Candidate Tree Codes via Pascal Determinant Cubes

Inbar Ben Yaacov
The Blavatnik School of Computer Science, Tel-Aviv University, Israel

Gil Cohen
The Blavatnik School of Computer Science, Tel-Aviv University, Israel

Anand Kumar Narayanan
CISPA Helmholtz Center for Information Security, Saarbrücken, Germany

Abstract

Tree codes are combinatorial structures introduced by Schulman [23] as key ingredients in interactive coding schemes. Asymptotically-good tree codes are long known to exist, yet their explicit construction remains a notoriously hard open problem. Even proposing a plausible construction, without the burden of proof, is difficult and the defining tree code property requires structure that remains elusive. To the best of our knowledge, only one candidate appears in the literature, due to Moore and Schulman [19].

We put forth a new candidate for an explicit asymptotically-good tree code. Our construction is an extension of the vanishing rate tree code by Cohen-Haeupler-Schulman [7], and its correctness relies on a conjecture that we introduce on certain Pascal determinants indexed by the points of the Boolean hypercube. Furthermore, using the vanishing distance tree code by Gelles *et al.* [12] enables us to present a construction that relies on an even weaker assumption. We furnish evidence supporting our conjecture through numerical computation, combinatorial arguments from planar path graphs and based on well-studied heuristics from arithmetic geometry.

2012 ACM Subject Classification Theory of computation → Error-correcting codes

Keywords and phrases Tree codes, Sparse polynomials, Explicit constructions

Digital Object Identifier 10.4230/LIPIcs.APPROX/RANDOM.2021.54

Category RANDOM

Related Version *Extended Version*: `https://eccc.weizmann.ac.il/report/2020/141/`

Funding *Inbar Ben Yaacov*: Funded by the Israel Science Foundation (grant number 1569/18).
Gil Cohen: Funded by the Israel Science Foundation (grant number 1569/18) and by the Azrieli Faculty Fellowship.
Anand Kumar Narayanan: Supported by the European Union's H2020 Programme (grant agreement #ERC-669891).

Acknowledgements The second author wishes to thank Roni Con, Shir Peleg-Schatzman, Noam Peri, Tal Roth, and Shahar Samocha for interesting discussions on tree codes.

1 Introduction

Coding theory addresses the problem of communication over an imperfect channel. In the classic setting studied in the seminal work of Shannon [26], Alice wishes to communicate a message to Bob over a channel that may induce errors. The question then is: how should Alice encode her message so that if the amount of errors is not excessive, Bob can recover her message? Around the same time, Hamming [14] introduced the notion of an error-correcting code. A function $C\colon \Sigma^k \to \Sigma^n$ is an *error-correcting code* with distance δ if for every distinct $x, y \in \Sigma^k$, the respective images $C(x), C(y)$ have relative Hamming distance at least δ. The

rate of information transmission $\rho = \frac{k}{n}$ and the fraction of errors corrected (roughly $\delta/2$) are competing quantities with a tradeoff between them. Among the most basic questions in coding theory is to obtain explicit *asymptotically good codes*, that is, codes over fixed Σ with constant distance $\delta > 0$ and constant rate $\rho > 0$. By "explicit" we mean that C can be evaluated in time $\text{poly}(n)$. Justensen [15] was the first to devise such an explicit construction. Since then, several explicit constructions have appeared, including using algebraic geometry codes [28] and expander graphs [27].

While error-correcting codes can be used to solve the problem of sending a single message from Alice to Bob over an imperfect channel, in some settings, the two parties interact with each other, sending multiple messages where a message depends on previous messages that were exchanged. Interactive coding addresses the subtler problem of enabling such dynamic interaction over an imperfect channel. In this far more challenging setting, standard codes do not offer a satisfactory solution.

Tree codes are powerful combinatorial structures, defined by Schulman [23, 25] as key ingredients for achieving interactive coding schemes. They play a role analogous to that error-correcting codes take in the single message setting. Tree codes, as their name suggests, are trees with certain distance properties. To give the formal definition, we set some notation. Let T be a rooted binary tree that is endowed with an edge coloring from some ambient color set (or alphabet) Σ. For vertices u, v of equal depth let w be their least common ancestor and denote the distance, in edges, from u to w by ℓ. Let $p_u, p_v \in \Sigma^\ell$ be the sequences of colors on the path from w to u and to v, respectively. We define $h(u, v)$ to be the relative Hamming distance between p_u and p_v. Informally, $h(u, v)$ measures the distance between the two color sequences obtained by following the paths from the root to each of u and v, excluding the "non-interesting" common prefix. A tree code is any coloring that has a lower bound on this quantity. Formally,

▶ **Definition 1** (Tree codes [23]). *Let T be the complete rooted binary tree of depth n. The tree T, together with an edge-coloring of T by a color set Σ is called a* tree code with distance *δ if for every pair of vertices u, v with equal depth it holds that $h(u, v) \geq \delta$.*

It is not clear at all that there exists a universal constant $\delta > 0$ such that for every n there exists a depth-n tree code with distance δ. Namely, it is not clear that there is a family of tree codes $(T_n)_{n\in\mathbb{N}}$, where T_n has depth n, such that the color set Σ is common to all trees in the family, and every T_n has distance δ. We refer to such a family as a tree code with distance δ over the color set Σ.

Three different proofs were provided by Schulman, showing that for any constant $\delta < 1$ there exists a tree code with alphabet size $|\Sigma| = O_\delta(1)$ achieving distance δ. More recently, based on Schulman's ideas, it was shown that there is a tree code with only 4 colors, having positive distance (in particular, distance $\delta = 0.136$) [8] and, moreover, 3 colors do not suffice to guarantee any constant distance $\delta > 0$. All of these proofs rely on the probabilistic method and thus are not explicit. The problem of constructing asymptotically-good tree codes has drawn substantial attention [24, 6, 19, 21, 12, 7, 20], but has endured as a difficult challenge.

Given this difficulty, it is natural to construct, for a given distance parameter $\delta > 0$, a family of tree codes $(T_n)_{n\in\mathbb{N}}$ for which T_n is allowed to use some $c(n)$ number of colors. The goal is to obtain an asymptotically slowly-growing function c. Note that constructing a tree code family with $c(n) = 2^n$ colors is trivial. Indeed, having so many colors at hand, one can encode the entire path leading to a vertex on the edge preceding it, yielding distance $\delta = 1$. In an unpublished manuscript, Evans, Klugerman and Schulman [24] constructed a tree code with $c(n) = n^{O_\delta(1)}$ colors. The state-of-the-art construction [7] achieves $c(n) = (\log n)^{O_\delta(1)}$. See [20] for alternative constructions achieving the same parameters as well as decoding algorithms, and [4] for an account relating [7] and [21].

Despite this progress, constructing asymptotically-good tree codes is wide open. Curiously, even candidate constructions are rare. This is mostly because a tree code is *not* a pseudorandom object. Its defining property requires structure that remains elusive. For this reason, even proposing a plausible construction, without the burden of proof, requires further insight and is not an easy task. To the best of our knowledge, there is a single candidate in the literature, due to Moore and Schulman [19]. The construction's distance property relies on an intriguing open conjecture about certain exponential sums that the authors introduce. The Moore-Schulman conjecture was verified computationally for small instances, and the hope is that these represent the general case.

1.1 Our Contribution

In this work we put forth a candidate construction of asymptotically-good tree codes. Namely, for some universal constant $c \geq 1$ and for every integer $n \geq 1$ we give an explicit construction of a depth-n binary tree code with c colors. The distance of the tree code is bounded below by some constant $\delta > 0$, independent of n, provided a conjecture that we introduce on certain Pascal determinants associated with the points of the Boolean hypercube holds. We give independent supporting evidence for our conjecture: first through the combinatorics of planar path graphs underlying our construction and then based on well-studied heuristics from arithmetic geometry. Furthermore, we verify the conjecture computationally on small values.

Our candidate tree code is an extension of the [7] construction. We set the stage in Section 2 with a discussion of [7] followed by a description of our contributions in Section 3. Underlying the [7] construction is a key online uncertainty principle for the Newton basis: a consequence of non-vanishing of Pascal (binomial) sub-matrix determinants, proved by invoking the combinatorial Lindström-Gessel-Viennot lemma. These determinants are in fact positive numbers growing exponentially with the depth of the tree, forcing the [7] construction to require poly-logarithmic number of colors. With the intent of reducing the number of colors, one may try to work modulo a prime in hopes the non-vanishing is still preserved. In Section 3.1 we reason the contrary is true: it is unlikely to work for primes small enough to guarantee a constant number of colors. There are exponentially many Pascal sub-matrix determinants, at least one of which is likely to vanish "accidentally" modulo the chosen prime.

Our main technical contribution is an extension of the [7] construction, which we present as a candidate asymptotically-good tree code. The construction extends ideas of [7] and further makes use of the vanishing-distance tree code by Gelles *et al.* [12], which allows us to relax our assumption. An informal description of the main ideas is in Section 3.3 with a formal treatment of the more intricate aspects deferred to Section 5. In our construction, the role of each Pascal sub-matrix determinant is recast as a bundle of Pascal sub-matrix determinants, parametrized by points on the Boolean hypercube of high enough dimension (hence the term "Pascal determinant cube" in the title). We then work modulo a prime p of appropriate size. Instead of worrying about a determinant vanishing modulo p, we only have to worry about the whole associated cube of Pascal determinants vanishing modulo p. Informed by computation, combinatorics and arithmetic, we formulate the Conjecture 4 in Section 3.2 that the cube of determinants never vanishes modulo our chosen prime. We prove that if the conjecture (or even an asymptotic version of it, Conjecture 5) holds then our construction is indeed asymptotically good.

In Appendix A, we investigate our conjecture through a combinatorial lens. Each determinant bundle in the conjecture can be encoded as an integer polynomial whose evaluation at the points of the Boolean hypercube gives the bundle. Through the Lindström-

Gessel-Viennot lemma, in Appendix A.1 we prove that the polynomial never vanishes on any point of the Boolean hypercube. For the conjecture to fail, all these exponentially many evaluations must be divisible by our chosen prime number, which we reason is likely impossible for our chosen parameters. This very scenario is reformulated in terms of Boolean functions in Appendix A.2, by multi-linearizing the aforementioned polynomial. Conjecture 4 is then rephrased as the non vanishing of an $\mathbb{F}_p$-valued Boolean function, furthering our belief in the conjecture.

In Appendix B, we look to deep results from arithmetic geometry to claim the plausibility of our conjecture. If the hypersurface of zeroes of the aforementioned polynomial encoding the bundle of determinants intersects with the Boolean hypercube generically, our conjecture holds true. Following Fouvry [10], we investigate this intersection deploying Katz-Laumon exponential sums. The bounds on Katz-Laumon sums and Fouvry's point counting technique fall short of quantitatively proving our conjecture. Yet, we show they suffice to prove a nontrivial relaxation of our conjecture: with the Boolean hypercube extended to hypercubes of side length $\approx p^{3/4}$. Despite falling short of proving our conjecture, the methods are illuminating and suggest there are no arithmetic obstructions to our conjecture.

1.2 Recent Developments

Since posting our report online, there have been several exciting developments. Brakerski, Kalai and Saxena [5] published a major advancement on the use of tree codes in interactive coding. They demonstrate how a tree code with an efficient encoding algorithm suffices in obtaining efficient and deterministic interactive coding schemes against adversarial errors. In particular, they completely eliminate the necessity of the tree code possessing an efficient decoding algorithm. Our candidate tree code clearly has an efficient encoding algorithm and seamlessly fits their needs. Therefore, proving our tree codes are indeed asymptotically good would immediately imply efficient and deterministic interactive coding schemes.

Pudlák gleaned an abstraction of our construction and reduced the problem of constructing asymptotically good tree codes to constructing block matrices of the following form [22]. Consider an n by n block matrix whose entries are 2 by 1 column vector blocks. Say it is triangular, meaning all blocks above the diagonal are $\binom{0}{0}$ vectors. For every k by k block sub matrix (where the sorted column indices are never ahead of the row indices), Pudlák demands that the $2k$ by k matrix induced by forgetting the block structure is full rank. This rank criterion is a relaxation of our determinant bundle non vanishing. Pudlák further proves that random triangular block matrices with entries from a finite field of size quadratic in n satisfy the rank criterion with high probability. Explicit deterministic construction of such block matrices beckons, with our construction being the only currently proposed candidate.

2 Cohen-Haeupler-Schulman Tree Codes

For the sequel, it is convenient to think of a tree code as an online version of a regular error correcting code. Recall that a tree code consists of a complete rooted, depth-n binary tree in which each edge is labeled by a symbol from an alphabet Σ. This naturally induces a one-to-one mapping assigning each binary string s to a path starting at the root, where s indicates which child is taken in each of the steps. Such a path maps to a string over Σ, namely, the concatenation of symbols along the path. This way, a tree code T encodes any binary string s into an equally long string $T(s)$ over Σ. This encoding has the online property because the encoding of any prefix does not depend on later symbols. Thus, one

can view a binary tree code as an online function $T : \{0,1\}^n \to \Sigma^n$. It is useful to consider input alphabets other than binary (which corresponds to a larger arity of the tree). In [7], the input symbols are elements of $\mathbb{Z}$ rather than $\{0,1\}$.

The distance property of a tree code can be phrased as follows when viewed as a function $T : \Sigma_{\mathsf{in}}^n \to \Sigma_{\mathsf{out}}^n$. For every pair of distinct strings $m = (m_0, \ldots, m_{n-1})$, $m' = (m'_0, \ldots, m'_{n-1})$, c being the least integer such that $m_c \neq m'_c$, the following holds. For every $\ell \in [0, n-c)$ (for integers $a < b$ we write $[a,b)$ for $\{a, a+1, \ldots, b-1\}$) the strings $(T(m)_c, \ldots, T(m)_{c+\ell})$, $(T(m')_c, \ldots, T(m')_{c+\ell})$ are at Hamming distance at least $\delta(\ell+1)$.

The Newton basis. [7] makes use of the Newton basis for real polynomials. This basis consists of polynomials of the form $\binom{x}{k} \in \mathbb{R}[x]$ for $k \in \mathbb{N}$, where $\binom{x}{k} = \frac{x(x-1)\cdots(x-(k-1))}{k!}$. It is easy to verify that for every $d \in \mathbb{N}$, the set $\{\binom{x}{k} \mid k = 0, 1, \ldots, d\}$ forms a basis for the space of univariate real polynomials of degree at most d. The feature which makes the Newton basis suitable for constructing tree codes unlike, say the standard basis, is its online nature with respect to $\mathbb{N}$. Let $m_0, \ldots, m_t \in \mathbb{R}$. Let $f(x) = \sum_{i=0}^{t} a_i x^i$ be the least degree polynomial that interpolates on the points $(0, m_0), \ldots, (t, m_t)$. Then, generally, given a new point $(t+1, m_{t+1})$, the least degree polynomial, $g(x) = \sum_{i=0}^{t+1} b_i x^i$, that interpolates on $(0, m_0), \ldots, (t+1, m_{t+1})$ will have a completely different sequence of coefficients (i.e., $a_i \neq b_i$). By contrast, using the Newton basis, the coefficients that were already "recorded" stay intact given the new point $(t+1, m_{t+1})$. More precisely, if $f(x) = \sum_{i=0}^{t} \gamma_i \binom{x}{i}$ then $g(x) = f(x) + \gamma_{t+1}\binom{x}{t+1}$ for some $\gamma_{t+1} \in \mathbb{R}$. Thus, for every t, the coefficient γ_t is determined by $m_0, m_1, \ldots, m_t$. Another convenient property of the Newton basis, not shared by the standard basis, is that if $m_0, \ldots, m_t$ are all integers, so are the coefficients $\gamma_0, \ldots, \gamma_t$.

The [7] tree code over the integers. In [7], for every integer $n \geq 1$ a function $\mathsf{TC}_{\mathbb{Z}} \colon \mathbb{Z}^n \to (\mathbb{Z} \times \mathbb{Z})^n$ is constructed as follows. Given $m = (m_0, \ldots, m_{n-1}) \in \mathbb{Z}^n$, let $f \in \mathbb{R}[T]$ be the least degree real polynomial that interpolates on $(0, m_0), \ldots, (n-1, m_{n-1})$. Expand f in the Newton basis $f(T) = \sum_{t=0}^{n-1} \gamma_t \binom{T}{t}$. With this notation, for every $t \in [0, n)$, define $\mathsf{TC}_{\mathbb{Z}}(m)_t = (m_t, \gamma_t)$. In words, at time t, both the t^{th} input symbol is outputted as well as the "new" coefficient γ_t.

Analysis. To argue about the distance of $\mathsf{TC}_{\mathbb{Z}}$, using the fact that it is $\mathbb{R}$-linear, one has to prove that if $c \in [0, n)$ is the least integer for which $m_c \neq 0$ then for every $\ell \in [0, n-c)$, at least δ-fraction of the indices in $[c, c+\ell]$ satisfies that $\mathsf{TC}_{\mathbb{Z}}(m)_t$ is nonzero (as a pair). If we write, for $d \in [0, n)$, $f_d(T) = \sum_{t=0}^{d} \gamma_t \binom{T}{t}$ then the number of non-zeros in the sequence $\gamma_c, \gamma_{c+1}, \ldots, \gamma_{c+\ell}$ is precisely the sparsity of $f_{c+\ell}$ in the Newton basis. This, together with the fact that for every $i \leq t$, $m_i = f_t(i)$, implies that to "break" the construction $\mathsf{TC}_{\mathbb{Z}}$, one must come up with a sparse polynomial $f_{c+\ell}$, with respect to the Newton basis, that has many roots in $I = \{c, c+1, \ldots, c+\ell\}$. Indeed, if $f_{c+\ell}$ is not sparse, then many of the γ-entries of $(\mathsf{TC}_{\mathbb{Z}}(m)_t)_{t \in I}$ will be nonzero. On the other hand, if $f_{c+\ell}$ has only few roots in I then many of the m-entries are nonzero. To this end, the main lemma proved in [7] is a bound on the numbers of distinct integral roots a real polynomial can have as a function of its sparsity in the Newton basis.

▶ **Lemma 2** ([7]). *Let $f \in \mathbb{R}[T]$ be a nonzero polynomial of sparsity $s \geq 1$ in the Newton basis. Let $c \geq 0$ be the least integer such that $f(c) \neq 0$. Then, f has at most $s-1$ distinct roots in $[c, \infty) \cap \mathbb{N}$.*

Lemma 2 implies that if the sparsity of $f_{c+\ell}$ is s then there can be at most $s-1$ zeros among the m-entries of $\{\mathsf{TC}_{\mathbb{Z}}(m)_t\}_{t \in I}$, establishing $\mathsf{TC}_{\mathbb{Z}}$ has distance at least $\frac{1}{2}$.

The Lindström-Gessel-Viennot Lemma. Lemma 2 is proved using a corollary of the Lindström-Gessel-Viennot Lemma. Let $\mathsf{t} = (t_1, \ldots, t_s)$, $\mathsf{c} = (c_1, \ldots, c_s)$ be strictly increasing sequences of non-negative integers. Let $M_{\mathsf{t},\mathsf{c}}$ be the $s \times s$ matrix whose $(i,j)^{\text{th}}$ entry is given by $\binom{t_i}{c_j}$. We write $\mathsf{c} \leq \mathsf{t}$ if $c_i \leq t_i$ for every $i \in [s]$.

▶ **Lemma 3** ([13], Corollary 2). $\mathsf{c} \leq \mathsf{t} \iff \det M_{\mathsf{t},\mathsf{c}} \neq 0$.

For more recent treatments of the LGV Lemma see [1], Chapter 5.4 or [2], Chapter 25. This lemma is in fact much older, and we invite the reader to look at the appendix of [7] for more information regarding the history of this lemma.

The binary tree code. To reduce the alphabet to binary, [7] proves that if for every t, $|m_t| \leq 2^k$ for some k then $|\gamma_t| \leq 2^{t+k}$. Given a binary string $m = (m_0, \ldots, m_{n-1})$, partition m to $\sqrt{n}$ consecutive blocks of length $\sqrt{n}$, and interpret each block as a non-negative integer M_i of size at most $2^{\sqrt{n}}$. At this point, the tree code over the integers $\mathsf{TC}_{\mathbb{Z}} : \mathbb{N}^{\sqrt{n}} \to (\mathbb{Z} \times \mathbb{Z})^{\sqrt{n}}$ can be applied to $M_0, \ldots, M_{\sqrt{n}-1}$. By the above bound, $|\gamma_t| \leq 2^{t+\sqrt{n}} \leq 2^{2\sqrt{n}}$. Hence, an output symbol (m_t, γ_t) can be encoded using $3\sqrt{n}$ bits. Of course, these bits cannot be output on the fly as one must write a symbol only after all of the $\sqrt{n}$ bits of the corresponding input symbol have been read. This creates a "lag" of length $\sqrt{n}$ that can be resolved by using a depth-$\sqrt{n}$ tree code which is obtained recursively. As the recursive depth is $O(\log \log n)$ and since for every bit read one writes $O(1)$ bits per recursive call, the number of bits written per input bit is $O(\log \log n)$. Hence, the $\text{poly}(\log n)$ alphabet size.

3 Our Contribution

3.1 The Unlikeliness of an LGV-Like Lemma Over Small Fields

The reason that the [7] construction is not asymptotically-good is that their tree code is constructed over the integers, and the alphabet reduction that is invoked has a cost that is exponential in the depth of the recursion. The recursion's depth is directly affected by the magnitude of the γ_t symbols which, unfortunately, are exponential in t. Taking $\sqrt{n}$-length blocks yields the best trade-off, resulting in depth $O(\log \log n)$.

One can show that resorting to such recursion could have been avoided if the construction was carried over a prime field $\mathbb{F}_p$ with $p = \text{poly}(n)$. That is, instead of outputting γ_t, output its reduction modulo p. To be precise, for the construction to work, one must take $p \geq n$ due to other considerations. However, as long as $p < n^e$ for some constant e, standard techniques can be used to obtain an asymptotically-good binary tree code, where the constant e will affect the rate of the resulted tree.

A very similar approach to this was raised by Pudlák [21]. On this, we quote a sentence from the conclusion part of [21]: "This seems to be a very difficult problem and we do not dare to conjecture that p may be of polynomial size". At this point, Pudlák suggests studying restricted cases for which small fields suffice and try to base tree code constructions on such results, but we digress.

In consensus with Pudlák, we too believe that the approach of working over $\mathbb{F}_p$ as suggested above is not likely to work. That is, it seems very plausible to us that the LGV Lemma does not have an analog over a field of size $\text{poly}(n)$. More precisely, we suspect that for every constant $e \geq 1$, there exists $n_0 = n_0(e)$ such that for every $n \geq n_0$ and $p \leq n^e$, there exists a pair $\mathsf{t}, \mathsf{c} \in [0, n)^s$, for some $s \in [n]$, satisfying $\mathsf{c} \leq \mathsf{t}$, such that $\det M_{\mathsf{t},\mathsf{c}} \equiv_p 0$.

To get some intuition as to why we believe this is the case, fix some prime p and $s \in [n]$. There are between $\binom{n}{s}$ and $\binom{n}{s}^2$ pairs of sequences t, c to consider. Unless some structure is present, one would expect that roughly $\frac{1}{p}$-fraction of pairs t, c would satisfy $p \mid \det M_{\mathsf{t},\mathsf{c}}$.

By that heuristic, we do not expect that p can be taken much smaller than $\binom{n}{s}$. As we are interested in s that can be as large as $\Omega(n)$, this heuristic points against the existence of a "good" prime $p = 2^{o(n)}$, let alone $p = \text{poly}(n)$.

This heuristic is supported by a computational search that we carried. Let $\mathcal{P}_1 : \mathbb{N} \to \mathbb{N}$ be the function that maps $n \in \mathbb{N}$ to the least prime p that satisfies the following property. For every $s \in [n]$ and strictly increasing sequences $\mathsf{c} = (c_1, \ldots, c_s), \mathsf{t} = (t_1, \ldots, t_s) \in [0, n)^s$ with $\mathsf{c} \leq \mathsf{t}$ it holds that $\det M_{\mathsf{t},\mathsf{c}} \not\equiv_p 0$. Informally, $\mathcal{P}_1$ maps n to the smallest prime p that is "good" for n. An exhaustive search we have conducted for hundreds of computer hours seems to suggest that $\mathcal{P}_1(n)$ grows exponentially with n.

Table 1 Values of $\mathcal{P}_1(n)$ obtained using a computer search.

n	6	7	8	9	10	11	12	13	14	15	16
$\mathcal{P}_1(n)$	13	17	47	89	241	641	2,687	6,521	15,401	74,257	$> 250,000$

Since posting our preprint online, Karingula and Lovett [16] consider the non singularity of submatrices of triangular matrices modulo p in a different context. They too arrive at our conclusion, in fact, conjecturing a stronger claim ([16], Conjecture 1.5): for every triangular integer matrix, a fraction of the determinants (corresponding to index sequences t, c, as above) are likely to vanish modulo p unless the field size p grows exponentially.

3.2 A Conjecture

The informal heuristic presented above makes the point that no poly(n)-size prime is likely to work against all $\exp(n)$ many pairs of sequences as we have no evidence for a structural phenomenon to support the seemingly unlikely alternative. The main contribution of this work is a tree code construction–a variant of [7]–whose distance analysis relies on what we believe is a plausible statement which we put forth as a conjecture. To formally state our conjecture some preparation is required.

As before, let $\mathsf{c} = (c_1, \ldots, c_s), \mathsf{t} = (t_1, \ldots, t_s) \in [0, n)^s$ be a pair of strictly increasing sequences with $\mathsf{c} \leq \mathsf{t}$. For symbolic variables $X_1, \ldots, X_s$, define the $s \times s$ (symbolic) matrix $M_{\mathsf{t},\mathsf{c}}(X_1, \ldots, X_s)$ whose $(i,j)^{\text{th}}$ entry is given by $\binom{X_i + t_i}{c_j}$. Define $\Phi_{\mathsf{t},\mathsf{c}}(X_1, \ldots, X_s) \triangleq \det M_{\mathsf{t},\mathsf{c}}(X_1, \ldots, X_s) \in \mathbb{Z}[X_1, \ldots, X_s]$. For a prime p, let $\Phi^p_{\mathsf{t},\mathsf{c}}(X_1, \ldots, X_s) \in \mathbb{F}_p[X_1, \ldots, X_s]$ denote the reduction of $\Phi_{\mathsf{t},\mathsf{c}}$ at p. That is, every coefficient of $\Phi_{\mathsf{t},\mathsf{c}}$ is taken modulo p to form $\Phi^p_{\mathsf{t},\mathsf{c}}$. With this notation, to ensure that the [7] tree code works over $\mathbb{F}_p$, one must establish that $\Phi^p_{\mathsf{t},\mathsf{c}}(0, \ldots, 0) \neq 0$ for all t, c in question. Put differently, the [7] construction fails if for some pair t, c as above, $\Phi^p_{\mathsf{t},\mathsf{c}}$ evaluates to 0 at the origin. Our main contribution is an explicit construction which fails only if $\Phi^p_{\mathsf{t},\mathsf{c}}$ evaluates to 0 on the *entire* Boolean hypercube $\{0,1\}^s$. Equivalently, our construction is asymptotically-good if

$$\exists (x_1, \ldots, x_s) \in \{0,1\}^s \quad \Phi_{\mathsf{t},\mathsf{c}}(x_1, \ldots, x_s) \not\equiv_p 0. \tag{3.1}$$

3.2.1 Preliminary Informal Discussion on the Plausibility of Equation (3.1)

To start with, consider a very informal point of view on the plausibility of Equation (3.1), a discussion similar in spirit to the one conducted for arguing against the plausibility of taking the [7] construction over $\mathbb{F}_p$. Heuristically, and very informally, one may think of the 2^s conditions in Equation (3.1) as 2^s trials that are "generated by s independent

random variables" $X_1, \ldots, X_s$. Unless some structural obstruction is in place, the "event" in Equation (3.1) is expected to have probability of about p^{-s}. Continuing this informal line of reasoning, by a union bound, one would expect that for a choice of p satisfying $p^{-s}\binom{n}{s}^2 \ll \frac{1}{n}$, Equation (3.1) holds for every pair $\mathsf{t}, \mathsf{c} \in [0, n)^s$, for every $s \in [n]$. The latter holds by taking $p \gg n^3$.

Another informal argument supporting the validity of Equation (3.1) is as follows. Note that $\Phi_{\mathsf{t},\mathsf{c}}$ has total degree $d \leq sn \leq n^2$. In fact, as we only care about $\Phi_{\mathsf{t},\mathsf{c}}$ restricted to $\{0,1\}^s$, we may assume that $\Phi_{\mathsf{t},\mathsf{c}}$ is multi-linear and so $d \leq s \leq n$. One can show that for $p > n$, $\Phi^p_{\mathsf{t},\mathsf{c}}$ is a nonzero polynomial; thus, by Schwartz-Zippel, $\Phi^p_{\mathsf{t},\mathsf{c}}$ has at most $\frac{d}{p} \leq \frac{n}{p}$ fraction of roots in $\mathbb{F}_p^s$. By taking, say, $p \geq n^2$, the roots of $\Phi_{\mathsf{t},\mathsf{c}}$ occupy at most $\frac{1}{\sqrt{p}}$-fraction of $\mathbb{F}_p^s$. Now, for the heuristic part, one may conjecture that $\{0,1\}^s$ "looks random" to the zero set $V_{\mathsf{t},\mathsf{c}}$ of $\Phi_{\mathsf{t},\mathsf{c}}$. As a weak consequence, $\{0,1\}^s$ is not contained in $V_{\mathsf{t},\mathsf{c}}$, which is the content of Equation (3.1).

3.2.2 The Conjecture

There is one small technical issue we need to address before presenting our formal conjecture. Note that if $t_{i+1} = t_i + 1$ for some i then $\Phi_{\mathsf{t},\mathsf{c}}(x_1, \ldots, x_s) = 0$ whenever $x_i = 1$ and $x_{i+1} = 0$ for the simple reason that two of the rows of $M_{\mathsf{t},\mathsf{c}}(x_1, \ldots, x_s)$ are identical. Informally, from the heuristic point of view discussed above, when $t_{i+1} = t_i + 1$, the events associated with the variables X_i, X_{i+1} are dependent. To exclude these trivial roots of $\Phi_{\mathsf{t},\mathsf{c}}(x_1, \ldots, x_s)$ we assume in the conjecture (and guarantee in the construction) that t, c only have even entries. In Appendix A.1 we prove that, having done so, $\Phi_{\mathsf{t},\mathsf{c}}$ has no root in $\{0,1\}^s$. That is, when considering t, c with even entries, $\Phi_{\mathsf{t},\mathsf{c}}(x_1, \ldots, x_s) \neq 0$ for every $(x_1, \ldots, x_s) \in \{0,1\}^s$, and so it is only the reduction modulo p that may yield roots. With this, we are finally ready to state our conjecture.

▶ **Conjecture 4** (The Pascal determinant cubes (PDC) conjecture)**.** *There exists a universal constant $e_p \geq 1$ such that for every integer $n \geq 1$ and prime $p \geq n^{e_p}$ the following holds. For every $s \in [n]$ and a pair of strictly increasing sequences $\mathsf{t} = (t_1, \ldots, t_s), \mathsf{c} = (c_1, \ldots, c_s) \in ([0, n) \cap 2\mathbb{Z})^s$ satisfying $\mathsf{c} \leq \mathsf{t}$, $\exists (x_1, \ldots, x_s) \in \{0,1\}^s \quad \Phi^p_{\mathsf{t},\mathsf{c}}(x_1, \ldots, x_s) \neq 0$.*

3.2.3 Experiments Supporting Conjecture 4

To support Conjecture 4 and, more fundamentally, to verify that there is no "structure" obstructing our heuristic arguments, we ran a computer search. Let $\mathcal{P}_2 : \mathbb{N} \to \mathbb{N}$ be the function that maps $n \in \mathbb{N}$ to the least prime p that satisfies the following property. For every $s \in [n]$, and every pair of strictly increasing sequences $\mathsf{t} = (t_1, \ldots, t_s), \mathsf{c} = (c_1, \ldots, c_s) \in ([0, n) \cap 2\mathbb{Z})^s$ satisfying $\mathsf{c} \leq \mathsf{t}$, it holds that $\Phi^p_{\mathsf{t},\mathsf{c}}(x_1, \ldots, x_s) \neq 0$ for some $(x_1, \ldots, x_s) \in \{0,1\}^s$. Informally, $\mathcal{P}_2$ maps n to the least prime that is "good" for n in our conjecture. In comparison with $\mathcal{P}_1$, for every t, c in question, $\mathcal{P}_1$ provides $\Phi^p_{\mathsf{t},\mathsf{c}}$ a single trial by evaluating it over the origin, while $\mathcal{P}_2$ evaluates it over the entire Boolean hypercube of dimension s, and accepts the smallest prime that for every such t, c, $\Phi^p_{\mathsf{t},\mathsf{c}}$ doesn't vanish on at least one of its points.

An exhaustive search we have conducted, spanned over hundreds of computer hours, verifies at least for small numbers, that unlike $\mathcal{P}_1(n)$, the function $\mathcal{P}_2(n)$ grows very slowly with n. In fact, the data collected in Table 2 shows that for $7 \leq n \leq 30$, $\mathcal{P}_2(n)$ equals the least prime number $p \geq n - 1$. [1]

[1] This is tight, namely, for every $n \geq 7$, $\mathcal{P}_2(n) \geq n - 1$. Indeed, take $p < n - 1$ a prime. If $p \geq 5$, consider

Table 2 Values of $\mathcal{P}_2(n)$ obtained using a computer search. Note that for an even n, $\mathcal{P}_2(n) = \mathcal{P}_2(n-1)$ as t, c have even entries. Thus, only the data of odd n's is collected.

n	5	7	9	11	13	15	17	19	21	23	25	27	29
$\mathcal{P}_2(n)$	3	7	11	11	13	17	17	19	23	23	29	29	29

We do not expect $\mathcal{P}_2(n)$ to grow so slowly and we certainly do not expect it to have such a simple formula. While we could not compute $\mathcal{P}_2(n)$ for $n > 29$, we were able to show that $\mathcal{P}_2(127) > 131$ by eliminating the first two "potential" primes $127, 131$. To see that, say, $\mathcal{P}_2(127) \neq 131$ we invite the diligent reader to verify that $\mathsf{c} = (0, 4, 10)$, $\mathsf{t} = (64, 68, 74)$ yields a counterexample. That is,

$$\begin{vmatrix} \binom{64+x_1}{0} & \binom{64+x_1}{4} & \binom{64+x_1}{10} \\ \binom{68+x_2}{0} & \binom{68+x_2}{4} & \binom{68+x_2}{10} \\ \binom{74+x_3}{0} & \binom{74+x_3}{4} & \binom{74+x_3}{10} \end{vmatrix} \equiv_{131} 0$$

for every $(x_1, x_2, x_3) \in \{0,1\}^3$.

3.2.4 Asymptotic Version of Conjecture 4

For the informal heuristic argument used in Section 3.2.1 the point made is that while the number of "tests" (t, c) grows exponentially with s, so does the number of "trials" $(x_1, \ldots, x_s)$. Thus, when considering such a heuristic, s is thought of as an asymptotic parameter. However, Conjecture 4 is stated for every $s \geq 1$. While it may very well be the case that our conjecture holds as is, we prefer to base our construction on a more robust conjecture that avoids the possible "irregularities" that may be present for small values of s.

A natural relaxation is to bound s from below by some parameter s_0 that is may even be allowed to grow with n. However, note that this should be done with some care. Indeed, if Conjecture 4 can be falsified for some value s, it is immediately false for larger values of s. To see this, take the counterexample $\mathsf{c} = (c_1, \ldots, c_s)$, $\mathsf{t} = (t_1, \ldots, t_s) \in [0, n)^s$ and consider $\mathsf{c}' = (c_1, \ldots, c_s, c_{s+1})$, $\mathsf{t}' = (t_1, \ldots, t_s, t_{s+1}) \in [0, n)^s$ where c_{s+1}, t_{s+1} are chosen so that $t_s < c_{s+1} \leq t_{s+1}$. Observe that this has the effect of "embedding" $M_{\mathsf{t},\mathsf{c}}(X_1, \ldots, X_s)$ as the top-left sub matrix of $M_{\mathsf{t}',\mathsf{c}'}(X_1, \ldots, X_{s+1})$. Furthermore, all but the lowest entry of the rightmost column are 0. In particular,

$$\Phi_{\mathsf{t}',\mathsf{c}'}(X_1, \ldots, X_{s+1}) = \binom{t_{s+1} + X_{s+1}}{c_{s+1}} \cdot \Phi_{\mathsf{t},\mathsf{c}}(X_1, \ldots, X_s).$$

Thus, if $\Phi_{\mathsf{t},\mathsf{c}}$ vanishes on $\{0,1\}^s$ then $\Phi_{\mathsf{t}',\mathsf{c}'}$ vanishes on $\{0,1\}^{s+1}$.

The "correct" way of formalizing a relaxation of Conjecture 4 in which only sufficiently large s are of interest is to restrict to pairs t, c for which not only $s \geq s_0$ but also $t_1 \geq c_{s_0}$. Observe that under this condition, counterexamples of size less than s_0 cannot be embedded as in the above discussion. We state below a variant of Conjecture 4 on which our candidate constructions rely. However, when discussing our conjecture we do not distinguish between Conjecture 4 and Conjecture 5 unless such a distinction is essential.

---

the sequences $\mathsf{t} = (0, p+1)$ and $\mathsf{c} = (0, 4)$. Note that $\Phi^p_{\mathsf{t},\mathsf{c}}(x_1, x_2) = \binom{p+1+x_2}{4}$, and that p divides both $\binom{p+1}{4}$ and $\binom{p+2}{4}$. For $p = 2, 3$ one can use $\mathsf{t} = (0, 2p)$, $\mathsf{c} = (0, 2)$. By Table 2, the assertion is false for $n < 7$.

▶ **Conjecture 5** (Asymptotic PDC conjecture). *There exist universal constants $e_p, e_s \geq 1$ such that for every integer $n \geq 1$ and prime $p \geq n^{e_p}$ the following holds. For every $s \geq s_0 \triangleq (\log n)^{e_s}$ and every pair of strictly increasing sequences $\mathsf{t} = (t_1, \ldots, t_s), \mathsf{c} = (c_1, \ldots, c_s) \in ([0, n) \cap 2\mathbb{Z})^s$ satisfying $\mathsf{t} \geq \mathsf{c}$ and $t_1 \geq c_{s_0}$, it holds that $\exists (x_1, \ldots, x_s) \in \{0,1\}^s \quad \Phi^p_{\mathsf{t},\mathsf{c}}(x_1, \ldots, x_s) \neq 0$.*

We will overcome this relaxation with the aid of the explicit tree code by Gelles *et al.* [12], which will provide some structure to the polynomials we need to analyze to prove the correctness of our construction.

3.2.5 Structural Factors of $\Phi_{\mathsf{t},\mathsf{c}}$ and Its Linearization

Conjecture 4 only concerns with the evaluation of $\Phi_{\mathsf{t},\mathsf{c}}$ at the Boolean hypercube which, recall, we prove never vanishes in Appendix A.1. But, as defined, $\Phi_{\mathsf{t},\mathsf{c}}$ does not encode this in any way. In this section, we identify and remove certain factors of $\Phi_{\mathsf{t},\mathsf{c}}$ that are, in a sense, "outside" the Boolean hypercube, and so are of no interest to us.

For sequences t, c as in Conjecture 4, consider the matrix $M_{\mathsf{t},\mathsf{c}}(X_1, \ldots, X_s)$. Take distinct $i, j \in [s]$ with $i > j$. The substitution $X_i = X_j + t_j - t_i$ turns the i^{th} and j^{th} rows identical, resulting in an identically zero determinant. By Hilbert's Nullstellensatz, $X_i - X_j + t_i - t_j$ divides $\Phi_{\mathsf{t},\mathsf{c}}(X_1, \ldots, X_s)$ in $\mathbb{Q}[X_1, \ldots, X_s]$. Therefore the determinant polynomial is of the form

$$\Phi_{\mathsf{t},\mathsf{c}}(X_1, \ldots, X_s) = \Xi_{\mathsf{t},\mathsf{c}}(X_1, \ldots, X_s) \cdot \prod_{i>j} (X_i - X_j + t_i - t_j)$$

for some polynomial $\Xi_{\mathsf{t},\mathsf{c}}[X_1, \ldots, X_s] \in \mathbb{Q}[X_1, \ldots, X_s]$. In fact, by Gauss's lemma for GCD domains, $\Xi_{\mathsf{t},\mathsf{c}}[X_1, \ldots, X_s]$ is in $\mathbb{Z}[X_1, \ldots, X_s]$ since $\Phi_{\mathsf{t},\mathsf{c}}$ and the structural factor are both primitive. Thus, we can consider reduction modulo a prime p. Since t_i, t_j are distinct even numbers in $[0, n)$, the structural factors do not vanish at any point of the Boolean hypercube, even when reduced modulo a prime $p > n$. Therefore, studying the zeros of $\Phi^p_{\mathsf{t},\mathsf{c}}$ in the Boolean hypercube is equivalent to studying those of $\Xi_{\mathsf{t},\mathsf{c}}$, even modulo a prime $p > n$.

Observe that the linearization of the univariate polynomial $\binom{X+t}{c}$, for $c \geq 1$ takes the nice form $\binom{X+t}{c} = \binom{t}{c-1}X + \binom{t}{c}$ as can be seen using Pascal's identity. In Appendix A.2 we take these ideas a step further and obtain a reformulation of Conjecture 4 which, informally, states that a certain polynomial $\Psi^p_{\mathsf{t},\mathsf{c}}$ is nonzero (as an element of the ring $\mathbb{F}_p[X_1, \ldots, X_s]$). That is to say, while the [7] tree code fails over $\mathbb{F}_p$ if a certain polynomial has a root at the origin, via its reformulation, Conjecture 4 is false only if a certain polynomial is the zero polynomial. An asymptotic version, equivalent to Conjecture 5 is immediate.

In Appendix B we suggest a stronger variant of Conjecture 4 and further study the plausibility of Conjecture 4 and its stronger variant based on deep results in arithmetic geometry. In particular, we reason about the distribution of values attain by $\Phi_{\mathsf{t},\mathsf{c}}$ on the Boolean hypercube by considering the exponential sum $\sum_{(x_1,\ldots,x_s)\in\{0,1\}^s} \zeta_p^{\Phi_{\mathsf{t},\mathsf{c}}(x_1,\ldots,x_s)}$, where ζ_p is a p^{th} root of unity in $\mathbb{C}$, and collect computational data that appear in a longer version of the paper [3]. However, we wrap up this preliminary discussion on our conjecture and its variants. In the next section we go back to the problem of constructing tree codes, and give an informal presentation of our construction and its analysis, based on Conjecture 4 or, more precisely, based on the asymptotic variant, Conjecture 5.

3.3 The Candidate Tree Code

Our candidate construction is a variant of the construction discussed in Section 2. In fact, for obtaining distance larger than $\frac{1}{2}$, [7] modified their original construction so that at time t, not one but some $r \geq 1$ number of evaluations of the "current" polynomial f_t is recorded. This enabled them to achieve distance $1 - \frac{1}{r+1}$. Our candidate construction is closely related to that variant. We make use of this idea of multiple evaluations not for improving the distance, but rather for relaxing the analysis so that it is plausible that the reduction modulo a small prime p yields non-vanishing distance and, in particular, follows by Conjecture 5.

Recall, however, that Conjecture 5 holds only for pairs of length $s \geq s_0$ for which $t_1 \geq c_{s_0}$. Therefore, we need to introduce some mechanism to the construction so that its correctness does not rely on the behaviour when applied with small values of s (nor on invalid pairs). To this end, we make use of an explicit tree code construction by [12]. For every $n \geq 1$, an explicit tree code $\mathsf{TC}' : [n^2]^n \to [2n^2]^n$ having distance $\frac{1}{\log n}$ is given (see Corollary 11 for a precise statement). Although TC' has a vanishing distance, it suffices for our needs as we will not use TC' directly for arguing about the distance; rather, we invoke TC' to guarantee some structure on the polynomials we need to analyze.

Take $p > 2n^2$ a prime, and think of $\mathsf{TC}' : [n^2]^n \to \mathbb{F}_p^n$ in the natural way. Our construction proceeds as follows. Given $m = (m_0, m_1, \ldots, m_{n-1}) \in [n^2]^n$ we first apply TC' to obtain $\left(\gamma_0, \gamma_2, \gamma_4, \ldots, \gamma_{2(n-1)}\right) = \mathsf{TC}'(m)$. For $t \in [0, n)$, we define $f_t(T) \in \mathbb{F}_p[T]$ by $f_t(T) = \sum_{i=0}^{t} \gamma_{2i} \binom{T}{2i}$. At time $t \in [0, n)$, our tree code $\mathsf{TC} : [n^2]^n \to (\mathbb{F}_p^3)^n$ outputs $\mathsf{TC}(m)_t = (\gamma_{2t}, f_t(2t), f_t(2t+1))$. As mentioned, as the alphabet is of size $\mathrm{poly}(n)$, standard techniques can then be used to obtain an explicit binary tree code with comparable parameters. We thus have,

▶ **Theorem 6.** *Assume that Conjecture 5 holds with parameters e_p, e_s. Then, there exist $c = c(e_p, e_s) \in \mathbb{N}$ and $\delta = \delta(e_p, e_s) \in (0, 1)$ such that the following holds. For every $n \in \mathbb{N}$ there exists an explicit tree code $\mathsf{TC} : \{0,1\}^n \to [c]^n$ with distance δ.*

3.3.1 Sketch of the Analysis

As for the analysis, consider distinct $m = (m_0, m_1, \ldots, m_{n-1})$, $m' = (m'_0, m'_1, \ldots, m'_{n-1})$, and let $c \in [0, n)$ be the least integer for which $m_c \neq m'_c$. By the property of TC' we get that for every $\ell \in [0, n-c)$, when restricted to $[c, c+\ell]$, the strings $\gamma = \mathsf{TC}'(m)$, $\gamma' = \mathsf{TC}'(m')$ are of distance $s \geq \frac{\ell}{\log n}$. In particular, when considering $\ell \geq (\log n)^e$ for some constant $e > 1$, we have that $s \geq (\log n)^{e-1}$. Let us assume this bound on ℓ for the moment. Observe now that, by construction, s is precisely the sparsity of the polynomial $g(T) = f_{c+\ell}(T) - f'_{c+\ell}(T)$ with respect to the Newton basis. Thus, we can write $g(T) = \sum_{j=1}^{s} \gamma''_{2c_j} \binom{T}{2c_j}$, where $c = c_1 < c_2 < \cdots < c_s \leq c + \ell < n$ and $\gamma''_{2c_j} = \gamma_{2c_j} - \gamma'_{2c_j}$.

We wish to bound the number of integers $t \in [c, c+\ell]$ for which $g(2t) = g(2t+1) = 0$ as indeed for every such t, $\mathsf{TC}(m)_t$ and $\mathsf{TC}(m')_t$ agree when projected to the last two entries of the triplet. To get a bound of b on such indices t, the natural approach is to assume the existence of some $t_1 < t_2 < \cdots < t_b$ in $[c, c+\ell]$ with $g(2t_i) = g(2t_i + 1) = 0$ for every $i \in [b]$, and try to get a contradiction via Conjecture 5 for a sufficiently large value b. Recall, however, that for the conjecture it is required that $\mathsf{t} \geq \mathsf{c}$ which is not necessarily the case. In [7] this technical issue is resolved by observing that one can restrict to the longest prefixes $(c_1, c_2, \ldots, c_{s_1})$, $(t_1, t_2, \ldots, t_{s_1})$ of the original sequences for which $c_i \leq t_i$ for every $i \in [s_1]$. Such s_1 exists as $c_1 \leq t_1$.

Our analysis is somewhat trickier as we can only invoke Conjecture 5 starting from some s_0 (and under some restriction on the pair). In particular, in the notation of Conjecture 5, we have $s_0 = (\log(2n))^{e_s}$, and it may very well be the case that the longest prefix length

$s_1 < s_0$. To overcome this, and to satisfy the hypothesis of Conjecture 5, we first prove a bound of s on the number of t_i's in $[c_{s_0}, c+\ell]$ rather than in $[c, c+\ell]$. This can be done based on Conjecture 5 using a similar argument to that of [7] who invoke the LGV Lemma.

To bound the number of the remaining t_i's, namely, those in $[c, c_{s_0}]$ we bound the length of this interval. Had $c_1, \ldots, c_{s_0}$ been arbitrary, the interval's length could have been unbounded. However, recall that by construction, $c_1, \ldots, c_{s_0}$ are the indices in $[c, c_{s_0}]$ for which $\mathsf{TC}'(m), \mathsf{TC}'(m')$ disagree. Since TC' has distance $\frac{1}{\log n}$ it follows that $s_0 \geq \frac{c_{s_0}-c}{\log n}$, and so the interval's length is bounded by $c_{s_0} - c \leq s_0 \log n \leq (\log(2n))^{e_s+1}$. Hence, the total number of t_i's is bounded by $s + (\log(2n))^{e_s+1}$, and so the distance between $\mathsf{TC}(m)$ and $\mathsf{TC}(m')$ when restricted to $[c, c+\ell]$ is at least $\max(s, \ell - (s + (\log(2n))^{e_s+1}))$. By taking ℓ sufficiently large, the latter approaches $\frac{\ell}{3}$.

In the discussion above, we assumed ℓ is sufficiently large. In particular, $\ell > \ell_0 = (\log n)^e$ for some constant e. To resolve this "lag", namely, to handle also smaller values of ℓ, we use a standard technique in which an explicit tree code of length $O(\ell_0)$ is concatenated with the construction above.

4 Preliminaries

Let $n \geq 1$ be an integer and Σ some (finite or infinite) set. For a string $x = (x_1, \ldots, x_n) \in \Sigma^n$ and integers $1 \leq a \leq b \leq n$, we let $x_{[a,b]}$ denote the substring $(x_a, \ldots, x_b)$. Given $x, y \in \Sigma^n$, we write $\mathsf{dist}(x, y)$ for their Hamming distance. For an integer $n \geq 1$ write $[n]$ for $\{1, 2, \ldots, n\}$. For integers $a < b$ we denote $[a, b) = \{a, a+1, \ldots, b-1\}$. We use the conventions that the natural numbers are $\mathbb{N} = \{0, 1, 2, \ldots\}$, and that $\binom{a}{b} = 0$ for integers $0 \leq a < b$.

Tree codes, as their name suggest, are trees with certain distance properties. However, as discussed in Section 2, we use an equivalent definition of tree codes that more explicitly specifies their online characteristic. Recall that a function $f\colon \Sigma_{\mathsf{in}}^n \to \Sigma_{\mathsf{out}}^n$ is said to be *online* if for every $i \in [n]$ and $x \in \Sigma_{\mathsf{in}}^n$, $f(x)_i$ is determined by $x_1, \ldots, x_i$. For a pair of distinct $x, y \in \Sigma^n$, we define $\mathsf{split}(x, y)$ as the least integer $s \in [n]$ such that $x_s \neq y_s$.

▶ **Definition 7** ([23]). *An online function* $\mathsf{TC}\colon \Sigma_{\mathsf{in}}^n \to \Sigma_{\mathsf{out}}^n$ *is a* tree code with distance δ *if for every distinct* $x, y \in \Sigma_{\mathsf{in}}^n$*, with* $s = \mathsf{split}(x, y)$*, and every* $\ell \in [0, n-s)$*,*

$$\mathsf{dist}\left(\mathsf{TC}(x)_{[s,s+\ell]}, \mathsf{TC}(y)_{[s,s+\ell]}\right) \geq \delta(\ell+1).$$

We refer to n *as the* depth *of* TC*. We refer to* $\Sigma_{\mathsf{in}}, \Sigma_{\mathsf{out}}$ *as the input alphabet and output alphabet, respectively.*

We are interested in some further properties of tree codes.

▶ **Definition 8.** *Let* $\mathsf{TC}\colon \Sigma_{\mathsf{in}}^n \to \Sigma_{\mathsf{out}}^n$ *be a tree code.*

- *We say that* TC *is a* binary tree code *if* $\Sigma_{\mathsf{in}} = \{0, 1\}$.
- *We say that* TC *is* explicit *if it can be evaluated on every input* $m \in \Sigma_{\mathsf{in}}^n$ *in polynomial time in the bit complexity of* m.

5 Proof of Theorem 6

In this section we present our candidate tree code and prove Theorem 6. Our construction is obtained in several steps, where the main part is to construct a relaxation of tree codes, called a lagged tree code. Informally, this is a tree code whose distance property holds only after a certain time interval.

▶ **Definition 9** ([7]). *An online function* $\mathsf{TC}\colon \Sigma_{\mathsf{in}}^n \to \Sigma_{\mathsf{out}}^n$ *is a* lagged tree code with lag ℓ_0 and distance δ *if for every distinct* $x, y \in \Sigma_{\mathsf{in}}^n$*, with* $s = \mathsf{split}(x, y)$*, and every* $\ell \in [\ell_0, n - s)$*,*

$$\mathsf{dist}\left(\mathsf{TC}(x)_{[s,s+\ell]}, \mathsf{TC}(y)_{[s,s+\ell]}\right) \geq \delta(\ell + 1).$$

Note that a tree code is a lagged tree code with lag parameter $\ell_0 = 0$. It is straightforward to transform any lag-ℓ_0 tree code to a tree code using a second tree code of length $O(\ell_0)$. Our construction of lagged tree codes, given below by Proposition 12, has lag $\ell_0 = \mathrm{poly}(\log n)$. A result by Braverman [6] provides, for every constant $\varepsilon \in (0, 1)$ and integer m an asymptotically-good tree code of length m in time $2^{O(m^\varepsilon)}$. Thus, asymptotically-good tree codes of length ℓ_0 can be obtained in time $\mathrm{poly}(n)$. The obtained tree code (as well as the lagged tree code that is given by Proposition 12) is over a $\mathrm{poly}(n)$-size alphabet. It is well-known how to reduce the alphabet to binary, obtaining tree codes with comparable parameters (see, e.g., [21], Proposition 3.1).

In light of the discussion above, we turn to present our candidate construction of $\mathrm{poly}(\log n)$-lagged tree codes over $\mathrm{poly}(n)$-size alphabet. Our construction makes use of a tree code construction by [12].

▶ **Lemma 10** (Lemma 5.1 in [12]). *There exists an absolute constant* $k_0 \in \mathbb{N}$ *such that the following hold for every* $\varepsilon > 0$ *and integers* $k, n \in \mathbb{N}$ *such that* $\frac{k_0 \cdot \log n}{\varepsilon} \leq k \leq n$*. There exists an explicit tree code* $C : \Sigma_{\mathsf{in}}^k \to \Sigma_{\mathsf{out}}^k$ *with* $\Sigma_{\mathsf{in}} = \{0,1\}^{\frac{\log n}{\varepsilon}}$*,* $\Sigma_{\mathsf{out}} = \{0,1\}^{\frac{\log n}{\varepsilon}+1}$*, rate* $\rho' = \frac{1}{1+\varepsilon/\log n}$ *and relative distance at least* $\delta' = \frac{1}{1+2\log(n)/\varepsilon}$*.*

The following is a straightforward corollary of Lemma 10 obtained by taking $\varepsilon = \frac{1}{2}$. Note that the factors of 4 and 8 in the alphabet size of TC' in Corollary 11 are for obtaining a tree code for every n, not just a power of two as in Lemma 10.

▶ **Corollary 11.** *There exists a universal constant* $n_0 \geq 1$ *such that for every integer* $n \geq n_0$ *there exists an explicit tree code* $\mathsf{TC}' : [4n^2]^n \to [8n^2]^n$ *with distance* $\delta = \frac{1}{5 \log n}$*.*

Given an integer $n \geq n_0$ we proceed as follows. Let p be the least prime number larger than $\max(8n^2, (2n)^{e_p})$, where e_p is the constant from Conjecture 5. By Corollary 11, there exists an explicit tree code $\mathsf{TC}' : [4n^2]^n \to [8n^2]^n$ with distance $\frac{1}{5 \log n}$. As $p > 8n^2$ we can embed the output symbols of TC' in $\mathbb{F}_p$ by identifying them with the field elements $1, \ldots, 8n^2$ of $\mathbb{F}_p$. Hence, we may think of TC' as a function of the form $\mathsf{TC}' : [4n^2]^n \to \mathbb{F}_p^n$.

Define the function $\mathsf{TC} : [4n^2]^n \to (\mathbb{F}_p^3)^n$ as follows. Let $m = (m_0, m_1, \ldots, m_{n-1}) \in [4n^2]^n$. Compute $\mathsf{TC}'(m) = (\gamma_0, \gamma_2, \gamma_4, \ldots, \gamma_{2(n-1)}) \in \mathbb{F}_p^n$. For $t = 0, 1, \ldots, n-1$ define the polynomial $f_t(T) \in \mathbb{F}_p[T]$ by

$$f_t(T) = \sum_{i=0}^{t} \gamma_{2i} \binom{T}{2i}. \tag{5.1}$$

Finally, for $t = 0, 1, \ldots, n-1$, define

$$\mathsf{TC}(m)_t = \left(\gamma_{2t}, f_t(2t), f_t(2t+1)\right). \tag{5.2}$$

▶ **Proposition 12.** *Assume that Conjecture 5 holds with parameters* e_p, e_s*. Then,* TC *as defined in Equation* (5.2) *is an* ℓ_0*-lagged tree code, where* $\ell_0 = 15(\log(2n))^{e_s+1}$*, having distance* $\frac{1}{3}$ *and rate at least* $\frac{1}{2\max(2,e_p)}$*.*

Proof. That the rate is bounded below by $\frac{1}{2\max(2,e_p)}$ is a straightforward calculation. We turn to analyze the distance. Note that TC is not linear and so, for the distance analysis, we consider two distinct messages. Let $m = (m_0, \ldots, m_{n-1}) \in [4n^2]^n$ and $m' = (m'_0, \ldots, m'_{n-1}) \in [4n^2]^n$

distinct. Let $0 \le c \le n-1$ be the least integer for which $m_c \neq m'_c$, and let $\ell \in [\ell_0, n-c)$. Denote

$$\gamma = (\gamma_0, \gamma_2, \ldots, \gamma_{2(n-1)}) = \mathsf{TC}'(m),$$
$$\gamma' = (\gamma'_0, \gamma'_2, \ldots, \gamma'_{2(n-1)}) = \mathsf{TC}'(m').$$

Since TC' has distance $\frac{1}{5\log n}$ it holds that

$$\begin{aligned} s &\triangleq \mathsf{dist}\left(\gamma_{[c,c+\ell]}, \gamma'_{[c,c+\ell]}\right) \\ &= \mathsf{dist}\left((\gamma_{2c}, \gamma_{2(c+1)}, \ldots, \gamma_{2(c+\ell)}), (\gamma'_{2c}, \gamma'_{2(c+1)}, \ldots, \gamma'_{2(c+\ell)})\right) \\ &\ge \frac{\ell+1}{5\log n}. \end{aligned}$$

As $\ell \ge \ell_0$ we have that $s > s_0$, where $s_0 \triangleq (\log(2n))^{e_s}$. Similarly to Equation (5.1), we define for $t = 0, 1, \ldots, n-1$ the polynomial $f'_t(T) \in \mathbb{F}_p[T]$ by

$$f'_t(T) = \sum_{i=0}^{t} \gamma'_{2i} \binom{T}{2i}.$$

Observe that s is precisely the sparsity of $f_{c+\ell}(T) - f'_{c+\ell}(T)$ with respect to the Newton basis. Let $c \le c_1 < c_2 < \cdots < c_s \le c+\ell$ be all the integers such that $\gamma_{2c_j} \neq \gamma'_{2c_j}$ for every $j \in [s]$. As TC' is a tree code (with nonzero distance) $\gamma_{2c} = \mathsf{TC}'(m)_c \neq \mathsf{TC}'(m')_c = \gamma'_{2c}$, and so $c_1 = c$. By denoting $\gamma''_i = \gamma_i - \gamma'_i$, one can write the polynomial $f_{c+\ell}(T) - f'_{c+\ell}(T)$ as

$$g(T) = \sum_{j=1}^{s} \gamma''_{2c_j} \binom{T}{2c_j}.$$

Define $Z = \{t \in [c, c+\ell] \mid g(2t) = g(2t+1) = 0\}$.

▷ Claim 13. Assuming Conjecture 5, $|Z \cap [c_{s_0}, c+\ell]| < s$.

Proof. Assume by way of contradiction that there are distinct integers $t_1, \ldots, t_s \in [c_{s_0}, c+\ell]$ such that

$$\forall i \in [s] \quad g(2t_i) = g(2t_i+1) = 0. \tag{5.3}$$

Assume further that $t_1 < \cdots < t_s$. Let $s_1 \in \{s_0, s_0+1, \ldots, s\}$ be the largest integer with the property that for every $i \in \{s_0, s_0+1, \ldots, s_1\}$, $t_i \ge c_i$. Note that s_1 is well-defined as $t_{s_0} \ge c_{s_0}$ (and so the maximum is taken over a non-empty, finite, set). Let $M(X_1, \ldots, X_{s_1})$ be the $s_1 \times s_1$ matrix whose $(i,j)^{\text{th}}$ entry is

$$M_{i,j}(X_1, \ldots, X_{s_1}) = \binom{X_i + 2t_i}{2c_j},$$

where $X_1, \ldots, X_{s_1}$ are formal variables. Let $\Phi \in \mathbb{F}_p[X_1, \ldots, X_{s_1}]$ be the polynomial that is given by $\Phi(X_1, \ldots, X_{s_1}) = \det M(X_1, \ldots, X_{s_1})$. Denote $\mathsf{t} = (2t_1, \ldots, 2t_{s_1})$ and $\mathsf{c} = (2c_1, \ldots, 2c_{s_1})$. Note that Φ as defined above is precisely $\Phi^p_{\mathsf{t},\mathsf{c}}$ in the notation of Conjecture 5. Clearly, $\mathsf{t}, \mathsf{c} \in ([0, 2n) \cap 2\mathbb{Z})^{s_1}$. We turn to show that $\mathsf{c} \le \mathsf{t}$. Indeed, for $i \in \{s_0, s_0+1, \ldots, s_1\}$ we have that $t_i \ge c_i$ by the definition of s_1. Moreover, recall that for every $i \in [s]$, $t_i \ge c_{s_0}$, and so, for $i < s_0$ we have that $t_i \ge c_{s_0} > c_i$. Recall that $p \ge (2n)^{e_p}$, $s_1 \ge s_0 = (\log(2n))^{e_s}$, and $2t_1 \ge 2c_{s_0}$. Thus, the hypothesis of Conjecture 5 is met with s, n in the notation

the conjecture taken to be s_1 and $2n$ in our notation, respectively. Therefore, assuming the validity of Conjecture 5 we conclude the existence of $(x_1, \ldots, x_{s_1}) \in \{0,1\}^{s_1}$ such that $\Phi(x_1, \ldots, x_{s_1}) \neq 0$ in $\mathbb{F}_p$.

We now use $(x_1, \ldots, x_{s_1})$ to get a contradiction. Let $\Gamma \in \mathbb{F}_p^{s_1}$ be the vector with i^{th} entry $\Gamma_i = \gamma''_{2c_i}$. Observe that Γ is a nonzero vector. To see this, consider its first entry $\Gamma_1 = \gamma''_{2c_1} = \gamma''_{2c}$. Recall that $\gamma''_{2c} = \gamma_{2c} - \gamma'_{2c}$. As TC' is a tree code (with distance larger than 0) and since $m_c \neq m'_c$ we have that $\gamma_{2c} = \mathsf{TC}'(m)_c \neq \mathsf{TC}'(m')_c = \gamma'_{2c}$. Thus, $\Gamma_1 \neq 0$.

Since $\Phi(x_1, \ldots, x_{s_1}) \neq 0$ we have that $M(x_1, \ldots, x_{s_1})$ is nonsingular, and therefore $M(x_1, \ldots, x_{s_1})\Gamma$ is a nonzero vector. Let then $i \in [s_1]$ be such that $(M(x_1, \ldots, x_{s_1})\Gamma)_i \neq 0$. Note that

$$(M(x_1, \ldots, x_{s_1})\Gamma)_i = \sum_{j=1}^{s_1} \gamma''_{2c_j} \binom{x_i + 2t_i}{2c_j}. \tag{5.4}$$

Assume for the moment that $s_1 < s$. As $i \leq s_1$ we have that $i < s$ and so we may refer to t_{i+1}. As $x_i \in \{0,1\}$, we have that $2t_i + x_i \leq 2t_i + 1 < 2t_{i+1}$. Hence, as $i \leq s_1$, $2t_i + x_i < 2t_{s_1+1}$. By the definition of s_1 we have that $t_{s_1+1} < c_{s_1+1}$, and so $2t_i + x_i < 2c_{s_1+1}$. Hence, $\binom{x_i+2t_i}{2c_j} = 0$ for all $j \in \{s_1 + 1, \ldots, s\}$. Thus,

$$\sum_{j=1}^{s_1} \gamma''_{2c_j} \binom{x_i + 2t_i}{2c_j} = \sum_{j=1}^{s} \gamma''_{2c_j} \binom{x_i + 2t_i}{2c_j} = g(2t_i + x_i). \tag{5.5}$$

Equation (5.5) trivially follows also when $s_1 = s$, and so it holds in general, namely, without any assumption on s_1. Equations (5.4) and (5.5) together imply that

$$g(2t_i + x_i) = (M(x_1, \ldots, x_{s_1})\Gamma)_i \neq 0$$

which, as $x_i \in \{0,1\}$, stands in contradiction to Equation (5.3), and thus proving the claim. ◁

▷ Claim 14. $|Z| \leq s + 5(\log(2n))^{e_s+1}$.

Proof. As TC' is a tree code with distance $\frac{1}{5\log n}$, we have that

$$\begin{aligned} s_0 &= \mathsf{dist}\left((\gamma_{2c_1}, \gamma_{2(c_1+1)}, \ldots, \gamma_{2c_{s_0}}), (\gamma'_{2c_1}, \gamma'_{2(c_1+1)}, \ldots, \gamma'_{2c_{s_0}})\right) \\ &= \mathsf{dist}\left(\mathsf{TC}'(m)_{[c_1, c_{s_0}]}, \mathsf{TC}'(m')_{[c_1, c_{s_0}]}\right) \\ &\geq \frac{c_{s_0} - c_1 + 1}{5\log n} \\ &\geq \frac{c_{s_0} - c}{5\log n}. \end{aligned}$$

Now, $s_0 = (\log(2n))^{e_s}$, and so

$$c_{s_0} - c \leq 5(\log n)(\log(2n))^{e_s} \leq 5(\log(2n))^{e_s+1}.$$

This, together with Claim 13, implies that

$$\begin{aligned} |Z| &\leq (c_{s_0} - c) + |Z \cap [c_{s_0}, c + \ell]| \\ &\leq s + 5(\log(2n))^{e_s+1}. \end{aligned}$$

◁

▷ Claim 15. For every $t \in [c, c + \ell]$ and $x \in \{0,1\}$,

$$g(2t + x) = f_t(2t + x) - f'_t(2t + x).$$

Proof. Recall that $c_1, \ldots, c_s$ are precisely the indices in $[c, c+\ell]$ for which γ and γ' disagree. More precisely, for $i \in [c, c+\ell]$, $\gamma_{2i} \neq \gamma'_{2i}$ if and only $i \in \{c_1, \ldots, c_s\}$. Hence, for every $t \in [c, c+\ell]$ and $x \in \{0,1\}$,

$$\begin{aligned} f_t(2t+x) - f'_t(2t+x) &= \sum_{i=0}^{t} (\gamma_{2i} - \gamma'_{2i}) \binom{2t+x}{2i} \\ &= \sum_{\substack{j \in [s] \\ c_j \leq t}} \gamma''_{2c_j} \binom{2t+x}{2c_j} \\ &= \sum_{j=1}^{s} \gamma''_{2c_j} \binom{2t+x}{2c_j} \\ &= g(2t+x), \end{aligned}$$

where the penultimate equality follows since $\binom{2t+x}{2c_j} = 0$ for every $j \in [s]$ for which $c_j > t$. Indeed, if $c_j > t$ then $2c_j \geq 2t+2$ and so $\binom{2t}{2c_j} = \binom{2t+1}{2c_j} = 0$. ◁

By Claim 15, $t \in Z$ if and only if the last two entries of $\mathsf{TC}(m)_t$, namely, $f_t(2t), f_t(2t+1)$, agree with the corresponding entries, $f'_t(2t), f'_t(2t+1)$, of $\mathsf{TC}(m')_t$. As the third entry of $\mathsf{TC}(m)$ and $\mathsf{TC}(m')$, when restricted to $[c, c+\ell]$, disagree on exactly s indices, we have that the number of indices $t \in [c, c+\ell]$ for which $\mathsf{TC}(m)_t \neq \mathsf{TC}(m')_t$ (as a triplet) is bounded below by

$$\begin{aligned} \max\left(s, \ell + 1 - |Z|\right) &\geq \max\left(s, \ell + 1 - s - 5(\log(2n))^{e_s+1}\right) \\ &\geq \frac{\ell - 5(\log(2n))^{e_s+1} + 1}{2} \\ &\geq \frac{\ell+1}{3}, \end{aligned}$$

where the last inequality follows since $\ell \geq \ell_0 = 15(\log(2n))^{e_s+1}$. ◀

References

1. Martin Aigner. *A course in enumeration*, volume 238 of *Graduate Texts in Mathematics*. Springer, Berlin, 2007. doi:10.1145/1814370.1814375.
2. Martin Aigner and Günter M. Ziegler. *Proofs from The Book*. Springer, Berlin, sixth edition, 2018. See corrected reprint of the 1998 original [MR1723092], Including illustrations by Karl H. Hofmann. doi:10.1007/978-3-662-57265-8.
3. Inbar Ben Yaacov, Gil Cohen, and Anand Kumar Narayanan. Candidate tree codes via Pascal determinant cubes. *ECCC*, 2020. URL: https://eccc.weizmann.ac.il/report/2020/141/.
4. Siddharth Bhandari and Prahladh Harsha. A note on the explicit constructions of tree codes over polylogarithmic-sized alphabet. *arXiv preprint arXiv:2002.08231*, 2020. URL: https://arxiv.org/abs/2002.08231.
5. Zvika Brakerski, Yael Tauman Kalai, and Raghuvansh R. Saxena. Deterministic and efficient interactive coding from hard-to-decode tree codes. In *61st Annual IEEE Symposium on Foundations of Computer Science—FOCS 2020*, pages 446–457. IEEE Computer Soc., Los Alamitos, CA, 2020. doi:10.1109/FOCS46700.2020.00049.
6. Mark Braverman. Towards deterministic tree code constructions. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 161–167. ACM, New York, 2012. doi:10.1145/2090236.2090250.

7 Gil Cohen, Bernhard Haeupler, and Leonard J. Schulman. Explicit binary tree codes with polylogarithmic size alphabet. In *STOC'18—Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 535–544. ACM, New York, 2018. doi:10.1145/3188745.3188928.

8 Gil Cohen and Shahar Samocha. Palette-alternating tree codes. In *The 35th Computational Complexity Conference (CCC 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020. doi:10.4230/LIPIcs.CCC.2020.11.

9 Pierre Deligne. La conjecture de weil : I. *Publications Mathématiques de l'IHÉS*, 43:273–307, 1974. doi:10.1007/bf02684373.

10 Étienne Fouvry. Consequences of a result of N. Katz and G. Laumon concerning trigonometric sums. *Israel Journal of Mathematics*, 120:81–96, 2000. doi:10.1007/s11856-000-1272-z.

11 Étienne Fouvry and Nicholas Katz. A general stratification theorem for exponential sums, and applications. *Journal Fur Die Reine Und Angewandte Mathematik - J REINE ANGEW MATH*, 2001:115–166, January 2001. doi:10.1515/crll.2001.082.

12 Ran Gelles, Bernhard Haeupler, Gillat Kol, Noga Ron-Zewi, and Avi Wigderson. Towards optimal deterministic coding for interactive communication. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1922–1936. ACM, New York, 2016. doi:10.1137/1.9781611974331.ch135.

13 Ira Gessel and Gérard Viennot. Binomial determinants, paths, and hook length formulae. *Adv. in Math.*, 58(3):300–321, 1985. doi:10.1016/0001-8708(85)90121-5.

14 Richard W. Hamming. Error detecting and error correcting codes. *Bell System Tech. J.*, 29:147–160, 1950. doi:10.1002/j.1538-7305.1950.tb00463.x.

15 Jørn Justesen. Class of constructive asymptotically good algebraic codes. *IEEE Transactions on Information Theory*, 18(5):652–656, 1972. doi:10.1109/tit.1972.1054893.

16 Sankeerth R. Karingula and Shachar Lovett. Codes over integers, and the singularity of random matrices with large entries. *CoRR*, abs/2010.12081, 2020. URL: https://arxiv.org/abs/2010.12081.

17 Nicholas Katz and Gérard Laumon. Transformation de fourier et majoration de sommes exponentielles. *Publications Mathématiques de l'IHÉS*, 62:145–202, 1985. doi:10.1007/bf02698808.

18 Serge Lang and Andre Weil. Number of points of varieties over finite fields. *American Journal of Mathematics*, 76(4):819–827, 1954. URL: https://www.jstor.org/stable/2372655.

19 Cristopher Moore and Leonard J. Schulman. Tree codes and a conjecture on exponential sums. In *ITCS'14—Proceedings of the 2014 Conference on Innovations in Theoretical Computer Science*, pages 145–153. ACM, New York, 2014. doi:10.1145/2554797.2554813.

20 Anand Kumar Narayanan and Matthew Weidner. On decoding Cohen-Haeupler-Schulman tree codes. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1337–1356. SIAM, 2020. doi:10.1137/1.9781611975994.81.

21 Pavel Pudlák. Linear tree codes and the problem of explicit constructions. *Linear Algebra Appl.*, 490:124–144, 2016. doi:10.1016/j.laa.2015.10.030.

22 Pavel Pudlák. On matrices potentially useful for tree codes. *CoRR*, abs/2012.03013, 2020. URL: https://arxiv.org/abs/2012.03013.

23 Leonard J. Schulman. Deterministic coding for interactive communication. In *Proceedings of the 25th annual ACM Symposium on Theory of Computing*, pages 747–756, 1993. doi:10.1145/167088.167279.

24 Leonard J. Schulman. Postscript of 21 september 2003 to coding for interactive communication. http://users.cms.caltech.edu/ schulman/Papers/intercodingpostscript.txt, 1994.

25 Leonard J. Schulman. Coding for interactive communication. *IEEE Trans. Inform. Theory*, 42(6, part 1):1745–1756, 1996. Codes and complexity. doi:10.1109/18.556671.

26 Claude E. Shannon. A mathematical theory of communication. *Bell System Tech. J.*, 27:379–423, 623–656, 1948. doi:10.1002/j.1538-7305.1948.tb01338.x.

27 Michael Sipser and Daniel A. Spielman. Expander codes. *IEEE Transactions on Information Theory*, 42(6):1710–1722, 1996. doi:10.1109/18.556667.

28 Michael A. Tsfasman, Serge G. Vlăduţ, and Thomas Zink. Modular curves, Shimura curves, and Goppa codes, better than Varshamov-Gilbert bound. *Mathematische Nachrichten*, 109(1):21–28, 1982. doi:10.1002/mana.19821090103.

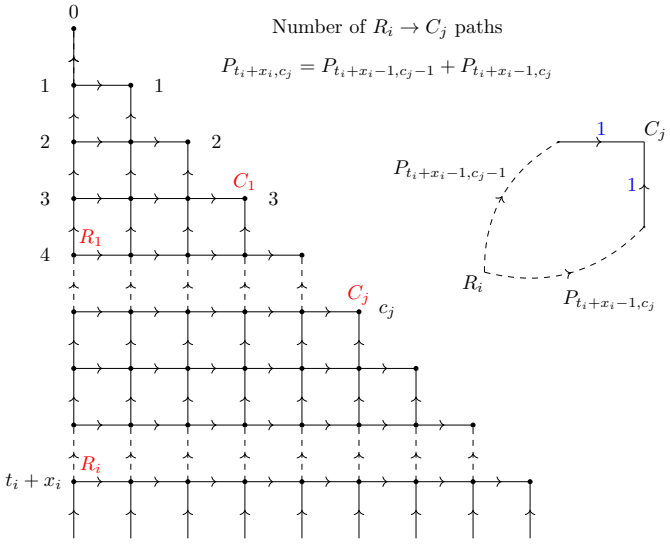
A Combinatorics Corroborating Conjecture 4

A.1 Non-Vanishing of $\Phi_{\mathsf{t},\mathsf{c}}$ on the Boolean Hypercube

In this section we prove that the integer polynomial $\Phi_{\mathsf{t},\mathsf{c}}$ as in Conjecture 4 does not vanish on any point of the Boolean hypercube. To this end, we make use of ideas similar to those used by [7] to prove that $\Phi_{\mathsf{t},\mathsf{c}}$ has no root at the origin. Fix sequences t, c as in Conjecture 4 for the remainder of this section. Consider a directed acyclic graph $G = (V, E)$ with edge weights $\{w(e) \mid e \in E\}$ coming from a commutative ring with identity, along with two ordered vertex sets $\mathcal{R} = \{R_1, R_2, \ldots, R_d\}, \mathcal{C} = \{C_1, C_2, \ldots, C_d\} \subseteq V$ of the same cardinality d. Associated to it is the path matrix M: the square matrix indexed by R, C with the $R \in \mathcal{R}, C \in \mathcal{C}$ entry $M_{R,C} \triangleq \prod_{P:R\to C} w(P)$ where the product is taken over all paths P from R to C and the weight $w(P)$ is the product of edge weights in the path P. Paths of length 0 are included and given the weight 1. A path system $\mathcal{P}$ from $\mathcal{R}$ to $\mathcal{C}$ consists of a permutation $\sigma \in S_d$ and a set of paths $\{P_i : R_i \to C_{\sigma(i)} \mid i \in [d]\}$. Let $\mathrm{sgn}(\mathcal{P})$ denote the sign of σ and $w(\mathcal{P})$ denote the product of the weights $\prod_{i=1}^{d} w(P_i)$. The path system is called vertex disjoint if its set of paths are vertex disjoint. The LGV Lemma is the expression for the determinant of the path matrix M in terms of the underlying path graph

$$\det(M) = \sum_{\substack{\text{vertex disjoint} \\ \text{path systems } \mathcal{P}}} \mathrm{sgn}(\mathcal{P}) w(\mathcal{P}).$$

Gessel and Viennot applied it to path graphs cut out from the square lattice and proved the non vanishing theorem for determinants of Pascal submatrices. We next show a non vanishing of determinants central to our construction using the same path graph but with vertices relabelled.



▶ **Lemma 16.** *For all strictly increasing non negative integer sequences* $\mathsf{c} = (c_1, \ldots, c_s), \mathsf{t} = (t_1, \ldots, t_s)$ *such that* $\mathsf{c} \leq \mathsf{t}$ *and* t_i *is even for all* $i \in [s]$*, and* $\forall (x_1, x_2, \ldots, x_s) \in \{0,1\}^s$*,*

$$\Phi_{\mathsf{t},\mathsf{c}}(x_1, \ldots, x_s) \neq 0.$$

Proof. Fix numbers $c_1, \ldots, c_s, t_1, \ldots, t_s, x_1, \ldots, x_s$ as in the statement. Consider the directed acyclic graph below with unit edge weights and distinguished (in red) vertex subsets $\{R_1, R_2, \ldots, R_s\}$ and $\{C_1, C_2, \ldots, C_s\}$. The $(t_1 + x_1)^{th}$ vertex on the first column is labelled R_1, the $(t_2 + x_2)^{th}$ vertex on the first column is labelled R_2 and so on. The labels R_is are well defined, for $(t_i + x_i)$s are distinct as t_is are even and x_is are in $\{0, 1\}$. The c_1^{th} vertex on the diagonal is labelled C_1, the c_2^{th} vertex on the diagonal is labelled C_2 and so on. To illustrate, $t_1 = 4, x_1 = 0, c_1 = 3$ in the diagram. The horizontal edges are directed from left to right and the vertical edges from bottom to top.

Since all the edge weights are 1, the $(i, j)^{th}$ entry $M_{i,j}$ of the path matrix is the number of paths $P_{t_i+x_i,c_j}$ from R_i to C_j. This satisfies the two term recurrence $P_{t_i+x_i,c_j} = P_{t_i+x_i-1,c_j-1} + P_{t_i+x_i-1,c_j}$ as evident from the picture on the right. This is Pascal's identity for binomials. The boundary conditions $\binom{t_i+x_i}{0} = 1$ and $\binom{t_i+x_i}{c_i} = 1$ for $t_i + x_i = c_i$ are consistent with the path formulation. We conclude that the associated path matrix is $M_{\mathsf{t},\mathsf{c}} = \left\{ \binom{t_i+x_i}{c_j} \;\middle|\; i, j \in [s] \right\}$, whose determinant $\Phi_{\mathsf{t},\mathsf{c}}(x_1, \ldots, x_s)$ is in question. The planar geometry forces all vertex disjoint path systems to have the identity permutation, which has sign 1. Hence the determinant is a positive number provided there is at least one vertex disjoint path system. By the condition $t_i + x_i \geq c_i$ for all i, there is at least one, namely for each $R_i \to C_i$, traverse c_i edges right before turning up. ◀

A.2 Reformulation of Conjecture 4

In this section we provide a reformulation of Conjecture 4. Fix sequences t, c as in Conjecture 4. Consider the variety $X_{\mathsf{t},\mathsf{c}}$ of intersection of the hypercube and the hypersurface generated by $\Phi_{\mathsf{t},\mathsf{c}}$. The variety $X_{\mathsf{t},\mathsf{c}}$ is generated by the ideal

$$\mathcal{I}_{\mathsf{t},\mathsf{c}} := \left\langle \Phi_{\mathsf{t},\mathsf{c}}(X_1, X_2, \ldots, X_s), X_1^2 - X_1, \ldots, X_s^2 - X_s \right\rangle.$$

Clearly, the intersection variety is zero dimensional (or empty), since the hypercube is zero dimensional and $\Phi_{\mathsf{t},\mathsf{c}}$ is nonzero. The degree of the polynomial defining the hypersurface can be reduced through the relations carving out the hypercube as follows. Let $\Psi_{\mathsf{t},\mathsf{c}}(X_1, X_2, \ldots, X_s) \in \mathbb{Z}[X_1, X_2, \ldots, X_s]$ be the unique lift of

$$\Phi_{\mathsf{t},\mathsf{c}}[X_1, X_2, \ldots, X_s] \bmod \left\langle X_1^2 - X_1, X_2^2 - X_2, \ldots, X_s^2 - X_s \right\rangle$$

with degree in each variable at most 1. Informally, $\Psi_{\mathsf{t},\mathsf{c}}$ is merely $\Phi_{\mathsf{t},\mathsf{c}}$ with every indeterminate X_i^* replaced by X_i. The $*$ in the superscript denotes some positive exponent. Since $\Phi_{\mathsf{t},\mathsf{c}}$ is nonzero, so is $\Psi_{\mathsf{t},\mathsf{c}}$. The respective hypersurfaces generated by $\Phi_{\mathsf{t},\mathsf{c}}$ and $\Psi_{\mathsf{t},\mathsf{c}}$ have the same intersection with the Boolean hypercube and hence we can work with either. We will proceed with $\Psi_{\mathsf{t},\mathsf{c}}$ as it has the form

$$\Psi_{\mathsf{t},\mathsf{c}}(X_1, X_2, \ldots, X_s) = \sum_{b=(b_1,b_2,\ldots,b_s)\in\{0,1\}^s} a_b X_1^{b_1} X_2^{b_2} \ldots X_s^{b_s}$$

familiar to Boolean functional analysts with possibly smaller degrees. Further, restricting to the Boolean cube removed the structural factors that concerned us in Section 3.2.5 from $\Psi_{\mathsf{t},\mathsf{c}}$. Let $\Psi_{\mathsf{t},\mathsf{c}}^p \in \mathbb{F}_p[X_1, X_2, \ldots, X_s]$ be the reduction of $\Psi_{\mathsf{t},\mathsf{c}}$ modulo the prime p.

Conjecture 4 amounts to $\Psi^p_{\mathsf{t},\mathsf{c}}$ being a nonzero polynomial. This is, at least one of the coefficients $a_b \bmod p, b \in \{0,1\}^s$ is nonzero. Equivalently, at least one of the evaluations $\Psi^p_{\mathsf{t},\mathsf{c}}(e), e \in \{0,1\}^s \subset \mathbb{F}_p^s$ is nonzero. Below we choose to reformulate the asymptotic version, Conjecture 5.

▶ **Conjecture 17** (Conjecture 5 reformulated). *There exist universal constants $e_p, e_s \geq 1$ such that for every integer $n \geq 1$, prime $p \geq n^{e_p}$, and $s \geq (\log n)^{e_s}$ the following holds. For every pair of strictly increasing sequences $\mathsf{t} = (t_1, \ldots, t_s), \mathsf{c} = (c_1, \ldots, c_s) \in ([0,n) \cap 2\mathbb{Z})^s$ satisfying $\mathsf{c} \leq \mathsf{t}$, it holds that $\Psi^p_{\mathsf{t},\mathsf{c}}(X_1, X_2, \ldots, X_s)$ is nonzero.*

B Arithmetic Geometry Heuristics Supporting Conjecture 4

We laboured through the whole previous section trying to argue that the restriction $\Psi_{\mathsf{t},\mathsf{c}}$ to the Boolean hypercube of $\Phi_{\mathsf{t},\mathsf{c}}$ is not identically zero modulo our chosen prime p. Our starting observation in this section is that the reduction $\Phi^p_{\mathsf{t},\mathsf{c}}$ of $\Phi_{\mathsf{t},\mathsf{c}}$ is non zero, since $\Phi_{\mathsf{t},\mathsf{c}}$ is primitive (it is apparent from the defining equation that the highest total degree term of $\Phi_{\mathsf{t},\mathsf{c}}$ is monic). Therefore, the zeroes of $\Phi^p_{\mathsf{t},\mathsf{c}}$ define a hypersurface (that is, of codimension 1). We study the intersection of the Boolean hypercube sitting inside $\mathbb{F}_p^s$ with this hypersurface using arithmetic geometry. Our analysis falls short of proving Conjecture 4 owing the failure to control some error terms. But we will prove Conjecture 4 holds when relaxed to accommodate hypercubes of side length growing with p.

It is convenient to be ambitious and target stronger versions of Conjecture 4 (or its asymptotic variant, Conjecture 5) which, arguably, are even more natural. First, the distribution of values obtained by evaluating $\Phi^p_{\mathsf{t},\mathsf{c}}$ on the Boolean hypercube $\{0,1\}^s$, for any t, c in question, is fairly balanced when p is taken sufficiently large compared to n. More precisely, we postulate the following conjecture.

▶ **Conjecture 18** (Strong form, value distribution). *There exist universal constants $e_p, e_s \geq 1$ and $\beta \in (0,1)$ such that for every integer $n \geq 1$, prime $p \geq n^{e_p}$, and $s \geq (\log n)^{e_s}$ the following holds. For every pair of strictly increasing sequences $\mathsf{t} = (t_1, \ldots, t_s), \mathsf{c} = (c_1, \ldots, c_s) \in ([0,n) \cap 2\mathbb{Z})^s$ satisfying $\mathsf{c} \leq \mathsf{t}$, it holds that*

$$\left| \sum_{(x_1,\ldots,x_s) \in \{0,1\}^s} \zeta_p^{\Phi_{\mathsf{t},\mathsf{c}}(x_1,\ldots,x_s)} \right| \leq 2^{\beta s}, \tag{B.1}$$

where ζ_p is a p^{th} root of unity in $\mathbb{C}$.

When the prime p exceeds the height of $\Phi_{\mathsf{t},\mathsf{c}}$, the sum concentrates in a wedge above the positive real axis disturbing the equidistribution. Despite not stating explicitly, we are only interested in (and only claim the conjecture) when p is small compared to the height of $\Phi_{\mathsf{t},\mathsf{c}}$. What really concerns us is the distribution of zeroes

$$\Phi_{\mathsf{t},\mathsf{c}}(\mathbb{F}_p, 2) := \left\{ (x_1, x_2, \ldots, x_s) \in \{0,1\}^s \subset \mathbb{F}_p^s \;\middle|\; \Phi^p_{\mathsf{t},\mathsf{c}}(x_1, x_2, \ldots, x_s) = 0 \right\}$$

of $\Phi^p_{\mathsf{t},\mathsf{c}}$ on the Boolean hypercube; suggesting another strengthening of Conjecture 5.

▶ **Conjecture 19** (Strong form, point count). *There exist universal constants $e_p, e_s \geq 1$ and $\beta \in (0,1)$ such that for every integer $n \geq 1$, prime $p \geq n^{e_p}$, and $s \geq (\log n)^{e_s}$ the following holds. For every pair of strictly increasing sequences $\mathsf{t} = (t_1, \ldots, t_s), \mathsf{c} = (c_1, \ldots, c_s) \in ([0,n) \cap 2\mathbb{Z})^s$ satisfying $\mathsf{c} \leq \mathsf{t}$, it holds that $|\Phi_{\mathsf{t},\mathsf{c}}(\mathbb{F}_p, 2)| \leq 2^{\beta s}$.*

We have gathered some data using a computer program to shed some more light on the exponential sum in Conjecture 18, presented in a longer version of the paper [3].

B.1 Pascal Determinant Hypersurfaces

Using arithmetic geometry, we next argue for the rarity of zeroes as stated in Conjecture 19. We start with the most naive yet convincing argument. Before addressing the intersection with the Boolean hypercube, consider the $\mathbb{F}_p$-rational points $\Phi_{\mathsf{t},\mathsf{c}}(\mathbb{F}_p) := \left\{ w \in \mathbb{F}_p \;\middle|\; \Phi^p_{\mathsf{t},\mathsf{c}}(w) = 0 \right\}$ on the hypersurface of dimension $s-1$ and degree $\leq ns$ in isolation. The Schwartz-Zippel Lemma implies $|\Phi_{\mathsf{t},\mathsf{c}}(\mathbb{F}_p)| \leq nsp^{s-1}$. If $\Phi^p_{\mathsf{t},\mathsf{c}}$ is irreducible or if it has only a few (say $N^p_{\mathsf{t},\mathsf{c}}$) irreducible components, the Lang-Weil bound gives the improved estimate [18]

$$\left||\Phi_{\mathsf{t},\mathsf{c}}(\mathbb{F}_p)| - N^p_{\mathsf{t},\mathsf{c}}\; p^{s-1}\right| = (ns-1)(ns-2)p^{s-3/2} + O(nsp^{s-2}).$$

With the unimportant structured factors removed from $\Phi_{\mathsf{t},\mathsf{c}}$, the remaining $\Xi_{\mathsf{t},\mathsf{c}}$ (which is also primitive, by Gauss's lemma) also has non zero reduction $\Xi^p_{\mathsf{t},\mathsf{c}}$. It is not always irreducible. For instance, if the index sets t,c are such that $c_j < t_{j+1}$ for some j, then the vertex disjoint paths connecting the first j vertices are decoupled from the rest: resulting in a factorization of $\Xi_{\mathsf{t},\mathsf{c}}$. But for the factorization induced by such decouplings, the reduction $\Xi^p_{\mathsf{t},\mathsf{c}}$ is likely to be irreducible. Better still, if (the homogenization of) $\Xi^p_{\mathsf{t},\mathsf{c}}$ is irreducible and defines a smooth projective variety, then deep results arising from Deligne's proof of the Weil conjectures [9, Thèoréme 8.1] imply the full *"square root cancellation"*

$$\left||\Xi_{\mathsf{t},\mathsf{c}}(\mathbb{F}_p)| - p^{s-1}\right| = O(b_{s-1}p^{\frac{s-1}{2}})$$

where $b_{s-1} \leq \frac{1}{2}s(s+1)(sn)^s$ is the $s-1^{th}$ Betti number. To derive our heuristic estimate, Schwartz-Zippel will suffice. For ease of exposition, we will use $\Phi_{\mathsf{t},\mathsf{c}}$ in the ensuing analysis, even though $\Xi^p_{\mathsf{t},\mathsf{c}}$ offers some minor gains degree wise. In spirit, the probability $\Phi^p_{\mathsf{t},\mathsf{c}}$ is zero at a point in $\mathbb{F}^s_p$ is centred at $\frac{N^p_{\mathsf{t},\mathsf{c}}}{p}$ with an error term depending on the smoothness. Irrespective of the smoothness, the error term is negligible compared to the estimate for p a big enough polynomial in n. We hypothesise that the hypersurface intersects generically with the Boolean hypercube and the number of intersection points is bounded as

$$|\Phi_{\mathsf{t},\mathsf{c}}(\mathbb{F}_p, 2)| \approx |\Phi_{\mathsf{t},\mathsf{c}}(\mathbb{F}_p)| \left(\frac{2}{p}\right)^s. \tag{B.2}$$

By the Schwartz-Zippel lemma

$$|\Phi_{\mathsf{t},\mathsf{c}}(\mathbb{F}_p, 2)| \approx |\Phi_{\mathsf{t},\mathsf{c}}(\mathbb{F}_p)| \left(\frac{2}{p}\right)^s = O\left(\frac{ns2^s}{p}\right) \tag{B.3}$$

suggesting Conjecture 19 holds for $p > n^2$.

B.2 Katz-Laumon Sums and Point Counting in Hypercubes

Through arithmetic geometric bounds on exponential sums, we argue our determinant hypersurfaces intersect generically with the Boolean hypercube. We show Conjecture 4 holds when relaxed to allow hypercubes of length (larger than 2) growing with the prime. Quantitatively, the bounds attained fall short of proving Conjecture 4. Yet, the methods are illuminating and suggest there are no arithmetic obstructions to our conjectures.

The key ingredient is the Katz-Laumon sum [17]. Building on Grothendieck's foundational trace formula for $\ell-$adic cohomology and Deligne's proof of the Weil conjectures, Katz and Laumon studied certain trigonometric sums over arbitrary high dimensional varieties over finite fields, parametrized by auxiliary points. They proved square root cancellation without any strong geometric assumption (such as smoothness) on the variety, for almost all choices

of the parameter. Fouvry [10] and Fouvry-Katz [11] extended Katz and Laumon's theorem to obtain a stratified theorem. Fouvry applied Katz-Laumon sums to count points of a variety on hypercubes (Boolean or more general). Fouvry and Katz extended this approach and proved better bounds provided more is assumed about the geometry of the variety. We adapt these techniques to bound the intersection of our hypersurfaces $\Phi_{\mathsf{t},\mathsf{c}}(\mathbb{F}_p)$ with Boolean hypercubes as (a proof of the bound is in a longer version of the paper [3])

$$|\Phi_{\mathsf{t},\mathsf{c}}(\mathbb{F}_p, 2)| = \left(\frac{2}{p}\right)^s |\Phi_{\mathsf{t},\mathsf{c}}(\mathbb{F}_p)| + O\left(p^{(s-1)/2}(\log p)^s + \frac{2^{s-1}\log p}{\sqrt{p}}\right). \tag{B.4}$$

The constant hidden in the asymptotic O notation may depend on s, but this dependence will be subsumed and removed in Equation (B.5). Our ultimate goal is to claim the right hand side is strictly less that 2^s, which would prove Conjecture 4. However, $p^{(s-1)/2}$ is too large and muddies the estimate. The bounds are good enough if the hypercube side length is extended to $b > 2$, since Fouvry [10] shows for every $\Phi_{\mathsf{t},\mathsf{c}}$, for large enough p,

$$|\Phi_{\mathsf{t},\mathsf{c}}(\mathbb{F}_p, b)| = \left(\frac{b}{p}\right)^s |\Phi_{\mathsf{t},\mathsf{c}}(\mathbb{F}_p)| + O\left(p^{(s-1)/2}(\log p)^s + \frac{b^{s-1}\log p}{\sqrt{p}}\right). \tag{B.5}$$

From the Schwartz-Zippel lemma bound Equation (B.3) on $|\Phi_{\mathsf{t},\mathsf{c}}(\mathbb{F}_p)|$,

$$|\Phi_{\mathsf{t},\mathsf{c}}(\mathbb{F}_p, b)| \leq \frac{ns2^s}{p} + O\left(p^{(s-1)/2}(\log p)^s + \frac{b^{s-1}\log p}{\sqrt{p}}\right).$$

For $b \gg p^{3/4}$, $|\Phi_{\mathsf{t},\mathsf{c}}(\mathbb{F}_p, b)| \ll b^s$. Fouvry's theorem applies to arbitrary varieties and the "for large enough p" clause is primarily in place to ensure the defining polynomials do not identically vanish modulo p. To us, $\Phi^p_{\mathsf{t},\mathsf{c}}$ is non zero, so the bounds should hold uniformly for all p. Therefore, with some work to ensure uniformity of bounds, these methods prove Conjecture 19 when relaxed to Boolean cubes of length growing $b \gg p^{3/4}$. A proof is deferred to the full version of this paper.

We believe the large error term in Equation (B.4) and Fouvry's theorem Equation (B.5) to be artefacts of proof techniques and not intrinsic to the quantities. The primary lesson we advocate from these arithmetic geometric techniques is qualitative and not quantitative. There should be no arithmetic obstruction to equidistribution of the zeroes of the hypersurfaces defined by our determinant polynomials in the Boolean hypercube, as claimed in the strong form of our conjecture.