $F : [N] \times [M] \to \{\pm 1\}$, be a data structure problem whose columns support a distribution which is γ-almost $\log N$-wise independent and $\frac{1}{4}$-biased, for $\frac{1}{\gamma} \ge N^{O(1)}$. Then any nonadaptive bit probe data structure for F with space S and time t must satisfy:*

$$S \ge \tilde{\Omega}(N^{\frac{1}{\frac{t}{2} - \frac{t-2}{2(t+2)}}})$$

Observe that for every $t > 2$, $\frac{t-2}{2(t+2)} > 0$, hence we obtain lower bounds of the form $S \ge N^{\frac{t}{2} + \Omega(1)}$ for every odd t. Recall also Lemma 1 from Section 2.1 in the preliminaries, which says that if a distribution is γ-biased then it is also γ-almost k-wise independent for any k. So we may strengthen our assumption on F to the simpler statement that its columns support an $N^{-O(1)}$-biased distribution. Our proof here will follow very closely the arguments in Section 9 of [GKM22].

Proof. Assume F has a data structure of space S and time $t = O(1)$, and let $\boldsymbol{f}$ be the random vector supported on the columns of F which is $\frac{1}{4}$-biased and N^{-c}-almost $\log N$-wise independence for c a suitably large constant. By Theorem 14 we may find $X \subseteq [N]$, $|X| \geq 2^{-2^t}N$, and XOR schemes $(\mathcal{C}_u)_{u\subseteq[t]}$ all indexed by X, each having $O(S)$ variables, so that $\mathcal{C}_u$ is a $|u|$-XOR scheme, and $\max_{u\subseteq[t],|u|<t} \mathrm{Val}(\mathcal{C}_u, f) \geq (1 - \mathrm{Val}(\mathcal{C}_{[t]}, f))2^{-t}$ for every column f of F. Using Observation 3 we may assume, up to increasing the number of variables of each XOR scheme by at most t, that $\mathcal{C}_u$ is a $(t-1)$-XOR scheme for every $|u| < t$. Set $L = S^{\frac{1}{t+2}}$, and assume that $N = \tilde{O}(1) \cdot \left(\frac{S}{L}\right)^{\frac{t}{2}} L$ where $\tilde{O}(1)$ is some sufficiently large $\mathrm{poly}\log N$ term. Applying Theorem 16, with probability at least $\frac{1}{5}$, we have

$$1 - \mathrm{OPT}(\mathcal{C}_{[t]}, \boldsymbol{f}) \geq \tilde{O}(1) \cdot \frac{S^{t/2}}{\left(\frac{S}{L}\right)^{\frac{t}{2}} L} = \tilde{O}(1) \cdot L^{-1}$$

On the other hand, using the fact that $\boldsymbol{f}$ is γ-almost $\log N$-wise independent and supported on the columns of F, with $\gamma \leq 2^{-O(tk)} = N^{-O(1)}$, we conclude using Theorem 10 (and the fact that $t-1$ is even) that for every $|u| < t$ we have:

$$\Pr[\mathrm{OPT}(\mathcal{C}_u, \boldsymbol{f}) \geq \delta] \leq N^{-1} << 2^{O(t)}, \quad \text{provided } N \geq \omega(\frac{1}{\delta^2} \log \frac{1}{\delta} S^{\frac{t-1}{2}} \log N)$$

Plugging in $\delta = \tilde{O}(1) \cdot 2^{-t} \cdot L^{-1}$, and using the fact $N = \omega(L^2 S^{\frac{t-1}{2}} \log N)$ as per our setting of L, there is a positive probability that

$$1 - \mathrm{OPT}(\mathcal{C}, \boldsymbol{f}) \geq \tilde{O}(1) \cdot L^{-1} > 2^{-t} \cdot \max_{0<|u|<t} \mathrm{Val}(\mathcal{C}_u, \boldsymbol{f})$$

and we reach a contradiction. Hence we cannot have $N \geq \tilde{O}(1) \cdot \left(\frac{S}{L}\right)^{\frac{t}{2}} L$, which by our setting of L yeilds the lower bound

$$S \geq \tilde{\Omega}(N^{\frac{1}{\frac{t}{2} - \frac{t-2}{2(t+2)}}})$$

□

The only reason we needed to assume t is odd is in our application of Theorem 10; if this theorem could be generalized to odd-arity XOR schemes (which seems attainable using methods for odd-arity CSPs, e.g. [GKM22]), the results here would be valid for all $t > 2$.

5 Algorithmic Implications of Our Lower Bounds

5.1 Range Avoidance

As referenced in the introduction, our main correspondence (Observation 1) immediately implies that explicit cell probe lower bounds yield NP oracle algorithms for range avoidance. We give a more formal version of this statement here:

Lemma 7. *Let $G : \Sigma^S \to \{0,1\}^N$ be a locally computable generator (resp. nonadaptive) with word length $w = \lceil \log |\Sigma| \rceil$ and time complexity t. There is a polynomial time* NP *oracle algorithm which, given G and a data structure problem $F \in \{0,1\}^{N \times M}$ which does not have space S, time t, word length w (resp. nonadaptive) data structures, solves the range avoidance problem for G.*

Proof. By Observation 1, the data structure lower bound for F implies that one of its columns must lie outside the range of G. We may use a NP oracle to test each column in order and find the lexicographically first which lies outside the range of G. □

Hence, our new explicit cell probe lower bounds immediately imply new polynomial time NP-oracle algorithms for range avoidance in a completely black box sense. For one of our lower bounds (Theorem 1), we will be able to analyze the argument in a non black box way and remove the NP oracle. For Theorem 2 we are unable to make such an improvement. We start with this latter result, where using known constructions of small bias probability spaces ([NN90]) we get:

Theorem (Theorem 3, restated)**.** *Let $t = O(1)$ be a fixed odd number. There is a polynomial time algorithm which outputs a list of strings in $\{0,1\}^m$, such that for any NC^0_t generator $G : \{0,1\}^n \to \{0,1\}^m$ with*

$$m \geq \tilde{O}(n^{\frac{t}{2} - \frac{t-2}{2(t+2)}})$$

one of the strings in the list must be a range-avoidance solution for G. In particular, there is a polynomial time NP-oracle algorithm for range avoidance on instances of the above form.

We now move on to our first lower bound from Theorem 1, which applies to more general adaptive generators with larger word lengths. Compared to Theorem 3, we will be able to remove the use of an NP oracle, and we will also be able to solve even the *remote point* problem for G, finding a string which is $(\frac{1}{2} - \epsilon)$-far in relative hamming distance from every string in the range of G for an arbitrarily small constant ϵ.

Theorem (Theorem 4, restated)**.** *Let t, w, n be given with $\log n > tw + 1$ and t even, and let $\epsilon \in (0, \frac{1}{2}]$ be given. There is a $m^{O(t^2 w)}$-time algorithm which, given any time-t locally-computable generator $G : [2^w]^n \to \{0,1\}^m$ with*

$$m \geq n^{\frac{t}{2}} \log n \cdot t^{\frac{t}{2}} 2^{(3+o(1))tw} \cdot \mathrm{poly}(\frac{1}{\epsilon})$$

outputs a string in $\{0,1\}^m$ which is $\geq (\frac{1}{2} - \epsilon)$-far in relative hamming distance from every string in the range of G.

We will set $N = m$, $S = n$ in the following, where m, n are as above. To obtain the hamming distance bound we require a strengthening of Theorem 7. Recall that Theorem 7 showed how to construct from a generator $G : \Sigma^S \to \{0,1\}^N$ a small collection of XOR schemes indexed by large subsets of $[N]$, so that whenever $f \in \mathrm{range}(G)$, f will induce a large value in one of the CSPs. The following strengthening will give us a similar conclusion under the weaker assumption that f is close in hamming distance to $\mathrm{range}(G)$. In the following, for two strings $f, g \in \{\pm 1\}^N$ and $X \subseteq [N]$, we use $\mathrm{Cor}_X(f,g) = |\mathbb{E}_{x \sim X} f_x g_x|$; by default $\mathrm{Cor}(\cdot,\cdot) = \mathrm{Cor}_{[N]}(\cdot,\cdot)$.

Theorem 17 (Correlation Variant of Theorem 7)**.** *Let $G : \Sigma^S \to \{\pm 1\}^N$ be a locally-computable generator with time complexity t, word length $w = \log |\Sigma|$. There exists a collection of sets $(X_j \subseteq [N])_{j \leq 2^{tw+1}}$, $|X_j| \geq \epsilon 2^{-tw-1} N$, and a collection of t-XOR schemes $(\mathcal{C}_j)_{j \leq 2^{tw+1}}$, with $\mathcal{C}_j$ a t-XOR scheme indexed by X_j and having $\leq tS$ variables, so that for every $f, g \in \{\pm 1\}^N$ with $g \in \mathrm{range}(G)$, $\mathrm{Cor}(f,g) \geq \epsilon$, there exists $j \leq 2^{tw+1}$ with $\mathrm{Val}(\mathcal{C}_j, f) \geq \epsilon 2^{-tw-1}$.*

Proof of Theorem 17. We follow the proof of Theorem 7 and indicate only those parts which need to change. The definitions of $\mathcal{C}^{(\pi,b)}$, $X^{(\pi,b)}$ remain as before; in this case we discard any such pair with $|X^{(\pi,b)}| < \epsilon 2^{-tw-1} N$. Let f, g be as in the statement of the theorem, and let E_g be a canonical

preimage of g under G. For each π let $\tilde{X}^{(\pi,g)} \subseteq [N]$ consist of those x such that $T_x(E_g)$ reaches the leaf corresponding to π. These sets (ranging over $\pi \in \Sigma^t$) partition $[N]$ and so we have

$$\epsilon \leq \mathrm{Cor}(f,g) = |\sum_{\pi \in \Sigma^t} \frac{|\tilde{X}^{(\pi,g)}|}{N} \mathop{\mathbb{E}}_{x \sim \tilde{X}^{(\pi,g)}} f_x g_x| \leq \sum_{\pi \in \Sigma^t} \frac{|\tilde{X}^{(\pi,g)}|}{N} \mathrm{Cor}_{\tilde{X}^{(\pi,g)}}(f,g)$$

There must then exist $\pi^{f,g}$ so that $\frac{|\tilde{X}^{(\pi^{f,g},g)}|}{N} \mathrm{Cor}_{\tilde{X}^{(\pi^{f,g},g)}}(f,g) \geq \epsilon 2^{-tw}$, in particular $|\tilde{X}^{(\pi^{f,g},g)}| \geq \epsilon 2^{-tw} N$ and $\mathrm{Cor}_{\tilde{X}^{(\pi^{f,g},g)}}(f,g) \geq \epsilon 2^{-tw}$. We take $\pi^{f,g}$ in place of "π^f" from the previous proof and $\tilde{X}^{(\pi^{f,g},g)}$ in place of "$\tilde{X}^{(\pi^f,f)}$," and and otherwise carry out essentially the same argument.

□

Using this in conjunction with the main ideas in Section 3.2 we can now prove Theorem 4.

Proof of Theorem 4. Given the generator G, we may apply Theorem 17 to compute a list of 2^{tw+1} XOR schemes, so that whenever a string in $\{0,1\}^m \leftrightarrow \{\pm 1\}^m$ makes the value of all XOR schemes $< \epsilon 2^{-tw-1}$, it must be $(\frac{1}{2} - \epsilon)$-far in relative hamming distance from range(G). We will use [NN90] to construct a polynomial size list of candidate strings, such that one of them must make all such values small simultaneously which will guarantee that it is far form range(G). To make our algorithm run in polynomial time, we need to certify that one of these strings from the list accomplishes this goal. To perform this certification, we transform the XOR schemes into matrix ensembles using Theorem 8 followed by Lemma 4, and then compute the associated spectral norms. We can observe easily that each of the transformations prescribed in Theorems 7 and 8 and in Lemma 4 can be performed efficiently. Using [NN90] as indicated above, we can construct in time $2^{O(ktw)}$ a list of $\leq m^{O(tw)}$ strings which support a $2^{-O(ktw)}$-almost k-wise distribution for $k = t \log N$. By Theorem 10, one of the strings in this list will pass all of the 2^{tw+1} spectral tests define above; our algorithm outputs the first such string. □

5.2 Linear Attacks on Local PRGs

Here we state formally our improvement to an old result of [MST03] on ϵ-biased generators in NC^0 referenced in the introduction:

Corollary (Corollary 1, restated)**.** *Let $t = O(1)$ be a fixed odd number. If $m \geq \tilde{O}(n^{\frac{t}{2} - \frac{t-2}{2(t+2)}})$, then for any NC^0_t generator $G : \{0,1\}^n \to \{0,1\}^m$ and any random variable $\boldsymbol{x} \in \{0,1\}^n$, $G(\boldsymbol{x})$ fails to be γ-biased for some $\gamma \geq n^{-O(1)}$.*

This is simply a reparameterization of Theorem 2 and does not require any further argument. Corollary 1 implies that for any odd $t = O(1)$ and any candidate NC^0_t PRG $G : \{0,1\}^n \to \{0,1\}^{n^{\frac{t}{2} - \epsilon_t}}$, there exists a *nonuniform distinguisher* for G, computable by a single parity gate, where $\epsilon_t > 0$ is a positive constant depending only on t. However, as noted in the introduction, our proof does not indicate how to find such a linear distinguisher, and hence has no bearing on the security of Goldreich's PRG in the stretch regime $n \mapsto n^{\frac{t}{2} - \epsilon_t}$ (even against nonuniform adversaries). It is therefore of interest to determine whether we can use our methods to find a *uniform distinguisher* for NC^0_t PRGs in this stretch regime.

One approach is to use the natural proofs paradigm suggested in Section 6.2. In particular, if we can define a uniformly computable natural property which proves nonadaptive bit probe lower bounds of the form $S \geq N^{\frac{t}{2} - \epsilon_t}$, we could use this in conjunction with a hybrid argument as in

Lemma 9 to uniformly distinguish such generators[6]. In particular, we showed that for any odd $t = O(1)$, if F supports a distribution which is $N^{-O(1)}$-almost $O(\log N)$-wise independent and $\frac{1}{4}$ biased, than it must satisfy a lower bound of the form $S \geq N^{\frac{t}{2}-\epsilon_t}$. For $M = \text{poly}(N)$ sufficiently large, we can easily certify the property of $N^{O(1)}$-almost $O(\log N)$-wise independence with all but negligable probability for a random $F : [N] \times [M] \to \{0,1\}$ in quasipolynomial time (iterate over every subset of $O(\log N)$ variables). Hence a natural proof (computable uniformly in $\text{poly}(N)$ time) that a matrix $F : [N] \times [N^{O(1)}] \to \{0,1\}$ is $\frac{1}{4}$ biased would yield a distinguisher. If we are a bit more careful in our argument, computing spectral norms of matrices depending on the given generator G as in the proof of Theorem 4 instead of naively certifying limited independence by brute force, we arrive at:

Lemma 8. *Say that there is a constant c and a polynomial time algorithm which can certify with nonnegligble probability that a random $F : [N] \times [N^c] \to \{0,1\}$ supports a $\frac{1}{4}$-biased distribution. Then there is a uniform polynomial time distinguisher for NC^0_t generators $G : \{0,1\}^n \to \{0,1\}^m$ whenever t is odd and $m \geq n^{\frac{t}{2}-\epsilon_t}$ for some $\epsilon_t > 0$ depending only on t. In particular, Goldreich's generator is insecure in this parameter regime.*

6 Towards Stronger Lower Bounds

In this section we discuss some of the broader implications of Observation 1 connecting locally-computable generators to cell probe lower bounds. Our focus here will be on the setting of nonadaptive bit probe data structures with time complexity $t = O(1)$, which correspond via Observation 1 to the class of NC^0 generators. More specifically, we are interested in the problem of obtaining lower bounds of two kinds, which we refer to as *strong* and *super-strong* respectively. In each setting, our goal is to construct lower bounds for a problem F with N queries; N will be the main complexity parameter. We seek lower bounds on the space as a function of N and the time t; subject to obtaining such a lower bound and the problem being explicit, we'd like to minimize the number of datapoints M of our problem as a function of N. We will parameterize each goal by a growth rate $M(N)$, where the goal gets more difficult the slower M grows:

Goal 1 (Strong Lower Bounds with $M(N)$ Datapoints)**.** *Exhibit an explicit data structure problem $F : [N] \times [M(N)] \to \{0,1\}$, and some universal constant $\epsilon > 0$, so that for any $t = O(1)$, F requires space $S \geq N^\epsilon$ for time-t nonadaptive bit probe data structures.*

Goal 2 (Super-Strong Bit Lower Bounds with $M(N)$ Datapoints)**.** *Exhibit an explicit data structure problem $F : [N] \times [M(N)] \to \{0,1\}$ so that F requires space $S \geq N - N^{o(1)}$ for data structures making 4 nonadaptive bit probes.*

In the most ambitious case we aim for $M(N) = \text{poly}(N)$, and in the most general case we aim for $M(N) = \exp(N^{o(1)})$. We observe that nonconstructive counting arguments imply that Goal 1 is achievable by a nonexplicit problem once $M \geq O(t \log N)$, while Goal 2 is achievable by a nonexplicit problem once $M \geq O(tN \log N)$, so even in the ambitious regime $M = \text{poly}(N)$ neither goal is vacuous in an information-theoretic sense. Note that our main lower bounds in Sections 3 and 4 are of the form $S \geq N^{\delta_t}$ where $\lim_{t\to\infty} \delta_t = 0$ and hence they do not achieve either of these goals. In this section we discuss some barriers and approaches to achieving lower bounds of these kinds, and connect these problems to other topics in complexity theory. More specifically we will

[6]We proved Lemma 9 for nonuniform distinguishers, however the hybrid argument is also valid in the uniform setting (with randomized distinguishers).

discuss barriers for both problems, while our "approaches" will only apply to the first (Goal 1); we will see shortly that super-strong lower bounds (Goal 2) are likely to be completely out of reach of current mathematical methods.

Before proceeding, we make a note of our use of the word *explicit*; for the rest of this section, we will use this word in the following formal sense:

Definition 12. *We say that a sequence of strings $(x_n)_{n\in\mathbb{N}}$ is explicit if there is a uniform algorithm which, given 1^n as input, will print x_n in time $\mathrm{poly}(n, |x_n|)$. In particular, if $(F_N : [N] \times [M(N)] \to \{0,1\})_{N\in\mathbb{N}}$ is a family of data structure problems, we say that F_N is explicit if there is a $\mathrm{poly}(N, M(N))$ time algorithm which, given 1^N, produces the truth table of F_N, i.e. the $N \times M(N)$ matrix consisting of all values $F(x, y)$.*

In this more formal terminology, our statement of Goal 1 should be interpreted as follows: define an explicit *family* of problems $F_N : [N] \times [M(N)] \to \{0,1\}$ for each N, so that $M(N) \leq \exp(N^{o(1)})$, and F_N requires space $\Omega(N^\epsilon)$ for time t data structures, where ϵ, t are as in the statement of Goal 1, and N goes to infinity after these are fixed. In all of our main results here, we may replace the assumption that F is defined for all N by the assumption it is defined only for N of some special forms, e.g. prime numbers, powers of 2, etc., provided this class of allowable values is efficiently recognizable, and not too sparse as a subset of $\mathbb{N}$. We stick to the above definition which requires the lower bound to hold for all N only for the purpose of simplicity.

As has been observed repeatedly in complexity theory, this formal notion of *explicitness* is perhaps overly broad compared to the typical *informal* use of the word; indeed, in many places it is preferred to use the more stringent notion of explicitness, requiring $F(x, y)$ to be computable uniformly in $\mathrm{poly}(\log N, \log M)$ time, which still captures apparently every data structure problem in the literature for which an interesting cell probe lower bound is currently known.

6.1 Consequences of Explicit Bit Probe Lower Bounds

Recall Lemma 7 from the previous section, which was a direct corollary of our main correspondence (Observation 1): every explicit data structure lower bound yields a NP oracle range avoidance algorithm for locally computable generators with corresponding parameters. We apply this as follows:

Theorem (Theorem 5, restated)**.**

1. *If explicit super-strong bit probe lower bounds with subexponentially many datapoints (Goal 2, $M = \exp(N^{o(1)})$) are achievable, then $\mathsf{EXP}^{\mathsf{NP}} \not\subseteq \mathsf{NC}^1$.*

2. *If explicit strong bit probe lower bounds with polynomially many datapoints (Goal 1, $M = \mathrm{poly}(N)$) are achievable, then there is a polynomial time NP oracle algorithm for NC^0 range avoidance with polynomial stretch.*

Proof. (1) If Goal 2 with $M(N) = \exp(N^{o(1)})$ is achievable than we can solve NC^0_t range avoidance in the stretch regime $n \mapsto n + n^{o(1)}$ in subexponential time with an NP oracle using Lemma 7. By [RSW22], this implies the separation $\mathsf{EXP}^{\mathsf{NP}} \not\subseteq \mathsf{NC}^1$.

(2) By well known composition reductions (see [Kor21]), we may reduce any NC^0 range avoidance $G : \{0,1\}^n \to \{0,1\}^{n^{1+\Omega(1)}}$ to a second NC^0 instance $G' : \{0,1\}^n \to \{0,1\}^{n^c}$ where $c = O(1)$ is an arbitrarily large constant in polynomial time with an NP oracle. If Goal 1 is achievable than by Lemma 7 we can solve range avoidance in this regime using an NP oracle in polynomial time. □

In (2) above, if we instead achieved Goal 1 with $M = \exp(\text{poly}\log N)$ we would obtain a quasipolynomial time algorithm; if we achieved it with $M = \exp(N^{o(1)})$ we would obtain a subexponential time algorithm. Obtaining P^{NP} algorithms for NC^0 range avoidance in the polynomial stretch regime was explicitly posed as an open problem in [RSW22].

6.2 Natural Bit Probe Lower Bounds Break NC^0 PRGs

We present here a natural proofs barrier to both "strong" and "super-strong" bit probe lower bounds. The natural proofs paradigm, introduced in the seminal work of Razborov and Rudich [RR97], identified the following pattern in all preceeding known lower bounds in complexity theory. For a lower bound problem, we have in mind in some resource-bounded computational model $\mathcal{C}$ for computing boolean valued functions $f : U \to \{0,1\}$, where U is a finite set with some structure (e.g. $U = \{0,1\}^n$ or $U = [N] \times [N]$). We would like to present an explicit function $f_{\text{explicit}} : U \to \{0,1\}$ which we can unconditionally prove requires a large amount of resources for computations in the class $\mathcal{C}$. Razborov and Rudich observed that, in those cases where we *can* prove such an explicit lower bound, the proof can be extended (with only minimal-to-moderate additional work) to have the following features: (1) the proof defines an efficiently decidable "property of functions $f : U \to \{0,1\}$" so that f_{explicit} has the property, and whenever f has this property it is hard for the class $\mathcal{C}$, and (2) not only does f_{explicit} have this property, but so does a *uniformly random* function $f : U \to \{0,1\}$ (with high probability). The "property of functions" referenced above is called a *natural property* in [RR97], and by "efficiently" we mean time polynomial/subexponential in the description size of f, which in this case is $|U|$. Razborov and Rudich then went on to show:

1. All previously known lower bounds are natural.
2. Assuming widely believed cryptographic assumptions, no natural proof can give superpolynomial lower bounds on circuit size.

Together these results give a strong indication that, to prove $f \notin \mathsf{P/poly}$ for some explicit function $f : \{0,1\}^n \to \{0,1\}$, we must develop methods of a fundamentally different nature than all results known up to the point of [RR97]. The situation has not changed much since the publication of [RR97], and this barrier is still considered to be of great significance in complexity theory. We now define formally the corresponding notion of natural proofs for the bit probe complexity measure:

Definition 13 (Natural Bit Probe Lower Bounds)**.** *A (N, M, S, t) natural property is an algorithm $\mathcal{A}$ which takes as input the description of a data structure problem $F : [N] \times [M] \to \{0,1\}$, and outputs a value in $\{\mathsf{Hard}, ?\}$, such that:*

1. *If $\mathcal{A}(F) = \mathsf{Hard}$, then F requires space $\geq S$ or time $> t$ for nonadaptive bit probe data structures.*
2. *If F is chosen uniformly at random amongst all functions $[N] \times [M] \to \{0,1\}$, then with probability at least $(NM)^{-O(1)}$, $\mathcal{A}(F) = \mathsf{Hard}$.*

We are interested in the existence of natural properties $\mathcal{A}$ which are computable by circuits of size $<< 2^N$, i.e. $\text{poly}(N)$ or $\exp(N^{o(1)})$. In our discussion we will have N as our main asymptotic complexity parameter, t fixed, and $S = S(N)$, $M = M(N)$ growing as a function of N. We will then refer to a $(N, M(N), S(N), t)$-natural property in plain english as a "natural proof of a space lower bound $\geq S(N)$ against time t data structures for a problem with $M(N)$ datapoints." If the natural property is computable by size k circuits, we say that it is a "k size computable" natural proof.

In the following, we will say that a cryptographic PRG $G : \{0,1\}^n \to \{0,1\}^m$ has "hardness λ" if circuits of size $\leq \lambda$ fail to distinguish G with advantage better than λ^{-1}.

Lemma 9. *Say that $G : \{0,1\}^S \to \{0,1\}^N$ is a cryptographic NC^0_t PRG with hardness λ. Then there does not exist an (N, M, S, t) natural property computable by circuits of size k unless $\lambda \leq k \cdot \mathrm{poly}(NM)$.*

Proof. Let G be a cryptographic NC^0_t PRG with hardness λ. Consider the generator $G^{\otimes M} : \{0,1\}^{S \cdot M} \to \{0,1\}^{N \cdot M}$ given by $G(E_1, \ldots, E_M) = (G(E_1), \cdots, G(E_M))$. A standard application of the hybrid argument implies that $G^{\otimes M}$ is also a cryptographic PRG with hardness $M^{-1}\lambda$. On the other hand, consider any string in the range of $G^{\otimes M}$. Interpreted as data structure problem $F : [N] \times [M] \to \{0,1\}$, this problem has time t nonadaptive bit probe data structures of space $\leq S$: if $G^{\otimes M}(E_1, \ldots, E_M) = F$, then $E_1, \ldots, E_M$ exhibit preimages under G for each column of F.

Say there existed a (N, M, S, t) natural property $\mathcal{A}$ computed by circuits of size k. Then $\mathcal{A}$ outputs Hard on a random $N \cdot M$ bit string with probability $\geq (NM)^{-O(1)}$ over the uniform distribution, and with probability zero over the range of our generator $G^{\otimes M}$. Hence $k^{-1}(NM)^{-O(1)}\lambda \leq O(1)$, i.e. $\lambda \leq k(NM)^{O(1)}$. □

To establish our natural properties barriers we need two results from NC^0 cryptography. The first is folklore and follows from standard PRG composition:

Lemma 10. *If there exists an NC^0 PRG $G : \{0,1\}^n \to \{0,1\}^m$ with arbitrary polynomial stretch ($m \geq n^{1+\Omega(1)}$) and hardness λ, then for every $c \in \mathbb{N}$ there exists an NC^0 PRG $G' : \{0,1\}^n \to \{0,1\}^{n^c}$ with hardness $\geq n^{-O(1)}\lambda$.*

Hence, from the existence of polynomial stretch NC^0 PRGs with superpolynomial hardness we obtain NC^0 PRGs with superpolynomial hardness of the form $G : \{0,1\}^{N^\epsilon} \to \{0,1\}^N$ for arbitrarily small constants $\epsilon > 0$. We also need the following from [AIK06]:

Lemma 11. *If there exists an NC^1 PRG with hardness λ then there exists an NC^0_4 PRG $G : \{0,1\}^m \to \{0,1\}^{m+m^{0.99}}$ for some $m \leq \mathrm{poly}(n)$ with hardness $\geq n^{-O(1)}\lambda$.*

Combining these two lemmas with our Lemma 9 we arrive at:

Theorem (Theorem 6, restated: Natural Proofs Barriers for Bit Probe Lower Bounds)**.**

1. *If NC^1 PRGs with superpolynomial hardness exist, then there is no polynomial-size computable natural proof of a lower bound $S \geq N - N^{0.99}$ against bit probe data structures making 4 nonadaptive bit probes for a problem with $M \leq \mathrm{poly}(N)$ datapoints.*

2. *If NC^0 generators with superpolynomial hardness exist, then for every ϵ there exists t so that there is no polynomial-size computable natural proof of a lower bound $S \geq N^\epsilon$ for against time t bit probe data structures for a problem with $M \leq \mathrm{poly}(N)$ datapoints.*

As in the last section, if we increase the hardness parameter in our cryptographic assumptions we obtain stronger conclusions in terms of the runtime of our natural proofs and the allowable number of datapoints M. In both cases it is plausible that the stated PRGs exists with hardness $\lambda = \exp(N^{\Omega(1)})$, in which case we conclude that natural proofs computable by circuits of size $\exp(N^{o(1)})$ cease to exist already in the regime $M = \exp(N^{o(1)})$.

As in [RR97], to give a convincing argument that this "barrier" ought to be taken seriously we need some indication that the known lower bound techniques have not already broken it. To this

end we consider known techniques that can prove the strongest lower bounds for explicit problems $F : [N] \times [M] \to \{0,1\}$ with $M \leq \exp(N^{o(1)})$ in the regime $t = O(1)$.

First we consider the communication bound, which can prove a space lower bound $S \geq \Omega(N^{\frac{1}{t}})$ already in the regime $M \leq \text{poly}(N)$. This lower bound holds provided F has maximal $(\log N)$ deterministic communication complexity, so it suffices to give a natural proof of strong communication lower bounds. For this we can use as our natural property "F has full rank" where the rank is measure either over $\mathbb{R}$ or $\mathbb{F}_2$. This property is computable in polynomial time, equals N for a random F with nonnegligable (actually very high) probability, and yields a maximal lower bound $\log N$ on the communication complexity of F. We make a note that lower bounds for stronger complexity measures for two party communication, including randomized and nondeterministic communication, can be made natural as well using the discrepency bound and related techniques. The original work of [RR97] proves this for the standard discrepancy bound which lower bounds both randomized and nondeterministic complexity. In the case of randomized communication one can use the more precise γ_2^∞-norm instead to obtain natural lower bounds with stronger parameters, see [LS07].

Next we turn to the cell sampling method used in [Sie89, Sie04, PTW10, Lar12] and our new lower bounds of the form $S \geq N^{\frac{2}{t}}$ from Section 3. Note that in the regime $M \leq \exp(N^{o(1)})$ the cell sampling bound does not give a significant improvement over the communication bound so it does not really require a separate treatment. Nonetheless, for both the cell sampling bound and our Theorem 1, the property we may use for the problem F is the following: the uniform distribution on columns of F is N^{-c}-almost $c \log N$-wise independent for some constant c depending on t. Setting $M = N^{O(c)}$, we can certify this property for a random F with high probability in time $N^{O(\log N)}$ by iterating over every subset of $c \log N$ rows of F. Note that this natural property runs in quasipolynomial time (rather than polynomial), however this is still interesting as the cryptographic assumptions in question are widely believed to have quasipolynomial (and much higher) hardness[7].

Finally we get to the nonadaptive bit probe lower bounds in Section 4 of the form $S \geq N^{\frac{1}{2}+\epsilon_t}$; here we actually *do not* know whether the lower bound in question can be made natural in the strict sense of Definition 13; this question is discussed further in Section 5. Nonetheless the lower bound in Theorem 2 still has the following form: for any random $\boldsymbol{f}$ which is *fools* $\mathbb{F}_2$ linear tests, we have $\boldsymbol{f} \notin \text{range}(G)$ with nonzero probability, for any NC^0_t generator G with large enough stretch. Stated contrapositively, we conclude that for any candidate NC^0_t generator $G : \{0,1\}^n \to \{0,1\}^{n^{t/2-\epsilon_t}}$, there exists a $\mathbb{F}_2$-linear test distinguishing its output from random, so in particular G cannot be secure as a PRG (this was Observed in Corollary 1 of the previous section). In other words, our lower bound still yields as a byproduct a method for distinguishing the associated PRGs, although in this case the existence of a distinguisher is shown nonconstructively. For this reason it seems that the lower bounds in Section 4 still have a degree of algorithmic content which cannot be present in any lower bound achieving Goals 1 and 2 (assuming the corresponding cryptographic generators exist).

6.3 Approaches to Strong Lower Bounds via Communication Complexity

Here we discuss some possible approaches to Goal 1: proving space lower bounds for time $t = O(1)$ nonadaptive bit probe data structures of the form $S \geq N^\epsilon$, where ϵ is a universal constant not

[7] Actually, we only need to assume that our cryptographic PRGs require super-quasipolynomial size circuits to be distinguished with polynomial advantage for a natural proof with these parameters to contradict their security, since $M = \text{poly}(N)$.

depending on t. The reader may observe that we would have arrived at an equivalent problem if we had used adaptive in place of nonadaptive bit probe data structures: any adaptive bit probe data structure with time complexity t can be converted into a nonadaptive data structure with time complexity 2^t by unrolling each query decision tree. We will make another simplifying assumption on our data structures, having no effect on Goal 1, which will streamline some of our arguments:

Definition 14. *We say that an NC^0_t generator is homogeneous if there is a single predicate $h : \{0,1\}^t \to \{0,1\}$, so that each output $G_x : \{0,1\}^S \to \{0,1\}^N$ is of the form $G_x(E) = h(E(p^1_x), \ldots, E(p^t_x))$ for some nonadaptive probe sequence $p^1_x, \ldots, p^t_x \in [S]$. We say that F has homogeneous data structures with time t and space S if there exists a homogeneous NC^0_t generator whose range contains every column of F.*

Just as the distinction between adaptive and nonadaptive data structures is irrelevant with respect to lower bounds achieving Goal 1, so too is the distinction between general and homogeneous data structures:

Lemma 12. *Let F be any data structure problem, $t \in \mathbb{N}$. There exist constants t', c depending only on t, so that if F has space S time t bit probe data structures, then F has space $c \cdot S$ and time t' homogeneous data structures.*

We will therefore focus specifically on homogeneous data structures with time $t = O(1)$. Homogeneous data structures have a rather appealing formulation in terms of communication complexity:

Definition 15 (Parallel Channel Model)**.** *Let $F : X \times Y \to \{0,1\}$ be a two-party communication problem. We say that F has a (t,k)-parallel protocol if it can be computed in the following way:*

1. *Alice receives $x \in X$; a council of t Bobs $B_1, \ldots, B_t$ receive (the same input) $y \in Y$; a referee Charlie receives no input.*
2. *(Round 1:) Alice sends a different k-bit message privately to each Bob*
3. *(Round 2:) Each Bob sends 1 bit to Charlie*
4. *(Round 3:) Charlie outputs the answer $F(x,y)$.*

The equivalence of this model (up to minor quantitative losses) to the homogeneous data structure model is a matter of definitional translation; the relation closely mirrors that between locally decodable codes and private information retrieval schemes:

Observation 4. *If F has homogeneous bit probe data structures with space S and time t, then F has a $(t, \log S)$-parallel protocol. Conversely, if F has a (t,k)-parallel protocol then it has a homogeneous data structure with time t and space $t \cdot 2^k$.*

We refer to this as equivalence, since w.r.t. Goal 1 the difference between a space lower bound S and tS is irrelevant.

Proof. We prove only the first direction. Let $G : \{0,1\}^S \to \{0,1\}^N$ be a homogeneous NC^0_t generator whose range covers the columns of F. Prior to communication, all of the Bobs prepare the data structure encoding E_y for y. Alice sends to B_j the message $p^j_x \in [S]$ which costs $\log S$ bits. After receiving this value, B_j sends Charlie the bit $E_y(B_j)$. Charlie, upon receiving bits $b_1, \ldots, b_t$ from the t Bobs, outputs $h(b_1, \ldots, b_t)$. □

We can then restate Goal 1 as follows:

Goal 3 (Strong Parallel Communication Lower Bounds for a problem with M Columns)**.** *Exhibit an explicit $F : \{0,1\}^n \times [M] \to \{0,1\}$ and some universal constant $\epsilon > 0$, so that F does not have $(O(1), \epsilon n)$-parallel protocols.*

In this setting we will set $M = 2^m$ for some m, and identify $[M]$ with $\{0,1\}^m$. In this case, achieving $m \leq O(n)$ for Goal 3 is equivalent to our most ambitious goal of achieving $M \leq \text{poly}(N)$ for Goal 1; we would also be happy with $m \leq 2^{o(n)}$ in Goal 3, which corresponds to our most general acceptable form of Goal 1 with $M \leq \exp(N^{o(1)})$.

While our primary intention in investigating the parallel channel model is as an avenue through which to derive new cell probe lower bounds, using Observation 4 we can use our lower bounds from Sections 3 and 4 to derive novel lower bounds in this communication model, for example:

Lemma 13. *Let $t = O(1)$ be an even number. Consider the following communication problem: Alice recieves $x \in \mathbb{F}_2^n$, and t Bobs receive a polynomial $p : \mathbb{F}_2^n \to \mathbb{F}_2$ of degree $\log n + \log\log n$; their goal is to compute $p(x)$. In the parallel channel model, Alice must send some Bob at least $\frac{2n}{t} - O(\log n)$ bits to solve this problem.*

We now discuss two avenues for approaching strong lower bound in the parallel channel model with $O(1)$ channels.

6.3.1 A Lifting-Based Approach To the Parallel Channel Model

For starters, we consider a restricted form of (t,k)-parallel protocol for a problem $F : \{0,1\}^n \times \{0,1\}^m \to \{0,1\}$, which we call a *projection protocol*. A (t,k) projection protocol is a special kind of (t,k)-parallel protocol, in which Alice is restricted to chose in advance sets $I_1, \ldots, I_t \subseteq [n]$ with $|I_j| \leq k$, and upon input x, she will always send the j^{th} Bob the values $x|_{I_j} = (x_i)_{i \in I_j}$. In other words, instead of sending the j^{th} Bob an arbitrary k-bit message depending on x, she is restricted to sending him some fixed set of k coordinates of x. For protocols of this form, we may observe the following:

Claim 1. *If F has (t,k) projection protocols then every column of $f : \{0,1\}^n \to \{0,1\}$ of F is expressible in the form $f(x) = h(g_1(x), \ldots, g_t(x))$ where $h : \{0,1\}^t \to \{0,1\}$ and g_j depends on at most k variables amongst $\{x_1, \ldots, x_n\}$.*

Boolean functions expressible in this way have been studied in [HR15, LRT22]. In particular, the minimum value t such that f can be expressed as $f = h(g_1, \ldots, g_t)$ for g_j depending on $\leq k$ variables is referred to as $C_k^2(f)$ in [HR15]. The "2" in the superscript refers to the interpretation of $C_k^2(f)$ as the smallest top-fanin of a depth 2 circuit with arbitrary gates and bottom fanin k computing f. Intriguingly, strong lower bounds on $C_k^2(f)$ for explicit functions f were proved unconditionally in [HR15]. For example, as a direct corollary of a classical result in [CFL83] on the number-on-forehead complexity of threshold functions, they observe:

Lemma ([CFL83, HR15])**.** *For any constant γ, $C^2_{(1-\gamma)n}(\text{MAJ}_n) = \omega(1)$.*

We are using MAJ_n to denote the n bit majority function. Quantitatively stronger lower bounds on $C_k^2(\cdot)$ in the same regime $k = (1-\gamma)n$ are derived in [HR15] for other more complicated explicit functions. In addition, an optimal lower bound on $C_k^2(\text{MAJ}_n)$ is derived in [LRT22] in a quite different parameter regime $k \leq n^{1-\epsilon}$. The above result implies that even the problem $F : \{0,1\}^n \times \{0,1\}^0 \to \{0,1\}$, given by $F(x, \emptyset) = \text{Maj}_n(x)$ is hard for $(O(1), 0.99n)$ projection protocols. This lower bound quantitatively matches (and in fact exceeds) the kind sought in Goal 3. Obviously in the general parallel channel model we cannot hope to obtain a lower bound

for a problem F with a single column (Alice could precompute $f(x)$ in this case); nonetheless we believe this toy model presents a promising avenue towards lower bounds for general protocols, via the machinery of *query-to-communication-lifting*.

Lifting is a powerful method used to obtain lower bounds in communication complexity (see [GPW20] for a prototypical example). The central idea is to start with an arbitrary base function, and "lift it" via function composition to get a new composed function whose *communication complexity* is essentially the same as the *query complexity* of the original base function. Proofs of lifting theorems typically involve a simulation theorem, showing from *any* protocol for the composed function, how to extract a simple protocol for the base function. Thus lifting simulation theorems give a constructive proof that the most efficient communication for the composed problem is essentially the protocol that simply mimics the optimal simple protocol, implying that separating the simpler query classes is not only necessary but also sufficient for proving the stronger communication separation.

We define here a standard lifting setup where the inner "gadget" function is the index function.

Definition 16 (Lifting with Index)**.** *Let $f : \{0,1\}^n \to \{0,1\}$ and ℓ be given. Let $X = [\ell]^n$, $Y = (\{0,1\}^\ell)^n$. We define the lifted function $f \circ \mathrm{Ind}_n^\ell : X \times Y \to \{0,1\}$ by $f \circ \mathrm{Ind}_n^\ell(x,y) = f(y_1(x_1), \ldots, y_n(x_n))$.*

With this definition, it is natural to view the parallel-channel model as the "communication analogue" of the complexity measure $C_k^2(\cdot)$. In particular, we can observe that F is computed by a (t,k) parallel protocol if and only if we can express it in the form:

$$F(x,y) = h(g_1(x,y), \ldots, g_t(x,y))$$

where $h : \{0,1\}^t \to \{0,1\}$ and $g_1, \ldots, g_t : X \times Y \to \{0,1\}$ have one way communication complexity (rows speaking to columns) $\leq k$. Compare this to the definition of $C_k^2(f) \leq t$, which means f can be expressed as:

$$f(x) = h(g_1(x), \ldots, g_t(x))$$

with $g_j : \{0,1\}^n \to \{0,1\}$ depending on at most k variables, i.e. g_j having *nonadaptive decision tree depth* at most k. Prior work has established that one-way communication complexity can be naturally viewed as a communication complexity analogue of "nonadaptive decision tree depth," and in particular very strong lifting theorems have been established (by quite simple arguments) which transfer nonadaptive decision tree lower bounds to one way communication lower bounds [MSS21]. This intuition leads us to make the following conjecture:

Conjecture 1 (Lifting C_k^2 To Parallel Channel Complexity)**.** *There is a fixed constant $\epsilon > 0$ and some $\ell = \mathrm{poly}(n)$ so that the following holds. For any $f : \{0,1\}^n \to \{0,1\}$, if $C^2_{(1-o(1))n}(f) = \omega(1)$, then $f \circ \mathrm{Ind}_m^n$ requires communication $\geq \epsilon n \log \ell$ in the parallel model with $O(1)$ Bobs.*

The above conjecture would imply a lower bound achieving Goals 1/3 for the specific explicit problem $\mathrm{MAJ}^n \circ \mathrm{Ind}_\ell^n$, $\ell \leq \mathrm{poly}(n)$. In this parameter setting, we would have a problem with $N = \ell^n = 2^{O(n \log n)}$ rows and $M = 2^{\mathrm{poly}(n)}$ columns, i.e. M is at most quasipolynomial in N. It would be reasonable to strengthen the conjecture, so that under the same assumption we conclude stronger lower bound $(1 - o(1))n \log m$ on communication; we state it in the weaker form above so as not to be too greedy.

6.3.2 Communication with Alternation and Modular Counting

Here we discuss two more standard complexity measures arising in two party communication complexity which provably capture the parallel-channel model; in principle, good enough lower bounds on either of these measures would be sufficient to achieve Goal 1. Via the natural properties connection in Section 6.2, this also implies that there is a natural proofs barrier to obtaining strong lower bounds for either of these complexity measures. The first measure we examine is a subclass of the communication polynomial hierarchy $\mathsf{PH}^{\mathsf{cc}}$ defined in [BFS86], which captures alternating protocols with two levels of alternation:

Definition 17. *For $F : \{0,1\}^n \times \{0,1\}^m \to \{0,1\}$, we define $\Sigma_2^{\mathsf{cc}}(F)$ as the minimal value $k_1 + k_2$ such that*

$$F(x,y) = \bigvee_{i \leq 2^{k_1}} \bigwedge_{j \leq 2^{k_2}} L_{i,j}(x) \vee R_{i,j}(y)$$

for some $L_{i,j} : \{0,1\}^n \to \{0,1\}$, $R_{i,j} : \{0,1\}^m \to \{0,1\}$

Lemma 14. *If $F : \{0,1\}^n \times \{0,1\}^m$ has (t,k)-parallel protocols then $\Sigma_2^{\mathsf{cc}}(F) \leq k + t + \log t$*

Proof. Say that $F(x,y) = h(g_1(x,y), \ldots, g_t(x,y))$ with g_j having one way communication complexity at most k. So $g_j(x,y) = R_j(\Pi_j(x), y)$ for some $\Pi_j : \{0,1\}^n \to \{0,1\}^k$, $R_j : \{0,1\}^k \times \{0,1\}^n \to \{0,1\}$. Hence

$$F(x,y) = \bigvee_{\xi \in h^{-1}(1)} \bigwedge_{\substack{j \leq t, \\ z \in \{0,1\}^k}} \mathbb{1}\{\Pi_j(x) \neq z\} \vee \mathbb{1}\{R(z,y) = \xi_j\}$$

□

This means that any explicit problem F satisfying $\Sigma_2^{\mathsf{cc}}(F) \geq \Omega(n)$ would automatically achieve Goal 1. This also yields, via Lemma 9, a natural properties barrier for the measure $\Sigma_2^{\mathsf{cc}}(\cdot)$:

Lemma 15. *Assuming the existence of polynomial stretch NC^0 generators, there is no natural proof of a lower bound $\Sigma_2^{\mathsf{cc}}(F) \geq \Omega(n)$ for $F : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$.*

We note that it is a notorious open problem to prove even a lower bound $\Sigma_2^{\mathsf{cc}}(F) \geq \omega(\log n)$ for an explicit problem $F : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$. It is also an open problem to prove lower bounds of the form $2^{\Omega(n)}$ for depth 3 AC^0 circuits computing n variable functions, which lower bounds of the above form would also imply. These facts gives some indication that proving $\Omega(n)$ lower bounds on Σ_2^{cc} complexity may far overshoot Goal 1 in terms of difficulty. We make an additional note that, using basically the same proof as above, it is also possible to strengthen Lemma 14 to general adaptive data structures with arbitrary word length:

Lemma 16 (Strengthening of Lemma 14)**.** *If F has general data structures with time t, wordsize w, and space S, then $\Sigma_2^{\mathsf{cc}}(F) \leq \log S + tw + \log t$*

With this strengthening, it follows that any lower bound $\Sigma_2^{\mathsf{cc}}(F) \geq (3 + \Omega(1))\sqrt{n}$ for a problem $F : \{0,1\}^n \times \{0,1\}^m \to \{0,1\}$ with $m \leq 2^{o(\sqrt{n})}$ would yield a new state of the art bit probe lower bound for general adaptive data structures in the setting $w = 1$, $t = \sqrt{n}$; this is a very different setting then Goal 1 and we do not know of a natural properties barrier which applies in such a regime. Nonetheless the cell probe lower bound would still be of interest in its own right.

Next we discuss a complexity measure associated to a subclass of $(\mathsf{AC}^0[\oplus_{\mathsf{p}}])^{\mathsf{cc}}$:

Definition 18. *Let p be a prime number and $F \in \{0,1\}^{N \times M}$ a boolean matrix. Define* r-rank$_p(F)$*, the "replacement-rank of F over $\mathbb{F}_p$," to be the least rank of a matrix $\tilde{F} \in \mathbb{F}_p^{N \times M}$ such that*

$$F(x,y) = \phi(\tilde{F}(x,y))$$

for some $\phi : \mathbb{F}_p \to \{0,1\}$.

In other words, r-rank$_p(F) \leq r$ if and only if there is a rank $\leq r$ matrix $\tilde{F} \in \mathbb{F}_p^{N \times M}$ such that the value of $F(x,y)$ is completely determined by the value $\tilde{F}(x,y)$.

Lemma 17. *Say that $F \in \{0,1\}^{N \times M}$ has $(O(1), \log S)$ parallel protocols. Then there exists a prime $p \leq 2^{t+2}$ so that* r-rank$_p(F) \leq tS$.

Proof. Say that $F(x,y) = h(g_1(x,y), \ldots, g_t(x,y))$ with $g_j \in \{0,1\}^{N \times M}$ having one way communication complexity $\leq \log S$. Choose $p > 2^{t+1}$ prime and consider each g_j to live in $\mathbb{F}_p^{N \times M}$. Then g_j has at most S distinct rows, and hence its rank is at most S. Define $\tilde{F} \in \mathbb{F}_p^{N \times M}$ by $\tilde{F} = \sum_{j \leq t} 2^j \cdot g_j$. Then $\text{rank}_{\mathbb{F}_p}(\tilde{F}) \leq \sum_j \text{rank}_{\mathbb{F}_p}(g_j) \leq tS$. On the other hand for any $\xi \in \{0,1\}^t$, $\sum_j 2^j \xi_j$ completely determines ξ (here we use that $p > 2^{t+1}$). □

Hence, exhibiting an explicit matrix $F \in \{0,1\}^{N \times M}$ satisfying r-rank$_p(F) \geq N^\epsilon$ for some fixed $\epsilon > 0$ and all primes $p = O(1)$ would solve Goal 1. Currently the best lower bound we know of for r-rank$_p(\cdot)$ is via the inequality r-rank$_p(A) \leq p \cdot \text{rank}(A)^{(p-1)}$ which follows by expressing the predicate $\phi : \mathbb{F}_p \to \{0,1\}$ as a degree $p-1$ polynomial over $\mathbb{F}_p$ using Fermat's little theorem; this never gives a bound better than $N^{\frac{1}{p-1}}$ which is clearly insufficient for our purposes. As in the case of Σ_2^{cc} complexity, combining Lemma 17 with Lemma 9 yields a natural properties barrier for lower bounding r-rank$_p(F)$.

Acknowledgment

The authors would like to give a special thanks to Pravesh Kothari, who, after hearing about a preliminary version of this work, made various suggestions which dramatically simplified and improved our core lower bound results in Section 3.

References

[AGK21] Jackson Abascal, Venkatesan Guruswami, and Pravesh K. Kothari. Strongly refuting all semi-random boolean csps. In *Proceedings of the Thirty-Second Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '21, page 454–472, USA, 2021. Society for Industrial and Applied Mathematics.

[AIK06] Benny Applebaum, Yuval Ishai, and Eyal Kushilevitz. Cryptography in nc^0. *SIAM Journal on Computing*, 36(4):845–888, 2006.

[AOW15] Sarah R Allen, Ryan O'Donnell, and David Witmer. How to refute a random csp. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 689–708. IEEE, 2015.

[App16] Benny Applebaum. Cryptographic hardness of random local functions. *Comput. Complex.*, 25(3):667–722, September 2016.

[BFS86] László Babai, Peter Frankl, and Janos Simon. Complexity classes in communication complexity theory. In *27th Annual Symposium on Foundations of Computer Science (sfcs 1986)*, pages 337–347. IEEE, 1986.

[CFL83] Ashok K Chandra, Merrick L Furst, and Richard J Lipton. Multi-party protocols. In *Proceedings of the fifteenth annual ACM symposium on Theory of computing*, pages 94–99, 1983.

[CM01] Mary Cryan and Peter Bro Miltersen. On pseudorandom generators in nc 0. In *Mathematical Foundations of Computer Science 2001: 26th International Symposium, MFCS 2001 Mariánské Lázne, Czech Republic, August 27–31, 2001 Proceedings 26*, pages 272–284. Springer, 2001.

[DGW19] Zeev Dvir, Alexander Golovnev, and Omri Weinstein. Static data structure lower bounds imply rigidity. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2019, page 967–978, New York, NY, USA, 2019. Association for Computing Machinery.

[dT22] Tommaso d'Orsi and Luca Trevisan. A ihara-bass formula for non-boolean matrices and strong refutations of random csps. *arXiv preprint arXiv:2204.10881*, 2022.

[EF75] Peter Elias and Richard A. Flower. The complexity of some simple retrieval problems. *J. ACM*, 22(3):367–379, July 1975.

[Fei07] Uriel Feige. Refuting smoothed 3cnf formulas. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 407–417. IEEE, 2007.

[FKO06] Uriel Feige, Jeong Han Kim, and Eran Ofek. Witnesses for non-satisfiability of dense random 3cnf formulas. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 497–508, 2006.

[GGNS23] Karthik Gajulapalli, Alexander Golovnev, Satyajeet Nagargoje, and Sidhant Saraogi. Range Avoidance for Constant Depth Circuits: Hardness and Algorithms. In Nicole Megow and Adam Smith, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2023)*, volume 275 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 65:1–65:18, Dagstuhl, Germany, 2023. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

[GGS23] Alexander Golovnev, Tom Gur, and Igor Shinkar. Derandomization of cell sampling. In *Symposium on Simplicity in Algorithms (SOSA)*, pages 278–284, 2023.

[GKM22] Venkatesan Guruswami, Pravesh K. Kothari, and Peter Manohar. Algorithms and certificates for boolean csp refutation: smoothed is no harder than random. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2022, page 678–689, New York, NY, USA, 2022. Association for Computing Machinery.

[GLW22] Venkatesan Guruswami, Xin Lyu, and Xiuhan Wang. Range avoidance for low-depth circuits and connections to pseudorandomness. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2022)*. Schloss-Dagstuhl-Leibniz Zentrum für Informatik, 2022.

[Gol00] Oded Goldreich. Candidate one-way functions based on expander graphs. Cryptology ePrint Archive, Paper 2000/063, 2000.

[GPW20] Mika Göös, Toniann Pitassi, and Thomas Watson. Query-to-communication lifting for bpp. *SIAM Journal on Computing*, 49(4):FOCS17–441, 2020.

[HKM23] Jun-Ting Hsieh, Pravesh K. Kothari, and Sidhanth Mohanty. *A simple and sharper proof of the hypergraph Moore bound*, pages 2324–2344. 2023.

[HR15] Pavel Hrubes and Anup Rao. Circuits with medium fan-in. In *30th Conference on Computational Complexity (CCC 2015)*, pages 381–391. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2015.

[IKOS08] Yuval Ishai, Eyal Kushilevitz, Rafail Ostrovsky, and Amit Sahai. Cryptography with constant computational overhead. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 433–442, 2008.

[JLS21] Aayush Jain, Huijia Lin, and Amit Sahai. Indistinguishability obfuscation from well-founded assumptions. In *Proceedings of the 53rd annual ACM SIGACT symposium on theory of computing*, pages 60–73, 2021.

[JLS22] Aayush Jain, Huijia Lin, and Amit Sahai. Indistinguishability obfuscation from lpn over f p, dlin, and prgs in nc 0. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 670–699. Springer, 2022.

[KKMP21] Robert Kleinberg, Oliver Korten, Daniel Mitropolsky, and Christos Papadimitriou. Total Functions in the Polynomial Hierarchy. In James R. Lee, editor, *12th Innovations in Theoretical Computer Science Conference (ITCS 2021)*, volume 185 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 44:1–44:18, Dagstuhl, Germany, 2021. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

[Kor21] Oliver Korten. The hardest explicit construction. In *62nd Annual Symposium on Foundations of Computer Science*, 2021.

[Lar12] Kasper Green Larsen. Higher cell probe lower bounds for evaluating polynomials. In *Proceedings of the 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, FOCS '12, page 293–301, USA, 2012. IEEE Computer Society.

[LPS97] Hanno Lefmann, Pavel Pudlák, and Petr Savicky. On sparse parity check matrices. *Designs, Codes and Cryptography*, 12(2):107–130, 1997.

[LRT22] Victor Lecomte, Prasanna Ramakrishnan, and Li-Yang Tan. The composition complexity of majority. *arXiv preprint arXiv:2205.02374*, 2022.

[LS07] Nati Linial and Adi Shraibman. Lower bounds in communication complexity based on factorization norms. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 699–708, 2007.

[Man19] Peter Manohar. *New Spectral Techniques in Algorithms, Combinatorics, and Coding Theory: The Kikuchi Matrix Method*. PhD thesis, Weizmann Institute, 2019.

[Mil93] Peter Bro Miltersen. The bit probe complexity measure revisited. In *STACS 93: 10th Annual Symposium on Theoretical Ascpects of Computer Science Würzburg, Germany, February 25–27, 1993 Proceedings 10*, pages 662–671. Springer, 1993.

[Mil94] Peter Bro Miltersen. Lower bounds for union-split-find related problems on random access machines. In *Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing*, STOC '94, page 625–634, New York, NY, USA, 1994. Association for Computing Machinery.

[MNSW98] Peter Bro Miltersen, Noam Nisan, Shmuel Safra, and Avi Wigderson. On data structures and asymmetric communication complexity. *Journal of Computer and System Sciences*, 57(1):37–49, 1998.

[MSS21] Nikhil S Mande, Swagato Sanyal, and Suhail Sherif. One-way communication complexity and non-adaptive decision trees. *arXiv preprint arXiv:2105.01963*, 2021.

[MST03] Elchanan Mossel, Amir Shpilka, and Luca Trevisan. On epsilon-biased generators in nc^ 0. In *Annual Symposium on Foundations of Computer Science*, volume 44, pages 136–145. Citeseer, 2003.

[NN90] J. Naor and M. Naor. Small-bias probability spaces: efficient constructions and applications. In *Proceedings of the Twenty-Second Annual ACM Symposium on Theory of Computing*, STOC '90, page 213–223, New York, NY, USA, 1990. Association for Computing Machinery.

[Pat08] Mihai Patrascu. Succincter. In *2008 49th Annual IEEE Symposium on Foundations of Computer Science*, pages 305–313. IEEE, 2008.

[PT06] Mihai Patrascu and Mikkel Thorup. Time-space trade-offs for predecessor search. In *Symposium on the Theory of Computing*, 2006.

[PTW10] Rina Panigrahy, Kunal Talwar, and Udi Wieder. Lower bounds on near neighbor search via metric expansion. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 805–814. IEEE, 2010.

[Raz88] Alexander Razborov. Bounded-depth formula over {AND,XOR} and some combinatorial problems (russian). *Problems of Cybernetics. Complexity Theory and Applied Mathematical Logic*, pages 149–166, 1988.

[RR97] Alexander A Razborov and Steven Rudich. Natural proofs. *Journal of Computer and System Sciences*, 55(1):24–35, 1997.

[RSW22] Hanlin Ren, Rahul Santhanam, and Zhikun Wang. On the range avoidance problem for circuits. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 640–650, 2022.

[Sav95] Petr Savickỳ. Improved boolean formulas for the ramsey graphs. *Random Structures & Algorithms*, 6(4):407–415, 1995.

[Sie89] A. Siegel. On universal classes of fast high performance hash functions, their time-space tradeoff, and their applications. In *30th Annual Symposium on Foundations of Computer Science*, pages 20–25, 1989.

[Sie04] Alan Siegel. On universal classes of extremely random constant-time hash functions. *SIAM J. Comput.*, 33(3):505–543, March 2004.

[Tro12] Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434, August 2012.

[Vio12] Emanuele Viola. The complexity of distributions. *SIAM Journal on Computing*, 41(1):191–218, 2012.

[Vio19] Emanuele Viola. Lower bounds for data structures with space close to maximum imply circuit lower bounds. *Theory of Computing*, 15(1):1–9, 2019.

[WEAM19] Alexander S Wein, Ahmed El Alaoui, and Cristopher Moore. The kikuchi hierarchy and tensor pca. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1446–1468. IEEE, 2019.

[Wit17] David Witmer. *Refutation of random constraint satisfaction problems using the sum of squares proof system*. PhD thesis, PhD thesis, Carnegie Mellon University, 2017.

[Yao81] Andrew Chi-Chih Yao. Should tables be sorted? *J. ACM*, 28(3):615–628, July 1981.

A Kikuchi Method

We prove here Theorem 8 from Section 3. This argument occurs in [WEAM19, GKM22] and we only need to adapt things to our notation:

Theorem 18 (Theorem 8)**.** *Let $\mathcal{C} = ([N], V, c)$ be a t-XOR scheme, $|V| = K$, and t even. For any $t \le L \le \frac{K}{8}$, we may associate with $\mathcal{C}$ a permutation ensemble $\mathcal{A}$ indexed by $[N]$ with the following properties:*

1. *$\mathcal{A}$ lives in $\mathbb{R}^{m\times m}$ for $m = \binom{K}{L}$*
2. *$d(\mathcal{A}) \ge \left(\frac{L}{K}\right)^{t/2} N$*
3. *For any $f \in \{\pm 1\}^N$, we have $\|\mathcal{A}[f]\|_{\infty\to 1} \ge d(\mathcal{A}) m \cdot \mathrm{OPT}(\mathcal{C}, f)$*

Proof. Let $\mathcal{C} = ([N], V, c)$ be t-XOR scheme, $|V| = K$, t even. Let $m = \binom{K}{L}$, and identify $[m]$ with $\binom{V}{L}$ canonically. We define a permutation ensemble in $\mathbb{R}^{m\times m}$, given by $A_x(a,b) = \mathbb{1}\{a \Delta b = c_x\}$ where $a\Delta b$ denotes the symmetric difference of a, b as elements of $\binom{X}{L}$. Observe that by symmetry, $\sum_{a,b} A_x(a,b)$ is the same value for all x; call this value r. Indeed we have $r = 2\binom{t}{t/2}\binom{K-t}{L-t/2}$. To see this, note that if if $|a| = |b| = L$, $a\Delta b = c$, $|c| = t$ then we must have $a = (w \cup u), b = (w \cup v)$ for some $|w| = L - t$, $|u| = |v| = \frac{t}{2}$, $u, v \subseteq c$. There are $\binom{t}{t/2}$ ways to choose disjoint subsets $u, v \subseteq c_x$, $|u| = |v| = \frac{t}{2}$, and $\binom{K-t}{L-\frac{t}{2}}$ ways to choose $w \in \binom{V\setminus\{u\cup v\}}{L-t/2}$. Since we will count $(a,b), (b,a)$ separately we must add in a factor of 2.

Now, let $f \in \{\pm 1\}^N$ and let $R : V \to \{\pm 1\}$ be such that $\mathrm{OPT}(\mathcal{C}, f) = \mathrm{Val}(\mathcal{C}, f, R)$. Define $\tilde{R} \in \{\pm 1\}^m = \{\pm 1\}^{\binom{V}{L}}$, with $\tilde{R}(a) = \prod_{i\in a} R(i)$. Then

$$\tilde{R}^\top \mathcal{A}[f]\tilde{R} = \sum_{a,b\in\binom{V}{L}} \tilde{R}(a)\tilde{R}(b)\mathcal{A}[f](a,b) = \sum_x f_x \sum_{a,b} A_x(a,b) \prod_{i\in a} R(i) \prod_{j\in b} R(j)$$
$$= \sum_{x\in X} \Big(\sum_{a,b} A_x(a,b)\Big) f(x) \prod_{i\in c_x} R(i) = r \sum_x f(x) \prod_{i\in c_x} R(i) = rN \cdot \mathrm{OPT}(\mathcal{C}, f)$$

Now, observe that $d(\mathcal{A}) = Nm^{-1}r$, in other words $r = d(\mathcal{A})N^{-1}m$, so we have the bound $\|\mathcal{A}[f]\|_{\infty\to 1} \ge d(\mathcal{A}) m \cdot \mathrm{OPT}(\mathcal{C}, f)$. It remains to lower bound $d(\mathcal{A})$, for which we observe

$$d(\mathcal{A}) = Nm^{-1}r \ge \frac{2\binom{t}{t/2}\binom{K-t}{L-t/2}}{\binom{K}{L}} N \ge \left(\frac{L}{K}\right)^{t/2} N$$

provided $t \le L \le \frac{K}{8}$, where the final inequality follows from standard manipulations of binomial coefficients (see [HKM23]). □

ECCC ISSN 1433-8092
https://eccc.weizmann.ac.il