New affine-invariant codes from lifting

Alan Guo* Swastik Kopparty † Madhu Sudan‡

November 22, 2012

Abstract

In this work we explore error-correcting codes derived from the "lifting" of "affine-invariant" codes. Affine-invariant codes are simply linear codes whose coordinates are a vector space over a field and which are invariant under affine-transformations of the coordinate space. Lifting takes codes defined over a vector space of small dimension and lifts them to higher dimensions by requiring their restriction to every subspace of the original dimension to be a codeword of the code being lifted. While the operation is of interest on its own, this work focusses on new ranges of parameters that can be obtained by such codes, in the context of local correction and testing. In particular we present four interesting ranges of parameters that can be achieved by such lifts, all of which are new in the context of affine-invariance and some may be new even in general. The main highlight is a construction of high-rate codes with sublinear time decoding. The only prior construction of such codes is due to Kopparty, Saraf and Yekhanin [33]. All our codes are extremely simple, being just lifts of various parity check codes (codes with one symbol of redundancy), and in the final case, the lift of a Reed-Solomon code.

We also present a simple connection between certain lifted codes and lower bounds on the size of "Nikodym sets". Roughly, a Nikodym set in $\mathbb{F}_q^m$ is a set S with the property that every point has a line passing through it which is almost entirely contained in S. While previous lower bounds on Nikodym sets were roughly growing as $q^m/2^m$, we use our lifted codes to prove a lower bound of $(1 - o(1))q^m$ for fields of constant characteristic.

*CSAIL, Massachusetts Institute of Technology, 32 Vassar Street, Cambridge, MA, USA. aguo@mit.edu. Research supported in part by NSF grants CCF-0829672, CCF-1065125, CCF-6922462, and an NSF Graduate Research Fellowship.

†Department of Computer Science & Department of Mathematics, Rutgers University, Piscataway NJ, USA. swastik.kopparty@rutgers.edu.

‡Microsoft Research New England, One Memorial Drive, Cambridge, MA 02139, USA. madhu@mit.edu.

1 Introduction

In this work we explore the "locality properties" of some highly symmetric codes constructed by "lifting" "affine-invariant" codes. We describe these terms below.

1.1 Basic terminology and background

We start with some standard coding theory preliminaries. Let $\mathbb{F}_q$ denote the finite field of cardinality q and for any finite set D, let $\{D \to \mathbb{F}_q\}$ denote the set of all functions from D to $\mathbb{F}_q$. In this work, a code on coordinate set D is a set of functions $\mathcal{F} \subseteq \{D \to \mathbb{F}_q\}$. A code $\mathcal{F}$ is said to be linear if it forms a vector space over $\mathbb{F}_q$, i.e., if for every $f, g \in \mathcal{F}$ and $\alpha \in \mathbb{F}_q$ the function $\alpha f + g \in \mathcal{F}$. We refer to $N = |D|$ as the length of the code. A second parameter of interest is the dimension of the code which is the dimension of $\mathcal{F}$ as a vector space. The dual of a code $\mathcal{F}$, denoted $\mathcal{F}^\perp$, is the set of functions $\{g : D \to \mathbb{F}_q | \langle f, g\rangle = 0 \ \forall f \in \mathcal{F}\}$, where $\langle f, g\rangle = \sum_{x \in D} f(x)g(x)$ denotes the standard inner product of vectors. Let $\mathrm{wt}(f) = |\{x \in D | f(x) \neq 0\}|$ denote the weight of f. Let $\delta(f, g) = |\{x \in D | f(x) \neq g(x)\}|/|D|$ denote the (normalized Hamming) distance between f and g. (So $\delta(f, g) = \mathrm{wt}(f - g)/|D|$.) We say f is δ-close to g if $\delta(f, g) \leq \delta$ and δ-far otherwise. We say f is δ-close to $\mathcal{F}$ if there exists $g \in \mathcal{F}$ that is δ-close to f and δ-far otherwise. We say $\mathcal{F}$ is a code of distance δ if every pair of distinct codewords in $\mathcal{F}$ are δ-far from each other. We use $\delta(\mathcal{F})$ to denote the maximum δ such that $\mathcal{F}$ is a code of distance δ.

In this work we explore some aspects of affine-invariant codes. In such codes the domain D is a vector space $\mathbb{F}_{q^n}^m$, i.e., an m-dimensional vector space over the n-dimensional extension field of the range $\mathbb{F}_q$. Let $Q = q^n$ and let $\mathbb{F}_Q$ denote the field of size Q. We say a function $A : \mathbb{F}_Q^m \to \mathbb{F}_Q^m$ is an affine function if $A(x) = M \cdot x + b$ for some matrix $M \in \mathbb{F}_Q^{m \times m}$ and vector $b \in \mathbb{F}_Q^m$. We say A is an affine permutation if M is invertible. A code $\mathcal{F} \subseteq \{\mathbb{F}_Q^m \to \mathbb{F}_q\}$ is said to be affine-invariant if for every affine permutation function $A : \mathbb{F}_Q^m \to \mathbb{F}_Q^m$ and for every $f \in \mathcal{F}$ the function $f \circ A$ given by $(f \circ A)(x) = f(A(x))$ is also in $\mathcal{F}$.[1]

Affine-invariant codes are of interest to us because they exhibit, under natural and almost necessary conditions, very good locality properties: they tend to be locally testable and locally correctible. We introduce these notions below. We say a code $\mathcal{F}$ is (k, δ)-locally correctible ((k, δ)-LCC) if there exists a probabilistic algorithm Corr that, given $x \in D$ and oracle access to a function $f : D \to \mathbb{F}_q$ which is δ-close to some $g \in \mathcal{F}$, makes at most k queries to f and outputs $g(x)$ with probability at least $2/3$. We say that $\mathcal{F}$ is (k, ϵ, δ)-locally testable ((k, ϵ, δ)-LTC) if $\mathcal{F}$ is a code of distance δ and there exists a probabilistic algorithm Test that, given oracle access to $f : D \to \mathbb{F}_q$, makes at most k queries to f and accepts $f \in \mathcal{F}$ with probability one, while rejecting f that is τ-far from $\mathcal{F}$ with probability at least $\epsilon \cdot \tau$.

1.2 This work: Motivation and Results

As noted above affine-invariant lead naturally to locally decodable codes and locally testable codes. In this work we use a certain lifting operation to exhibit codes with very good locality. We start by defining the lifting operation. For a function $f : \mathbb{F}_Q^m \to \mathbb{F}_q$ and set $S \subseteq \mathbb{F}_Q^m$ let $f|_S$ denote the restriction of f to the domain S.

[1] In some of the earlier works invariance is defined with respect to all affine functions and not just permutations. In Section A we show that the two notions are equivalent and so we use invariance with respect to permutations in this paper.

Definition 1.1 (Lifting). *For a code $\mathcal{F} \subseteq \{\mathbb{F}_Q^t \to \mathbb{F}_q\}$, and integer $m \geq t$ its m-dimensional lift* $\mathrm{Lift}_m(\mathcal{F}) \subseteq \{\mathbb{F}_Q^m \to \mathbb{F}_q\}$ *is the code*

$$\{f : \mathbb{F}_Q^m \to \mathbb{F}_q \mid f|_V \in \mathcal{F} \text{ for every } t\text{-dimensional affine subspace } V \subseteq \mathbb{F}_Q^m\}.$$

(Note that the definition above assumes some canonical way to equate t-dimensional subspaces of $\mathbb{F}_Q^m$ with $\mathbb{F}_Q^t$. But for affine-invariant families $\mathcal{F}$ the exact correspondence does not matter as long as the map is an isomorphism.)

The lift is a very natural operation on affine-invariant codes, and builds long codes from shorter ones. Indeed, lifts may be interpreted as the basic operation that leads to the construction of "(Generalized) Reed-Muller" codes, codes formed by m-variate polynomials over $\mathbb{F}_q$ of total degree at most d: Such codes are the "lifts" of t-variate polynomials of degree at most d, for $t = \lceil \frac{d+1}{q-q/p} \rceil$ where p is the characteristic of q. (This follows from the "characterization" of polynomials as proven in [30].) While the locality properties (testability and correctability) of Reed-Muller codes are well-studied [36, 3, 4, 35, 1, 30, 27, 10, 26], they are essentially the only rich class of symmetric codes that are well-studied. The only other basic class of symmetric codes that are studied seem to be sparse ones, i.e., ones with few codewords.

In this work we explore the lifting of codes as a means to building rich new classes of *dense* symmetric codes. (In Theorems 1.2 - 1.5 below we describe some of the codes we obtain this way, and contrast them with known results.) Along the way we also initiate a systematic study of lifts of codes. Lifts of codes were introduced first in [7], who explored it to prove negative results — specifically, to build "symmetric LDPC codes" that are not testable. (Their definition was more restrictive than ours, and also somewhat less clean.) Our work is the first to explore positive use of lifts.

We remark that all codes constructed by lifting have relative distance of at least Q^{-t} and are $(Q^t, Q^{-t}/3)$-LCC's and $(Q^t, \Omega(Q^{-2t}), Q^{-t})$-LTC's. The local correctability follows directly from their definition, while the local testability is a consequence of the main result of [31, Theorem 2.9]. (See also Proposition 2.10.) This general feature suffices for three of our code construction, while in the fourth case we have to analyze the decodability a little more carefully.

An example. Let q be a power of 2, let $d = (1-\delta)q$ and let us consider the lift of the set of all univariate polynomials over $\mathbb{F}_q$ of degree at most d to $\mathbb{F}_q^2$. Explicitly, we mean the code $\mathcal{F}$ consisting of all functions $f : \mathbb{F}_q^2 \to \mathbb{F}_q$ such that the restriction of f to any line of $\mathbb{F}_q^2$ is a univariate polynomial of degree at most d. $\mathcal{F}$ is an affine-invariant linear space.

By construction, it is clear that $\mathcal{F}$ has a lot of local structure; this leads to a simple local-correction algorithm for $\mathcal{F}$ based on picking random lines and performing noisy univariate polynomial interpolation (i.e., Reed-Solomon decoding). We will show that in fact $\mathcal{F}$ also has large dimension (when δ is small). This leads to a high-rate locally correctable code.

Which functions $f : \mathbb{F}_q^2 \to \mathbb{F}_q$ lie in $\mathcal{F}$? We will give an answer to this question later in the paper, in terms of the polynomial representation $f(X,Y) = \sum_{0 \leq i,j < q} a_{ij} X^i Y^j$. Here since we are interested in showing that $\dim \mathcal{F}$ is large, it will suffice for us to show that there are many linearly independent elements in $\mathcal{F}$. To do this, we will study when a monomial $g(X,Y) = X^i Y^j$ is in $\mathcal{F}$. Note that if we restrict g to a line $\ell(T) = (\alpha_1 T + \alpha_0, \beta_1 T + \beta_0)$, we get the function

$$g|_\ell(T) = (\alpha_1 T + \alpha_0)^i (\beta_1 T + \beta_0)^j = \sum_{r \leq i} \sum_{s \leq j} \alpha_1^r \alpha_0^{i-r} \beta_1^s \beta_0^{j-s} \binom{i}{r} \binom{j}{s} T^{r+s}.$$

This function will equal a univariate polynomial of degree at most d at all points of $\mathbb{F}_q$ if, when we reduce it mod $T^q - T$, we see no monomials of degree $> d$. Reducing the above polynomial mod $T^q - T$ amounts to replacing T^{r+s} in the above expression with $T^{r+s \pmod{q}}$ (where $a \pmod{q} = 0$ if $a = 0$ and $a \pmod{q} = b \in \{1, \ldots, q-1\}$ if $a \neq 0$ and $a = b \pmod{q-1}$). This will happen if i, j satisfy the following criterion: for every $r \leq i, s \leq j$, if $\binom{i}{r} \neq 0 \mod 2$ and $\binom{j}{s} \neq 0 \mod 2$, then $r + s \pmod{q} \leq d$. Via Lucas' theorem (which gives a characterization of when $\binom{a}{b} = 0 \mod 2$, we deduce that the monomial $X^i Y^j$ is in $\mathcal{F}$ if (i, j) lies in the set:

$$S = \{(i, j) \mid \forall r \leq_2 i, j \leq_2 s, r + s \pmod{q} \leq d\},$$

where $a \leq_2 b$ means that set of coordinates that equal 1 in the binary representation of a is a subset of the set of coordinates that equal 1 in the binary representation of b. Finally, an analysis of the set S shows that its size is $\geq (1 - \epsilon_\delta) \cdot q^2$, where $\epsilon_\delta \to 0$ as $\delta \to 0$. Thus the dimension of $\mathcal{F}$ is at least $(1 - \epsilon_\delta) \cdot q^2$.

We will formally treat this example in greater generality in a later section. Before that, we will build up the theory of lifts of multivariate codes. In Proposition 2.2 we will see that affine-invariant codes are completely characterized by (and in fact spanned by) the monomials in the code; thus the dimension of the code above exactly equals $|S|$.

The constructions. For simplicity most codes are described for the case of fields of characteristic two, while the construction does generalize to other fields. (The main exception is in Theorem 1.3 where the code is later applied in other cases, so we describe the more general result.) The codes in the first three theorems below are obtained by the lifting of the parity-check code. By making appropriate choices of Q and t we get codes with different locality (and distance). The fourth code works over large fields only and is obtained by lifting the Reed-Solomon code.

Our first code has constant locality k, for k being a power of 2. If the length of the code is N (in our setting $N = Q^m$), then the code has dimension $\Omega_k((\log N)^k)$.

Theorem 1.2. *For every positive integer t and $k = 2^t$, there exists a constant $c_k > 0$ such that for every positive integer m and $N = 2^m$, there exists a binary code of length N, dimension at least $c_k(\log N)^{k-2}$ which is a $(k-1, k^{-1}/3)$-LCC, and a $(k, \Omega(k^{-2}), k^{-1})$-LTC.*

To contrast this with other known codes, essentially the only symmetric binary code known in this regime is the Reed-Muller code, which has dimension $\Omega((\log N)^{\log k})$ for locality k. Thus our code has significantly greater dimension in this regime. Our results are also asymptotically optimal for affine-invariant codes, by a result of Ben-Sasson and Sudan [9] which shows that any affine-invariant code with such local correctability or testability must have dimension $(\log N)^{k+O(1)}$.

For local correctability, these codes asymptotically match the performance of best-known codes, which would be obtained by taking Generalized Reed-Muller codes over a field of size roughly k and then composing it with some binary code. Our codes are simpler to describe and the symmetry comes without any loss of parameters. Furthermore, for really small constants, say $k = 4$ or $k = 8$, these codes seem to be better than previously known locally correctible codes.

Our next two codes consider relatively large locality (growing with N). The advantage with these codes is that the redundancy (the difference between the length and the dimension) grows exceedingly slowly. The first of these two codes considers the setting where the locality is N^ϵ for some positive (but tiny) ϵ. In such cases, we get codes of dimension $N - N^{1-\epsilon'}$ where $\epsilon' > 0$ if $\epsilon > 0$. Thus the dimension is extremely close to the length.

Theorem 1.3. *For every $\epsilon > 0$ and prime p, there exists $\epsilon' > 0$ such that for infinitely many N, there is a p-ary code of length N, dimension $N - N^{1-\epsilon'}$ which is a $(N^{\epsilon}, N^{-\epsilon}/3)$-LCC and a $(N^{\epsilon}, \Omega(N^{-2\epsilon}), N^{-\epsilon})$-LTC.*

The codes from Theorem 1.3 are not new. These codes, and in particular their exact dimension are well-known in the literature in combinatorics [11, 38]. Their locality was first noted by Yekhanin [39] who noticed in particular that they are LCCs. Our main contribution is to note that these are (naturally) obtained from lifts. In the process we get that these are affine-invariant codes and so are also LTCs, a fact that was not known before. Finally, our bounds while cruder, give better asymptotic sense of the redundancy of these codes (and in particular note that the redundancy is sublinear in the code length).

We remark that these codes have very poor distance and very poor error-correcting capability. However, in the context of applications such as constructions of PCPs (probabilistically checkable proofs, see e.g., [2]) one does not need distance or error-correction capability per se. All one seems to need is the local correction and decoding capability. So the theorem above motivates the search for extremely efficient PCPs, where the difference between the length of the PCP and the length of the classical proof is sublinear, while allowing for sublinear query complexity. Such a result, if at all possible, would really be transformative in the use of PCPs as a positive concept. We also note that these codes play a useful role in giving lower bounds on the size of Nikodym sets — we will elaborate on this shortly.

Next, we consider codes of locality $\Omega(N)$, so linear in the length of the code. This range of parameters was motivated by the recent result of Barak et al. [5] who used such codes (with additional properties that we are not yet able to prove) to build "small-set expanders" with many "large eigenvalues". We won't describe the application here, but instead turn to the parameters they sought. They wanted codes of length N with locality ϵN and dimension $N - \mathrm{poly}(\log N)$. The codes they used were Reed-Muller codes. By exploring lifts we are able to suggest some alternate codes. These codes do have slightly better dimension, though unfortunately, the improvement is not asymptotically significant (and certainly not close to any known limits). Nevertheless we report the codes below.

Theorem 1.4. *For every $\epsilon > 0$ and for infinitely many N, there is a binary code of length N, dimension $N - O_{\epsilon}((\log N)^{\log 1/\epsilon})$, which is a $(\epsilon N, \frac{1}{3}(\epsilon N)^{-1})$-LCC and a $(\epsilon N, \Omega((\epsilon N)^{-2}), (\epsilon N)^{-1})$-LTC.*

We note that Barak et al. also require the codes to be "absolutely testable", a strong notion of testability that we do not achieve in this work. Indeed, it is unclear if the codes as described above will turn out to be absolutely testable. In followup work to ours, Haramaty et al. [25], do show that some codes constructed by the above principle (but not all) are absolutely testable. The dimensions of their codes are somewhere between those of Barak et al. and those from the above theorem (so are still of no asymptotic significance).

Finally, we describe the most interesting choice of parameters. Our final code has locality N^{δ} for arbitrarily small $\delta > 0$, while achieving dimension $(1 - \epsilon)N$ for arbitrarily small $\epsilon > 0$. While the dimension of this code is smaller than that of the codes of Theorem 1.4, it corrects a constant positive fraction of errors.

Theorem 1.5. *For every $\epsilon, \delta > 0$ there exists $\tau > 0$ such that for infinitely many N, there is a q-code of length N over $\mathbb{F}_Q$, for $Q \approx N^{\delta}$, of dimension $(1 - \epsilon)N$ which is a (N^{δ}, τ)-LCC, for some $q \approx N^{\delta}$.*

Till 2010, no codes achieving such a range of parameters were known. In particular no code was known that achieved dimension greater than $N/2$ while achieving $o(N)$ locality to correct constant fraction of errors. In 2010, Kopparty et al. [33] introduced what they called the "multiplicity codes" which manage to overcome the rate 1/2 barrier. Other than their codes, no other constructions were known that achieved the parameters of Theorem 1.5 and our construction provides the first alternate. We remark that while qualitatively our theorem matches theirs, the behavior of τ as a function of ϵ and δ is much worse in our construction. Nevertheless for concrete values of N, ϵ and δ our construction actually seems to perform quite well. Also, whereas in the basic codes of [33] are over larger alphabets than N, our codes are naturally over much smaller alphabets. (Of course, one can always use concatenation to reduce alphabet sizes, but such operations do result in a loss in concrete settings of parameters.)

Theorems 1.2-1.5 are proved in Section 3. While each of the codes above may be of interest on their own, the underlying phenomenon, of constructing codes with interesting parameters by lifting shorter codes is an important one. Given our belief that lifting is an important operation that deserves study, we also do some systematic analysis of lifts. In particular in this work we show that lifting of a base code essentially preserves distance. This preservation is not exact and we give examples proving this fact.

Bounds on the size of Nikodym sets. One of the applications of our results is to bounding, from below, the size of "Nikodym sets" over finite fields (of small characteristic). We define this concept before describing our results.

A set $N \subseteq \mathbb{F}_q^m$ is said to be a *Nikodym set* if every point x has a line passing through it such that all points of the line, except possibly the point x itself, are elements of N. More precisely, N is a Nikodym set if for every $x \in \mathbb{F}_q^m$ there exists $y \in \mathbb{F}_q^m \setminus \{\mathbf{0}\}$ such that $\{x + ty | t \in \mathbb{F}_q^*\} \subseteq N$.

Nikodym sets are closely related to "Kakeya sets" — the latter contain a line in every direction, while the former contain almost all of a line through every point. A lower bound for Kakeya sets was proved by Dvir [12] using the polynomial method and further improved by using "method of multiplicities " by Saraf and Sudan [37] and Dvir et al. [13]. Kakeya sets have seen applications connecting its study to the study of randomness extractors, esp. [14, 15]. Arguably Nikodym sets are about as natural in this connection as Kakeya sets.

Previous lower bounds on Kakeya sets were typically also applicable to Nikodym sets and led to bounds of the form $|N| \geq (1 - o(1))q^m/2^m$ where the $o(1)$ term goes to zero as $q \to \infty$[2]. In particular previous lower bounds failed to separate the growth of Nikodym sets from those of Kakeya sets. In this work we present a simple connection (see Proposition 4.1) that shows that existence of (high-rate) affine-invariant codes that are lifts of non-trivial univariate codes yield (large) lower bounds on the size of Nikodym sets. Using this connection we significantly improve the known lower bound on the size of Nikodym sets over fields of constant characteristic.

Theorem 1.6. *For every prime p, and every integer m, there exists $\epsilon = \epsilon(p, m) > 0$ such that for every finite field $\mathbb{F}_q$ of characteristic p, if $N \subseteq \mathbb{F}_q^m$ is a Nikodym set, then $|N| \geq q^m - q^{(1-\epsilon)m}$. In particular if $q \to \infty$, then $|N| \geq (1 - o(1)) \cdot q^m$.*

Thus whereas previous lower bounds on the size of Nikodym sets allowed for the possibility that the density of the Nikodym sets vanishes as m grows, ours show that Nikodym sets occupy almost

[2] In the $m = 2$ case, better bounds are known for Nikodym sets [16, 34].

all the space. One way to view our results is that they abstract the polynomial method in a more general way, and thus lead to stronger lower bounds (in some cases).

Previous work on affine-invariance. The study of invariance, and in particular affine-invariance, in property testing was initiated by Kaufman and Sudan [31] and there have been many subsequent works [9, 21, 22, 20, 32, 29, 6, 7, 28, 8, 24]. Most of the works, with the exceptions of [32, 29], study the broad class with the aim of characterizing all the testable properties. The exceptions, Kaufman and Wigderson [32] and Kaufman and Lubotzky [29], are the few that attempt to find new codes using invariance. While the performance of their codes is very good, unfortunately they do not seem to lead to local testability and the performance is too good to be locally decodable (or locally correctible). Our work seems to be the first in this context to explore new codes that do guarantee some locality properties.

A second, more technical, point of departure is that our work refocusses attention on invariance of "multivariate properties". Since the work of [31] most subsequent works focussed on univariate properties. While this study seemed to be without loss of generality, for the purpose of constructions it seems necessary to go back to the multivariate setting. One specific contribution in this direction is that we show that invariance under general affine-transformations and under affine-permutations lead to the same set of properties (see Section A).

Organization. In Section 2 we present some of the background material on affine-invariant codes and present some extensions in the multivariate setting. In Section 3 we describe our codes and analyze them. In Section 4 we describe our application to lower bounding Nikodym sets. In Section 5 we describe how distance of lifted codes behave. Some of the technical proofs are deferred to the appendix.

Version. A previous version of this paper appeared, as [23]. The main difference in the results is the addition, in this version, of lower bounds on the size of Nikodym sets (Theorem 1.6).

2 Preliminaries

In this section we describe some basic aspects of affine-invariant properties, specifically their degree sets. We mention in particular the fact that the size of degree sets determines the dimension of a given affine-invariant code. Finally we conclude by relating the degree set of a base code to the degree set of a lifted code. In later sections we will use this relationship to lower bound the size of the degree set of lifted codes, and thus lower bound their dimension. We note that the results of this section are described for general q (and not for the special case of $q = 2$).

For a function $f : \mathbb{F}_Q^m \to \mathbb{F}_q$, we associate with it the unique polynomial in $\mathbb{F}_Q[x_1, \ldots, x_m]$ of degree at most $Q - 1$ in each variable that evaluates to f. (We abuse notation by using the same notation to refer to a function and the associated polynomial.) For $\mathbf{d} = \langle d_1, \ldots, d_m \rangle$ and $\mathbf{x} = \langle x_1, \ldots, x_m \rangle$, let $\mathbf{x}^{\mathbf{d}}$ denote the monomial $\prod_{i=1}^m x_i^{d_i}$. For a function $f = \sum_{\mathbf{d}} c_{\mathbf{d}} \mathbf{x}^{\mathbf{d}}$, let its support, denoted $\mathrm{supp}(f)$, be the set of degrees with non-zero coefficients in f, i.e., $\mathrm{supp}(f) = \{\mathbf{d} \mid c_{\mathbf{d}} \neq 0\}$.

Definition 2.1 (Degree set)**.** *For a code $\mathcal{F} \subseteq \{\mathbb{F}_Q^m \to \mathbb{F}_q\}$, its degree set, denoted* $\mathrm{Deg}(\mathcal{F})$*, is the set* $\mathrm{Deg}(\mathcal{F}) = \cup_{f \in \mathcal{F}} \mathrm{supp}(f)$*. For a set $D \subseteq \{0, \ldots, Q-1\}^m$, let its code, denoted* $\mathrm{Fam}(\mathcal{F})$ *be the set* $\mathrm{Fam}(\mathcal{F}) = \{f : \mathbb{F}_Q^m \to \mathbb{F}_q \mid \mathrm{supp}(f) \subseteq D\}$.

For an affine-invariant code, its degree set uniquely determines the code and in particular the following proposition holds.

Proposition 2.2. *For linear affine-invariant codes $\mathcal{F} \subseteq \{\mathbb{F}_Q^m \to \mathbb{F}_q\}$, we have* $\mathrm{Fam}(\mathrm{Deg}(\mathcal{F})) = \mathcal{F}$.

We prove the proposition below. The proof uses some basic facts about linear affine-invariant codes that are proved in Section A. (We note that this would be the logical place to read/verify the contents.)

Proof. Trivially $\mathcal{F} \subseteq \mathrm{Fam}(\mathrm{Deg}(\mathcal{F}))$. For the other direction, consider $f \in \mathrm{Fam}(\mathrm{Deg}(\mathcal{F}))$. Express $f = \mathrm{Tr} \circ g$ (see, e.g., Proof of Lemma A.7), where $g \in \mathbb{F}_Q[\mathbf{x}]$ is chosen among all such to be minimal in its support. We have $\mathrm{supp}(g) \subseteq \mathrm{supp}(f) \subseteq \mathrm{Deg}(\mathcal{F})$. Suppose $g = \sum_{\mathbf{d} \in \mathrm{Deg}(\mathcal{F})} c_{\mathbf{d}} \mathbf{x}^{\mathbf{d}}$, then by Lemma A.7 we have $\mathrm{Tr}(c_{\mathbf{d}} \mathbf{x}^{\mathbf{d}}) \in \mathcal{F}$ for every $\mathbf{d}$. Now by linearity of $\mathcal{F}$ it follows that $\sum_{\mathbf{d}} \mathrm{Tr}(c_{\mathbf{d}} \mathbf{x}^{\mathbf{d}}) \in \mathcal{F}$, but by the linearity of the Trace function we have that this function is f. □

Our reason to study the degree sets is that the size of the degree set gives the dimension of a code exactly.

Proposition 2.3. *For a linear affine-invariant code $\mathcal{F} \subseteq \{\mathbb{F}_Q^m \to \mathbb{F}_q\}$, we have the dimension of $\mathcal{F}$ equals $|\mathrm{Deg}(\mathcal{F})|$.*

Proof. We generalize the proof of [6, Lemma 2.14] to the multivariate setting. For degree $\mathbf{d} \in \{0, \ldots, Q-1\}^m$, define $S(\mathbf{d}) = \{q^i \mathbf{d} \mid i \in \mathbb{Z}\}$. For every $\mathbf{d}, \mathbf{e}$, either $S(\mathbf{d}) = S(\mathbf{e})$ or $S(\mathbf{d}) \cap S(\mathbf{e}) = \varnothing$. Write $f \in \mathcal{F}$ as $f(\mathbf{x}) = \sum_{\mathbf{d} \in \mathrm{Deg}(\mathcal{F})} f_{\mathbf{d}} \mathbf{x}^{\mathbf{d}}$. Since $f^q = f$, it follows that $f_{q \cdot \mathbf{d}} = f_{\mathbf{d}}^q$ for all $\mathbf{d}$ and hence $f_{\mathbf{d}} \in \mathbb{F}_{q^{|S(\mathbf{q})|}}$. From each $S(\mathbf{d})$ pick a representative, and let S be the set of these representatives, so that $\mathrm{Deg}(\mathcal{F}) = \cup_{\mathbf{d} \in S} S(\mathbf{d})$ is a partition. Then we may write $f(\mathbf{x}) = \sum_{\mathbf{d} \in S} \mathrm{Tr}_{\mathbb{F}_{q^{|S(\mathbf{d})|}}, \mathbb{F}_q}(f_{\mathbf{d}} \mathbf{x}^{\mathbf{d}})$. For each $\mathbf{d} \in S$ there are $q^{|S(\mathbf{d})|}$ choices for $f_{\mathbf{d}}$, so the total number of choices for f is $\prod_{\mathbf{d} \in S} q^{|S(\mathbf{d})|} = q^{\sum_{\mathbf{d} \in S} |S(\mathbf{d})|} = q^{|\mathrm{Deg}(\mathcal{F})|}$ □

Next we attempt to describe how the degree set of a lifted code can be determined from the degree set of a base code. We start by mentioning a simple property of degree sets that will be quite useful in our analysis.

Let $(\mathrm{mod}^* Q)$ denote the operation that maps non-negative integers to the set $\{0, \ldots, Q-1\}$ as given by $a \ (\mathrm{mod}^* Q) = 0$ if $a = 0$ and $a \ (\mathrm{mod}^* Q) = b \in \{1, \ldots, Q-1\}$ if $a \neq 0$ and $a = b \pmod{Q-1}$. (Note that if $a \ (\mathrm{mod}^* Q) = b$, then $x^a = x^b \pmod{x^Q - x}$.)

For $Q = q^n$ and $\mathbf{e}, \mathbf{d} \in \{0, \ldots, Q-1\}^n$, we say that $\mathbf{e}$ is a q-shift of $\mathbf{d}$ if there exists j such that for every i, we have $e_i = q^j \cdot d_i \ (\mathrm{mod}^* Q)$. Note that $\mathbf{e}$ is a q-shift of $\mathbf{d}$ if and only if $\mathbf{d}$ is a q-shift of $\mathbf{e}$.

Proposition 2.4. *Let $\mathcal{F} \subseteq \{\mathbb{F}_Q^m \to \mathbb{F}_q\}$ be a linear affine-invariant code and let $D = \mathrm{Deg}(\mathcal{F})$ be its degree set. Then D is q-shift closed, i.e., if $\mathbf{d} \in D$ and $\mathbf{e}$ is a q-shift of $\mathbf{d}$ then $\mathbf{e} \in D$.*

Proof. Follows immediately from the fact that for every function $f : \mathbb{F}_Q^m \to \mathbb{F}_q$, we have $\mathbf{d} \in \mathrm{supp}(f)$ if and only if $\mathbf{e} \in \mathrm{supp}(f)$, which follows from the fact that $f(\mathbf{x})^{q^j} = f(\mathbf{x}) \pmod{\mathbf{x}^Q - \mathbf{x}}$ for every j. □

We now turn to identifying the degree sets of lifted codes. We start with the case of lifts of univariate codes, which are somewhat simpler to describe. The lifts of multivariate codes come from the same principles, but are messier to describe.

It turns out that the structure of the degree set (not every set D is the degree set of an affine-invariant code) is strongly influenced by the base p representation of its members, where p is the characteristic of q, the alphabet of our codes. We start with some notions related to such representations. For non-negative integers a and b, let $a^{(0)}, a^{(1)}, \ldots$, and $b^{(0)}, b^{(1)}, \ldots$, be their base p expansion, i.e., $0 \leq a^{(i)}, b^{(i)} < p$, $a = \sum_i a^{(i)} p^i$ and $b = \sum_i b^{(i)} p^i$. We say a is in the p-shadow of b, denoted $a \leq_p b$, if $a^{(i)} \leq b^{(i)}$ for every i. We extend the notion to vectors coordinate-wise. So for, $\mathbf{e}, \mathbf{d} \in \mathbb{Z}^n$, we say $\mathbf{e} \leq_p \mathbf{d}$ if $e_i \leq_p d_i$ for all $i \in [n]$.

Definition 2.5. *For a set $D \subseteq \{0, \ldots, Q-1\}$, its mth lift, denoted $\mathrm{Lift}_m(D)$ is given by*

$$\underset{m}{\mathrm{Lift}}(D) \triangleq \left\{ \mathbf{d} = \langle d_1, \ldots, d_m \rangle \in \{0, \ldots, Q-1\}^m | \forall \mathbf{e} \leq_p \mathbf{d}, \ \sum_{i=1}^{m} e_i \pmod^* Q \in D \right\}.$$

The following proposition makes the implied connection between lifts of codes and their degree sets explicit. We note that this proposition is implicit in [7].

Proposition 2.6. *For every linear affine-invariant code $\mathcal{F} \subseteq \{\mathbb{F}_Q \to \mathbb{F}_q\}$, and for every $m \geq 1$, we have $\mathrm{Lift}_m(\mathrm{Deg}(\mathcal{F})) = \mathrm{Deg}(\mathrm{Lift}_m(\mathcal{F}))$.*

Proof. Let $\mathbb{F} = \mathbb{F}_q$ and $\mathbb{K} = \mathbb{F}_Q$. In what follows we will use the notation $\mathbf{x}^{\mathbf{e}}$ to denote $\prod_{i=1}^n x_i^{e_i}$. And we use $\binom{\mathbf{d}}{\mathbf{e}}$ to denote $\prod_{i=1}^n \binom{d_i}{e_i}$.

Since $\mathcal{F}$ is linear, we have that there exist some $I \leq Q$ linear constraints given by $t_{i,j} \in \mathbb{K}$ and $\lambda_{ij} \in \mathbb{F}$ for $1 \leq i \leq I$ and $1 \leq j \leq J$ such that $f \in \mathcal{F}$ if and only if $\sum_{j \leq J} \lambda_{ij} f(t_{ij}) = 0$ for every $i \leq I$.

We now have the following equivalences:

$$\begin{aligned}
\mathbf{d} \in \mathrm{Deg}(\underset{m}{\mathrm{Lift}}(\mathcal{F})) \quad & \overset{\text{Lemma A.7}}{\iff} \quad \forall \lambda \in \mathbb{K} \quad \mathrm{Tr}(\lambda \mathbf{x}^{\mathbf{d}}) \in \underset{m}{\mathrm{Lift}}(\mathcal{F}) \\
& \iff \quad \forall \lambda \in \mathbb{K} \ \forall \mathbf{a} \in \mathbb{K}^m \ \forall \mathbf{b} \in \mathbb{K}^m \quad \mathrm{Tr}(\lambda (t \cdot \mathbf{a} + \mathbf{b})^{\mathbf{d}}) \in \mathcal{F} \\
& \iff \quad \forall \lambda, \mathbf{a}, \mathbf{b} \ \ \forall i \ \ \sum_j \lambda_{ij} \mathrm{Tr}(\lambda (t_{ij} \mathbf{a} + \mathbf{b})^{\mathbf{d}}) = 0 \\
& \overset{\text{Lemmas B.2, B.3}}{\iff} \quad \forall \lambda, \mathbf{a}, \mathbf{b} \ \ \forall i \ \ \mathrm{Tr}\left(\lambda \sum_{\mathbf{e} \leq_p \mathbf{d}} \binom{\mathbf{d}}{\mathbf{e}} \mathbf{a}^{\mathbf{e}} \mathbf{b}^{\mathbf{d}-\mathbf{e}} \sum_j \lambda_{ij} t_{ij}^{\sum_{\ell=1}^n e_\ell} \right) = 0 \\
& \iff \quad \forall \mathbf{a}, \mathbf{b} \ \ \forall i \ \ \sum_{\mathbf{e} \leq_p \mathbf{d}} \binom{\mathbf{d}}{\mathbf{e}} \mathbf{a}^{\mathbf{e}} \mathbf{b}^{\mathbf{d}-\mathbf{e})} \sum_j \lambda_{ij} t_{ij}^{\sum_\ell e_\ell} = 0 \\
& \iff \quad \forall \mathbf{e} \leq_p \mathbf{d} \ \ \forall i \ \ \sum_j \lambda_{ij} t_{ij}^{\sum_\ell e_\ell} = 0 \\
& \iff \quad \forall \mathbf{e} \leq_p \mathbf{d} \ \ \Sigma(\mathbf{e}) \pmod^* Q \in \mathrm{Deg}(\mathcal{F}).
\end{aligned}$$

□

We now extend the above definition and proposition to the case where the code being lifted is itself a multivariate one.

To this end we extend some of the notations from the previous parts to matrices. For matrices $\mathbf{A}, \mathbf{B} \in \mathbb{Z}^{n\times\ell}$ we say $\mathbf{A} \leq_p \mathbf{B}$ if $(\mathbf{A})_{ij} \leq_p (\mathbf{B})_{ij}$ for every pair $(i,j) \in [n] \times [\ell]$.

Next, we extend the notion to compare vectors to elements and matrices to vectors. For $\mathbf{e} \in \mathbb{Z}^\ell$ and $d \in \mathbb{Z}$ we say $\mathbf{e} \leq_p d$ if for every $\mathbf{f} \leq_p \mathbf{e}$ we have $\sum_{i\in[\ell]} f_i \leq_p d$. (This notion corresponds to the support of $(1 + \sum_{i=1}^{\ell} x_i)^d$: $\mathbf{x}^\mathbf{e}$ appears with a non-zero coefficient only if $\mathbf{e} \leq_p d$.) Extending to matrices and vectors, $\mathbf{A} \in \mathbb{Z}^{n\times\ell}$ with rows $(\mathbf{A})_j \in \mathbb{Z}^\ell$ and $\mathbf{d} = \langle d_1, \ldots, d_n\rangle \in \mathbb{Z}^n$ we say $\mathbf{A} \leq_p \mathbf{d}$ if $(\mathbf{A})_j \leq_p d_j$ for every $j \in [n]$.

Finally, we need one more piece of notation before defining the degree sets of multivariate lifts. For matrix $\mathbf{A} \in \mathbb{Z}^{n\times\ell}$, let $\Sigma(\mathbf{A}) \in \mathbb{Z}^\ell$ denote its row sum given by $\Sigma(\mathbf{A})_j = \sum_{i=1}^{n} (\mathbf{A})_{ij}$.

We are now ready to define the lifts of multivariate degree sets.

Definition 2.7 (Degree sets of lifts)**.** *For a set $D \subseteq \{0, \ldots, Q-1\}^t$, its mth* lift*, denoted* $\mathrm{Lift}_m(D)$*, is given by*

$$\{\mathbf{d} \in \{0, \ldots, Q-1\}^m \mid \forall\ \mathbf{E} \in \mathbb{Z}^{m\times t} \leq_p \mathbf{d},\ \text{we have } \Sigma(\mathbf{E}) \pmod^* Q) \in D\}.$$

The following proposition is the multivariate analog of Proposition 2.6.

Proposition 2.8. *For every linear affine-invariant code $\mathcal{F} \subseteq \{\mathbb{F}_Q^t \to \mathbb{F}_q\}$, and for every $m \geq t$, we have* $\mathrm{Lift}_m(\mathrm{Deg}(\mathcal{F})) = \mathrm{Deg}(\mathrm{Lift}_m(\mathcal{F}))$.

Proof. The proof is very similar to the proof of Proposition 2.6, with enriched notation. For a matrix $\mathbf{E}$, let $\mathbf{E}^\top$ denote the transpose of $\mathbf{E}$, so that $(\mathbf{E})_{ij} = (\mathbf{E}^\top)_{ji}$. For an integer d and vector $\mathbf{e} = \langle e_1, \ldots, e_t\rangle$, define $\binom{d}{\mathbf{e}} = \frac{d!}{e_1!\cdots e_t!(d-e_1-\cdots-e_n)!}$, the standard multinomial coefficient and also the coefficient of $\mathbf{x}^\mathbf{e}$ in the expansion of $(1 + \sum_{i=1}^{t} x_i)^d$. Extending this notation, for a vector $\mathbf{d} = \langle d_1, \ldots, d_m\rangle$ and a matrix $\mathbf{E} \in \mathbb{Z}^{m\times t}$ with rows $\mathbf{e}_1, \ldots, \mathbf{e}_m$, define $\binom{\mathbf{d}}{\mathbf{E}} = \prod_{i=1}^{m} \binom{d_i}{\mathbf{e}_i}$. We will use the fact that $\binom{\mathbf{d}}{\mathbf{E}} \not\equiv 0 \pmod p$ if and only if $\mathbf{E} \leq_p \mathbf{d}$ (see Lemma B.2). Finally, for two matrices $\mathbf{A}, \mathbf{E} \in \mathbb{Z}^{m\times t}$, define $\mathbf{A}^\mathbf{E} = \prod_{i,j} a_{ij}^{e_{ij}}$. For convenience, let $\mathbb{F} = \mathbb{F}_q$ and let $\mathbb{K} = \mathbb{F}_Q$.

Now, we begin the proof. There exist $\mathbf{t}_{ij} \in \mathbb{K}^t, \lambda_{ij} \in \mathbb{F}$ such that $f \in \mathcal{F} \iff \forall i\ \sum_j \lambda_{ij} f(\mathbf{t}_{ij}) = 0$. The assertion then follows from the following equivalences:

$$\begin{aligned}
\mathbf{d} \in \mathrm{Deg}(\underset{m}{\mathrm{Lift}}(\mathcal{F})) &\overset{\text{Lemma A.7}}{\iff} && \forall \lambda \in \mathbb{K} \quad \mathrm{Tr}(\lambda \mathbf{x}^\mathbf{d}) \in \underset{m}{\mathrm{Lift}}(\mathcal{F}) \\
&\iff && \forall \lambda \in \mathbb{K}\ \forall \mathbf{A} \in \mathbb{K}^{m\times t}\ \forall \mathbf{b} \in \mathbb{K}^m \quad \mathrm{Tr}(\lambda(\mathbf{A}\mathbf{x} + \mathbf{b})^\mathbf{d}) \in \mathcal{F} \\
&\iff && \forall \lambda, \mathbf{A}, \mathbf{b}\ \ \forall i\ \ \sum_j \lambda_{ij} \mathrm{Tr}(\lambda(\mathbf{A}\mathbf{t}_{ij} + \mathbf{b})^\mathbf{d}) = 0 \\
&\overset{\text{Lemmas B.2, B.3}}{\iff} && \forall \lambda, \mathbf{A}, \mathbf{b}\ \ \forall i\ \ \mathrm{Tr}\left(\lambda \sum_{\mathbf{E}\leq_p \mathbf{d}} \binom{\mathbf{d}}{\mathbf{E}} \mathbf{A}^\mathbf{E} \mathbf{b}^{\mathbf{d}-\Sigma(\mathbf{E}^\top)} \sum_j \lambda_{ij} {\mathbf{t}_{ij}}^{\Sigma(\mathbf{E})}\right) \\
&\iff && \forall \mathbf{A}, \mathbf{b}\ \ \forall i\ \ \sum_{\mathbf{E}\leq_p \mathbf{d}} \binom{\mathbf{d}}{\mathbf{E}} \mathbf{A}^\mathbf{E} \mathbf{b}^{\mathbf{d}-\Sigma(\mathbf{E}^\top)} \sum_j \lambda_{ij} {\mathbf{t}_{ij}}^{\Sigma(\mathbf{E})} = 0 \\
&\iff && \forall \mathbf{E} \leq_p \mathbf{d}\ \ \forall i\ \ \sum_j \lambda_{ij} {\mathbf{t}_{ij}}^{\Sigma(\mathbf{E})} = 0 \\
&\iff && \forall \mathbf{E} \leq_p \mathbf{d}\ \ \Sigma(\mathbf{E}) \pmod^* Q) \in \mathrm{Deg}(\mathcal{F})
\end{aligned}$$

□

The definition of $\mathrm{Lift}_m(D)$ is somewhat cumbersome and not easy to work with. However in the upcoming sections we will try to gain some combinatorial insights about it to derive bounds on the dimension of the codes of interest.

Finally, before concluding we mention explicitly the locality properties of lifted codes. We start with a simple observation.

Proposition 2.9. *Let $\mathcal{F} \subsetneq \{\mathbb{F}_Q^m \to \mathbb{F}_q\}$ be a linear affine-invariant code. The $\delta(\mathcal{F}) \geq 2 \cdot Q^{-m}$.*

Proof. For $\mathbf{a} \in \mathbb{F}_Q^m$ let $\Delta_{\mathbf{a}} : \mathbb{F}_Q^m \to \mathbb{F}_q$ be the function satisfying $\Delta_{\mathbf{a}}(\mathbf{a}) = 1$ and $\Delta_{\mathbf{a}}$ is zero everywhere else. For contradiction assume $\Delta_{\mathbf{a}} \in \mathcal{F}$ for some $\mathbf{a} \in \mathbb{F}_Q^m$. But then by affine-invariance we have $\Delta_{\mathbf{b}} \in \mathcal{F}$ for every $\mathbf{b} \in \mathbb{F}_Q^m$ and then by linearity we have every function in $\{\mathbb{F}_Q^m \to \mathbb{F}_q\}$ is contained in $\mathcal{F}$, contradicting our hypothesis on $\mathcal{F}$. □

Proposition 2.10. *Let $\mathcal{F} \subsetneq \{\mathbb{F}_Q^t \to \mathbb{F}_q\}$ be a linear affine-invariant code. Let $\mathcal{L} = \mathrm{Lift}_m(\mathcal{F})$ be its m-ary lift. Then $\mathcal{L}$ is a $(Q^t - 1, \frac{1}{3}Q^{-t})$-LCC and a $(Q^t, \Omega(Q^{-2t}), Q^{-t})$-LTC.*

Proof. Given $f : \mathbb{F}_Q^m \to \mathbb{F}_q$ that is $Q^{-t}/3$-close to $p \in \mathcal{L}$ and $\mathbf{a} \in \mathbb{F}_Q^m$, the local decoding algorithm works as follows: Pick random linearly independent $\mathbf{b}_1, \ldots \mathbf{b}_t \in \mathbb{F}_Q^m$ and let $h : \mathbb{F}_Q^t \to \mathbb{F}_q$ be given by $h(\mathbf{0}) = 0$ and $h(u_1, \ldots, u_t) = f(\mathbf{a} + u_1\mathbf{b}_1 + \cdots + u_t\mathbf{b}_t)$ for all other $u_1, \ldots, u_t$. Compute $g \in \mathcal{F}$ such that $g(\mathbf{u}) = h(\mathbf{u})$ for all $\mathbf{u} \in \mathbb{F}_Q^t \setminus \{\mathbf{0}\}$ and output $g(\mathbf{0})$.

It is clear that the decoder makes at most $Q^t - 1$ queries. We show that the decoder succeeds with high probability. Let $p \in \mathcal{F}$ satisfy $\delta(p, f) \leq Q^{-t}/3$. Let A be the t-dimensional subspace $A = \{\mathbf{a} + u_1\mathbf{b}_1 + \cdots + u_t\mathbf{b}_t \mid \mathbf{u} \in \mathbb{F}_Q^t\}$. For every $\mathbf{u} \in \mathbb{F}_Q^t \setminus \{\mathbf{0}\}$ we have $\Pr_{\mathbf{b}_1,\ldots,\mathbf{b}_t}[h(\mathbf{u}) \neq p|_A(\mathbf{u})] \leq Q^{-t}/3$. By a union bound $\Pr_{\mathbf{b}_1,\ldots,\mathbf{b}_t}[\exists \mathbf{u} \in \mathbb{F}_Q^t \setminus \{\mathbf{0}\} | h(\mathbf{u}) \neq p|_A(\mathbf{u})] \leq (Q^t - 1)/(3Q^t) < 1/3$. So, with probability at least $2/3$, we have that $p|_A \in \mathcal{F}$ agrees with h on all of $\mathbb{F}_Q^t - \mathbf{0}$. Furthermore, by the fact that $\delta(\mathcal{F}) \geq 2Q^{-t}$ (Proposition 2.9), $p|_A$ is the unique such function with this property. It follows that the decoder outputs $p|_A(\mathbf{0}) = p(\mathbf{a})$ with probability at least $2/3$ as desired.

The local testability follows directly from [31, Theorem 2.9]. □

3 Constructions

3.1 Codes of constant locality

In this section we prove Theorem 1.2 which promised binary codes of locality k and length N with dimension $\Omega_k(\log N)^{k-2}$.

The Code: Fix $k = Q = 2^\ell$ and $N = 2^{m\ell}$. Let $\mathcal{F}_1 \subseteq \{\mathbb{F}_Q \to \mathbb{F}_2\}$ be the code given by $\{f : \mathbb{F}_Q \to \mathbb{F}_2 \mid \sum_{\alpha \in \mathbb{F}_Q} f(\alpha) = 0\}$. Let $\mathcal{L}_1 = \mathrm{Lift}_m(\mathcal{F}_1)$. In what follows we verify that $\mathcal{L}_1$ has the properties claimed in Theorem 1.2.

We start with some obvious aspects.

Proposition 3.1. *$\mathcal{L}_1$ is a binary code of length N and a $(k-1, k^{-1}/3)$-LCC and a $(k, \Omega(k^{-2}), k^{-1})$-LTC.*

Proof. The length is immediate from the construction. The local correctability and testability follow from Proposition 2.10. □

The main aspect to be verified is the dimension of $\mathcal{L}_1$. We first describe the degree set of $\mathcal{F}_1$.

Claim 3.2. $\mathrm{Deg}(\mathcal{F}_1) = \{0, \ldots, Q-2\}$.

Proof. Write $f : \mathbb{F}_Q \to \mathbb{F}_2$ as $f(x) = \sum_{d=0}^{Q-1} f_d x^d$. Then $\sum_{\alpha \in \mathbb{F}_Q} f(\alpha) = \sum_{\alpha \in \mathbb{F}_Q} \sum_{d=0}^{Q-1} f_d \alpha^d = \sum_{d=0}^{Q-1} f_d \left(\sum_{\alpha \in \mathbb{F}_Q} \alpha^d \right) = -f_{Q-1}$ where we have used the fact that $\sum_{\alpha \in \mathbb{F}_Q} \alpha^d = -1$ if $d = Q-1$ and is equal to 1 otherwise. Therefore $f \in \mathcal{F}_1$ if and only if $\deg(f) < Q-1$. □

Remark: Note that the proof above applies without change to the case of the range being $\mathbb{F}_q$, for any q, provided $\mathbb{F}_Q$ extends $\mathbb{F}_q$.

The next claim interprets the definition of $\mathrm{Lift}_m(D)$ in our setting.

Claim 3.3. $\mathbf{d} \in \{0, \ldots, Q-1\}^m$ *is contained in* $\mathrm{Lift}_m(\mathrm{Deg}(\mathcal{F}_1))$ *if and only if for every* $\mathbf{e} \leq_2 \mathbf{d}$ *we have* $\sum_{i=1}^m e_i \pmod^* Q \neq Q-1$.

Proof. Follows immediately by applying Proposition 2.6 to Claim 3.2. □

Given the claim, it is simple to get a lower bound on the dimension of our code.

Lemma 3.4. *The dimension of* $\mathcal{L}_1$ *is at least* $\binom{m}{Q-2}$.

Proof. For $S \subseteq [m]$ let $\mathbf{d}_S$ denote the vector that is one on coordinates from S and zero outside. It is clear that for $|S| \leq Q-2$, $\mathbf{d}_S \in \mathrm{Lift}_m(\mathrm{Deg}(\mathcal{F}_1))$ and there are at least $\binom{m}{Q-2}$ such sets. □

Proof of Theorem 1.2. Theorem 1.2 follows Proposition 3.1 and Lemma 3.4 and plugging the values of m and Q from the construction. Specificalle we have that the dimension of the code is at least $\binom{m}{Q-2} \geq \frac{1}{k^{k-2}k!}(\log N)^{k-2}$. So the theorem follows for $c_k = \frac{1}{k^{k-2}k!}$. □

3.2 Codes of sublinear locality

Next we turn to Theorem 1.3, which asserts the existence of codes of locality N^ϵ with dimension $N - N^{1-\epsilon'}$.

The Code: Given $\epsilon > 0$ and prime p, let $m = \lceil 1/\epsilon \rceil$. Let ℓ be an integer such that $p^{m\ell} \geq N$. Let $Q = p^\ell$. Let $\mathcal{F}_2 \subseteq \{\mathbb{F}_Q \to \mathbb{F}_p\}$ be the code $\{f : \mathbb{F}_Q \to \mathbb{F}_p \mid \sum_{\alpha \in \mathbb{F}_Q} f(\alpha) = 0\}$. Let $\mathcal{L}_2 = \mathrm{Lift}_m(\mathcal{F}_2)$.

As usual we get the following proposition.

Proposition 3.5. $\mathcal{F}_2$ *is a* p*-ary code of length at least* N *and locality at most* N^ϵ. *Specifically it is a* $(N^\epsilon, \Omega(N^{-\epsilon}))$*-LCC and a* $(N^\epsilon, \Omega(N^{-2\epsilon}), N^{-\epsilon})$*-LTC.*

We now turn to the task of analyzing the dimension of this code. We first describe the degree sets of $\mathcal{F}_2$ and $\mathcal{L}_2$.

Claim 3.6. $\mathrm{Deg}(\mathcal{F}_2) = \{0, \ldots, Q-2\}$ *and*

$$\mathrm{Deg}(\mathcal{L}_2) = \{\mathbf{d} \in \{0, \ldots, Q-1\}^m \mid \forall \mathbf{e} \leq_p \mathbf{d}, \sum_i e_i \pmod^* Q \neq Q-1\}.$$

Proof. The first part follows from the proof of Claim 3.2 (see the remark following the proof). The second part follows immediately from Proposition 2.8. □

Lemma 3.7. *The dimension of* $\mathcal{L}_2$ *is at least* $N - N^{1-\epsilon'}$ *for some* $\epsilon' = \Omega(2^{-2/\epsilon})$.

Proof. Let $D = \mathrm{Deg}(\mathcal{F}_2)$. Let $\mathbf{e} = \langle e_1, \ldots, e_m \rangle$ and $e_i^{(0)}, e_i^{(1)}, \ldots, e_i^{(\ell-1)}$ denote the p-ary expansion of e_i.

Claim 3.8. *If there exists integer $s \in \{0, \ldots, \ell-1\}$ such that for every $i \in [m]$ and every $j \in [1 + \lceil \log m \rceil]$ we have $e_i^{(s+j \pmod \ell)} = 0$, then $\mathbf{e} \in \mathrm{Lift}_m(D)$.*

Proof. Recall, by Proposition 2.4 that $\mathbf{e} \in \mathrm{Lift}_m(D)$ if and only if $\mathbf{e}' \in D$ for every $\mathbf{e}'$ that is a p-shift of $\mathbf{e}$. Thus without loss of generality we can assume (by shifting $\mathbf{e}$ appropriately), that the block of zeroes are the most significant digits in the e_i's. (i.e., $s = \ell - \lceil \log m \rceil - 2$.)

With this assumption, we now have $e_i < p^{\ell - \log m - 1} < Q/(pm) < (Q-1)/m$. We thus conclude that for every $\mathbf{f} \leq_2 \mathbf{e}$, $\sum_{i=1}^m f_i \leq \sum_{i=1}^m e_i < Q - 1$ and so (by Claim 3.6) $\mathbf{e} \in \mathrm{Lift}_m(D)$. □

The lemma follows by an easy counting argument. Let $t = 1 + \lceil \log m \rceil$. We partition the set $[\ell]$ into ℓ/t blocks of t successive integers each. For each such block the number of possible assignments of digits that do not make the entire block zero in each e_i is $p^{mt} - 1$. Thus the total number of vectors $\mathbf{e}$ that do not have any of these blocks set to zero is $(p^{mt}-1)^{\ell/t} = p^{m\ell}(1 - p^{-mt})^{\ell/t} \approx p^{m\ell} e^{\ell/(tp^{mt})} = p^{m\ell(1-\Omega(1/mtp^{mt}))} = N^{1-\epsilon'}$ for $\epsilon' = 1/(mtp^{mt})$. Recalling that $\epsilon = 1/m$, we have $\epsilon' = \Omega(p^{-2/\epsilon})$. The lemma follows by noting that if $\mathbf{e} \notin \mathrm{Lift}_m(D)$ then in each of these blocks it must be non-zero somewhere (by Claim 3.8 above). □

Proof of Theorem 1.3. Theorem 1.3 follows immediately from Proposition 3.5 and Lemma 3.7. □

3.3 Codes of linear locality

Finally, we prove Theorem 1.4, which claims codes of locality ϵN with dimension $N - \mathrm{poly}\log N$. This construction is different from the previous two in that here we lift a multivariate code, whereas in both previous constructions we lifted univariate codes.

The Code: Let $\ell = \lceil \log 1/\epsilon \rceil$ (so that $2^{-\ell} \leq \epsilon$). Let $Q = 2^\ell$. For integer m let $N = 2^{m\ell}$ and let $t = m - 1$. Let $\mathcal{F}_3 \subseteq \{\mathbb{F}_Q^t \to \mathbb{F}_2\}$ be given by $\mathcal{F}_3 = \{f : \mathbb{F}_Q^t \to \mathbb{F}_2 \mid \sum_{\alpha \in \mathbb{F}_Q^t} f(\alpha) = 0\}$. Let $\mathcal{L}_3 = \mathrm{Lift}_m(\mathcal{F}_3)$.

Proposition 3.9. *$\mathcal{L}_3$ is a code of block length N with locality ϵN. Specifically, it is a $(\epsilon N, \frac{1}{3}(\epsilon N)^{-1})$-LCC and a $(\epsilon N, (\epsilon N) - 2, (\epsilon N)^{-1})$-LTC.*

The proposition below asserts that every degree except the vector that is $Q-1$ in every coordinate is in the degree set of $\mathcal{F}_3$. (Here $(Q-1)^t$ denotes the t-tuple all of whose entries is $Q-1$, rather than $(Q-1)$ exponentiated to the t-th power).

Proposition 3.10. $\mathrm{Deg}(\mathcal{F}_3) = \{0, \ldots, Q-1\}^t - \{(Q-1)^t\}$.

Proof. Write $f : \mathbb{F}_Q^t \to \mathbb{F}_2$ as $f(\mathbf{x}) = \sum_{\mathbf{d} \in \{0,\ldots,Q-1\}^t} f_{\mathbf{d}} \mathbf{x}^{\mathbf{d}}$. Then $\sum_{\alpha \in \mathbb{F}_Q^t} f(\alpha) = \sum_{\alpha \in \mathbb{F}_Q^t} \sum_{\mathbf{d}} f_{\mathbf{d}} \mathbf{x}^{\mathbf{d}} = \sum_{\mathbf{d}} f_{\mathbf{d}} \left(\sum_{\alpha \in \mathbb{F}_Q^t} \alpha^{\mathbf{d}} \right) = \sum_{\mathbf{d}} f_{\mathbf{d}} \prod_{i=1}^t \left(\sum_{\alpha \in \mathbb{F}_Q} \alpha^{d_i} \right) = (-1)^t f_{(Q-1)^t}$ where we have used the fact that $\sum_{\alpha \in \mathbb{F}_Q} \alpha^d = -1$ if $d = Q-1$ and is equal to 0 otherwise. Therefore $f \in \mathcal{F}_3$ if and only if $f_{(Q-1)^t} = 0$. □

While in general degree sets of lifts of multivariate families are not easy to characterize, in this particular case we have a clean characterization of the degree set.

Given $\mathbf{e} = \langle e_1, \ldots, e_m \rangle$ let $e_i^{(j)}$ denote the jth bit in the binary expansion of e_i. Let $M(\mathbf{e})$ denote the $m \times \ell$ matrix with entries $M(\mathbf{e})_{i,j} = e_i^{(j)}$.

Lemma 3.11. $\mathbf{e} \in \mathrm{Lift}_m(\mathrm{Deg}(\mathcal{F}_3))$ *if and only if there exists a column in* $M(\mathbf{e})$ *with at least two zeroes.*

Proof. As in the proof of Lemma 3.7 we have that $\mathbf{e} \in \mathrm{Lift}_m(\mathrm{Deg}(\mathcal{F}_3))$ if and only if $2\mathbf{e} \pmod^* Q) \in \mathrm{Lift}_m(\mathrm{Deg}(\mathcal{F}_3))$. So without loss of generality we can assume that $\mathbf{e}$ is shifted so that the two zeroes are in the most significant bits. Thus we have that $m-2$ of the e_i's, say $e_1, \ldots, e_{m-2}$, are at most $Q-1$ and the remaining two are at most $Q/2-1$. We thus have that $\sum_{i=1}^{m} e_i < (m-1)Q - 1$. Using this and applying Proposition 2.8 it is easy to verify that $\mathbf{e}$ is not in $\mathrm{Lift}_m(\mathrm{Deg}(\mathcal{F}_3))$. □

The following lemma now follows by simple counting.

Lemma 3.12. *The dimension of* $\mathcal{L}_3$ *is* $2^{m\ell} - (m+1)^\ell$.

Proof of Theorem 1.4. Follows by plugging in the values for the parameters, specifically by setting $\ell = \log 1/\epsilon$ and $m = (\log N / \log 1/\epsilon)$. We get that the dimension of $\mathcal{L}_3$ is $N - (1 + \log N/\log 1/\epsilon)^{\log 1/\epsilon}$. □

We remark that the construction in [5] is very close in parameters. In their construction (i.e., the Reed-Muller codes) the matrix $M(\mathbf{e})$ must have at least $\ell + 1$ zeroes. Since any such matrix must have two zeroes in a single column it follows that every matrix their construction admits is also admissible in ours, while our allow for other matrices also. However the difference between the length and dimension is at most a constant factor (depending on ℓ). (More precisely, the dimension of their code is $2^{m\ell} - \sum_{i=0}^{\ell} \binom{m\ell}{i} \approx 2^{m\ell} - (em)^\ell$.) Of course, for their application the code needs to have much better local testability than given here. But the local testability given here is just what follows immediately from the definition and previous works, and it is quite possible that better bounds can be achieved by more careful examination of this code.

3.4 High-rate high-error LCCs

Finally, we prove Theorem 1.5. This construction is a departure from the others in that the code is not binary, and the code being lifted is not the parity check code. Finally the decoding algorithm is a bit more complex to explain, though even this algorithm is by now folklore.

The code itself is a generalization of the classical multivariate polynomial code. Here we consider the set of all functions $f : \mathbb{F}_q^m \to \mathbb{F}_q$ such that the restriction of f to any line has degree d. As is well known, every multivariate polynomial of degree at most d is such a function. The remarkable fact is that if q has small characteristic, then there are many more such functions.

The Code: Recall that we are given δ, ϵ and some N_0 and we wish a code of length $N \geq N_0$ of dimension $(1-\epsilon)N$ and locality N^δ. Let $m = \lceil 1/\delta \rceil$ and s be such that $Q = 2^s \geq N_0^\delta$. Let $b = 1 + \lceil \log m \rceil$ and $c = \lceil b 2^{bm} \log 1/\epsilon \rceil$. Let $\gamma = 2^{-c}$ and $\tau = \gamma/6$ (so that $6\tau \leq \gamma \leq \epsilon^{-(1+\lceil \log m \rceil) 2^{m(1+\lceil \log m \rceil)}}$ and let $d = (1-2^{-c})Q$. Let $\mathcal{F}_4 = \{f : \mathbb{F}_Q \to \mathbb{F}_Q \mid \deg(f) \leq d\}$. Let $\mathcal{L}_4 = \mathrm{Lift}_m(\mathcal{F}_4)$. In words, it is the set of all degree m-variate functions that have degree at most d when restricted to a line.

Decoding: The general idea for decoding $\mathcal{L}_4$ is the same as that for multivariate polynomials, and in particular the algorithm from Gemmell et al. [18].

Given $f : \mathbb{F}_Q^m \to \mathbb{F}_Q$ that is τ-close to $p \in \mathcal{L}_4$ and $\mathbf{a} \in \mathbb{F}_Q^m$, the decoding algorithm works as follows: Pick a random $\mathbf{b} \in \mathbb{F}_Q^m$ and let $h : \mathbb{F}_Q \to \mathbb{F}_Q$ be given by $h(t) = f(\mathbf{a} + t\mathbf{b})$. Compute, using a Reed-Solomon decoder (see, for instance, [19, Appendix]), a polynomial $g \in \mathbb{F}_Q[t]$ of degree at most d such that $\delta(h, g) < \gamma/2$. Output $g(0)$.

Lemma 3.13. *$\mathcal{L}_4$ is a code of block length N with locality N^δ. Specifically, it is a $(N^\delta, \gamma/6)$-LCC.*

Proof. Let $L = \{\mathbf{a} + t\mathbf{b} \mid t \in \mathbb{F}_Q - \{0\}\}$ be the line through $\mathbf{a}$ with slope $\mathbf{b}$. We first claim that with probability at least $2/3$, the line L contains fewer that $\gamma/2$ fraction errors (i.e., points $t \neq 0$ such $h(t) \neq p|_L(t)$).

Claim 3.14. *For every* $\mathbf{a}$, $\Pr_{\mathbf{b}}[\delta(h, p|_L) \geq \gamma/2] < 2\tau/\gamma$.

The above claim follows easily from an application of Markov's inequality. Next we note that if the fraction of errors on L is less than $\gamma/2$ then the decoder satisfies $g = p|_L$ and so outputs $g(0) = p|_L(0) = p(\mathbf{a})$ as desired. □

Next we turn to the analysis of the dimension of $\mathcal{L}_4$ which is similar to the analysis of $\mathcal{L}_2$. First we note the obvious fact.

Proposition 3.15. $\mathrm{Deg}(\mathcal{F}_4) = \{0, \ldots, d\}$ *and*

$$\mathrm{Deg}(\mathcal{L}_4) = \left\{ \mathbf{d} \in \{0, \ldots, Q-1\}^m \mid \forall\ \mathbf{e} \leq_2 \mathbf{d}, \sum_{i=1}^m e_i \pmod{^* Q} \in \{0, \ldots, d\} \right\}.$$

Lemma 3.16. *The dimension of $\mathcal{L}_4$ is at least $(1 - \epsilon)N$.*

Proof. For non-negative integer b, let $b^{(j)}$ denote its binary expansion so that $b = \sum_j b^{(j)} 2^j$. Recall $d = (1 - 2^{-c})Q$. Letting $d^{(j)}$ denoting its binary expansion, we note an integer $e \in \{0, \ldots, Q-1\}$ is at most d if (and only if) one of the bits $e^{(s-c)}, \ldots, e^{(s-1)}$ is zero. We use this to reason about $\mathrm{Deg}(\mathcal{L}_4)$.

Let $\mathbf{d} = \langle d_1, \ldots, d_m \rangle$ and let $d_i^{(j)}$ denote the jth bit in the binary expansion of d_i.

Claim 3.17. *Let $b = 1 + \lceil \log m \rceil$. If there exists $j \in \{s - c, \ldots, s - b\}$ such that for every $i \in [m]$ and every $\ell \in \{0, \ldots, b-1\}$ we have $d_i^{(j+\ell)} = 0$, then $\mathbf{d} \in \mathrm{Deg}(\mathcal{L}_4)$.*

Proof. Let $\mathbf{e} = \langle e_1, \ldots, e_m \rangle \leq_2 \mathbf{d}$ and let $e = \sum_{i=1}^m e_i \pmod{^* Q}$. We claim that $e^{(j+b-1)} = 0$, which suffices to show that $e \leq d$. Let $\overline{e}_i = 2^{s-(j+b)} e_i \pmod{^* Q}$ for all $i \in [m]$, and let $\overline{e} = \sum_{i=1}^m \overline{e}_i \pmod{^* Q}$. For every $i \in [m]$ and every $k \in [s]$, $e_i^{(k)} = \overline{e}_i^{(k+s-(j+b) \pmod s)}$ and similarly $e^{(k)} = \overline{e}^{(k+s-(j+b) \pmod s)}$. Therefore it suffices to show that $\overline{e}^{(s-1)} = 0$ or equivalently $\overline{e} < 2^{s-1}$. By our assumption on $\mathbf{d}$, $e_i^{(j+\ell)} = 0$ for all $\ell \in \{0, \ldots, b-1\}$, so $\overline{e}_i^{(k)} = 0$ for all $k \in \{s-b, \ldots, s-1\}$ and thus $\overline{e}_i < 2^{s-b}$ for all $i \in [m]$. By our choice of b, $m \leq 2^{b-1}$, and thus $\sum_{i=1}^m \overline{e}_i < m 2^{s-b} \leq 2^{s-1}$. □

We now consider picking $\mathbf{d}$ at random. By partitioning the c most significant bits into disjoint blocks of b bits each, we get that any such block is all zero with probability at least 2^{-mb}. Thus the probability there exists a block which is all zero is at least $1 - (1 - 2^{-mb})^{c/b} \geq 1 - e^{-c/(b2^{mb})}$. By choice of c we have that $c/(b2^{mb}) \geq \ln(1/\epsilon)$ and so $e^{-c/(b2^{mb})} \leq \epsilon$ and thus the dimension is lower bounded by $(1 - \epsilon)N$. □

Proof of Theorem 1.5. Follows immediately from Lemmas 3.13 and 3.16. □

We remark that the construction of this section is somewhat contrary to folk belief, which tends to suggest that generalized Reed-Muller codes (evaluations of m-variate polynomials of degree at most d) are equivalently defined by requiring that their restriction to lines are Reed-Solomon codewords (evaluations of univariate degree d polynomials). As pointed out earlier this folk statement is true only with some restrictions on d and Q, and our construction benefits by violating the restrictions. While the fact that there exist functions that are not degree d polynomials, for $d \geq Q - Q/p$, which are degree d polynomials on every line has been known for a while [17], presumably it was suspected that the effect on the dimension of the lifted family was negligible. Fortunately for this work, this presumption turned out to be false.

We also give below an example of some concrete setting of parameters for which this construction works.

Example 3.18. For every $N = 2^{2n}$, for $n \geq 7$, there exists a code of length N over the alphabet $\mathbb{F}_{2^n}$ of dimension $.77N$ that is decodable from 0.26% fraction errors with $\sqrt{N}$ queries

The example is obtained by setting $c = 6$, $m = 2$ and $Q = 2^n$ in the construction. The fraction of errors is $2^{-6}/6 \approx 0.26\%$. The rate follows from the following claim.

Claim 3.19. *The dimension of the code is at least* $((4^c - (5/4)3^c + 1/4)/4^c)N$.

While the error-correction rate of the code is smaller than that in [33], it does seem to start working at much smaller lengths and with much smaller alphabet sizes.

4 Nikdoym sets

A *Nikodym set* $N \subseteq \mathbb{F}_q^m$ is a set such that for all $x \in \mathbb{F}_q^m$, there exists $y \in \mathbb{F}_q^m$ such that the punctured line $\{x + ty \mid t \in \mathbb{F}_q \setminus \{0\}\} \subseteq N$.

The following proposition strengthens and generalizes the result usually obtained via the polynomial method [12].

Proposition 4.1. *If $\mathcal{L} \subset \{\mathbb{F}_q^m \to \mathbb{F}_q\}$ is the lift of some univariate linear affine-invariant family $\mathcal{F} \subsetneq \{\mathbb{F}_q \to \mathbb{F}_q\}$, and $N \subseteq \mathbb{F}_q^m$ is a Nikodym set, then $|N| \geq \dim \mathcal{L}$.*

Proof. Suppose for sake of contradiction that $|N| < \dim \mathcal{L}$. Then there exists nonzero $f \in \mathcal{L}$ such that $f|_N \equiv 0$. Let $x \in \mathbb{F}_q^m$. Then there is $y \in \mathbb{F}_q^m$ such that $x + ty \in N$ for every $t \in \mathbb{F}_q \setminus \{0\}$. Define $g(t) = f(x + ty)$. By definition of $\mathcal{L}$, we have $g \in \mathcal{F}$, and moreover $\mathcal{F}$ is a nontrivial, so by Proposition 2.9, either $g = 0$ or $\mathrm{wt}(g) \geq 2$. But $g(t) = 0$ for every $t \neq 0$, hence $g = 0$, and in particular $f(x) = g(0) = 0$. Since x was arbitrary, this shows that f is identically zero, a contradiction. □

We are now ready to prove Theorem 1.6.

Proof of Theorem 1.6. Follows immediately by applying Proposition 4.1 to the code $\mathcal{L}$ obtained from Theorem 1.3, i.e. the family of f taking values in $\mathbb{F}_p$ whose restrictions to lines are polynomials of degree at most $q - 2$. □

For comparison, the bound obtained by the polynomial method is $\binom{m+q-2}{m} \approx q^m/m!$, which can be improved to $q^m/2^m$ using the method of multiplicities. Other work on finite field Nikodym sets by Li [34] as well as Feng, Li, and Shen [16] obtain lower bounds that beat the standard polynomial method bound for $m = 2$. In particular, [16] obtains a bound of $q^2 - q^{3/2} - q$, which is actually better than our bound for two dimensions, which is $q^2 - O(q^{\log_2 3/4})$ for characteristic two. Moreover, their bound applies to q of any characteristic. However, our bounds are the best known and the only ones achieving $q^m(1 - o(1))$ for $m \geq 3$.

5 General investigation of lifting

The codes of the previous section simply picked some basic codes and lifted them to derive long codes of reasonable distance and interesting local testability and decodability. To go beyond this setting, we feel it is important to pick basic codes of possibly high distance and then lift them, and this could improve the performance of such codes. As may be observed from the previous section most of the work needed to analyze lifted codes is devoted to determining their dimension, and this can be a function of the exact code chosen. Features such as distance, decodability, and testability seem to follow more generically. In this section, we examine the simplest of these properties, namely the distance of the lifted code and prove some basic facts.

Theorem 5.1. *Let $\mathcal{F} \subseteq \{\mathbb{F}_Q^t \to \mathbb{F}_q\}$ and $\mathcal{L} = \mathrm{Lift}_m(\mathcal{F})$ for some $m \geq t$. We have the following:*

1. *$\delta(\mathcal{L}) \leq \delta(\mathcal{F})$.*
2. *$\delta(\mathcal{L}) \geq \delta(\mathcal{F}) - Q^{-t}$.*
3. *If $Q \in \{2, 3\}$ and $\delta(\mathcal{F}) > Q^{-t}$ then $\delta(\mathcal{L}) \geq \delta(\mathcal{F})$.*

5.1 Proof of Theorem 5.1

We divide the proof of Theorem 5.1 into several parts. We start by proving that distance does not increase under lifting (Theorem 5.1, Part 1).

Lemma 5.2. *Let $\mathcal{F} \subseteq \{\mathbb{F}_Q^t \to \mathbb{F}_q\}$ be a linear affine-invariant code with lift $\mathcal{L} = \mathrm{Lift}_m(\mathcal{F})$. Then $\delta(\mathcal{L}) \leq \delta(\mathcal{F})$.*

Proof. By induction, it suffices to show the assertion for the case $m = t + 1$. Let $f \in \mathcal{F}$ and let $\delta = \delta(f, 0)$. Let $\mathbf{x} = \langle x_1, \ldots, x_t \rangle$. Now consider the function $g(\mathbf{x}, y) = f(\mathbf{x})$. Clearly we have $\delta(g, 0) = \delta$. We claim that $g \in \mathcal{L}$, which completes the proof. To do so we will show that $g|_H \in \mathcal{F}$ for every t-dimensional affine subspace $H \subseteq \mathbb{F}_Q^m$. Fix such a subspace H and let $A : \mathbb{F}_Q^t \to \mathbb{F}_Q^m$ be an affine map whose image is H (such a map does exist). Note that $g|_H(\mathbf{z}) = f(A(\mathbf{z})_1, \ldots, A(\mathbf{z})_t)$. Thus if we let $A' : \mathbb{F}_Q^t \to \mathbb{F}_Q^t$ be the affine map given by the projection of A to its first t coordinates, we have that $g|_H = f \circ A'$. By Theorem A.1 $f \circ A' \in \mathcal{F}$ and so we have $g \in \mathcal{F}$ as claimed. (Note that we need to use Theorem A.1 since A' need not be an affine permutation but it is an affine transformation.) □

Next we prove Part 3 of Theorem 5.1 which asserts that the distance of non-trivial binary codes does not decrease with lifting.

Lemma 5.3. *If $\mathcal{F} \subseteq \{\mathbb{F}_2^t \to \mathbb{F}_2\}$ has distance $\delta(\mathcal{F}) > \frac{1}{2^t}$, then $\delta(\mathrm{Lift}_m(\mathcal{F})) \geq \delta(\mathcal{F})$ for all $m \geq t$.*

We prove the above lemma by stating and proving the following stronger lemma first.

Lemma 5.4. *For all $m \geq 2$, if $\delta > \frac{1}{2^{m-1}}$ and $f : \mathbb{F}_2^m \to \mathbb{F}_2$ such that $0 < \Pr_{x \in \mathbb{F}_2^m}[f(x) \neq 0] < \delta$, then there exists an $(m-1)$-dimensional affine subspace $H \subsetneq \mathbb{F}_2^m$ such that $0 < \Pr_{x \in H}[f(x) \neq 0] < \delta$.*

Proof. We proceed by induction on m. The base case $m = 2$ is straightforward to verify.

Now suppose $m > 2$ and our assertion holds for $m-1$. Let H_0, H_1 be the affine subspaces given by $x_m = 0$ and $x_m = 1$ respectively. Let δ_0, δ_1 denote $\delta(f|_{H_0}, 0), \delta(f|_{H_1}, 0)$ respectively. Note that $\delta > \delta(f, 0) = (\delta_0 + \delta_1)/2$. If both $\delta_0, \delta_1 > 0$, then by averaging we have $0 < \delta_i < \delta$ and so $H = H_i$ does the job. Otherwise, suppose w.l.o.g. that $\delta_1 = 0$. Note that $0 < \delta_0 < 2\delta$ and $2\delta > \frac{1}{2^{m-2}}$. Thus, by the induction hypothesis, there exists an $(m-2)$-dimensional affine subspace $H_0' \subset H_0$ such that $0 < \delta(f|_{H'}, 0) < 2\delta$. Let $H_1' = \{(a_1, \ldots, a_{m-1}, 1) \in \mathbb{F}_2^m \mid (a_1, \ldots, a_{m-1}, 0) \in H_0'\}$ be the translate of H_0' in H_1, and note that $\delta(f|_{H_1'}, 0) = 0$. Let $H = H_0' \cup H_1'$. Then H is an $(m-1)$-dimensional subspace of $\mathbb{F}_2^m$ such that $0 < \delta(f|_H, 0) = (\delta(f|_{H_0'}, 0) + \delta(f|_{H_1'}, 0))/2 < \delta$. □

Proof of Lemma 5.3. We prove the lemma by induction on $m - t$. Indeed the inductive step is straightforward since $\mathrm{Lift}_m(\mathcal{F}) = \mathrm{Lift}_m(\mathrm{Lift}_{m-1}(\mathcal{F}))$ and by induction both lifts on the RHS have smaller value of $m - t$ and so the distance does not reduce in either step. The main case is thus the base case with $m = t + 1$.

Suppose $f \in \mathrm{Lift}_m(\mathcal{F}) \subsetneq \{\mathbb{F}_2^m \to \mathbb{F}_2\}$ such that $0 < \delta(f, 0) < \delta(\mathcal{F})$. By Lemma 5.4, there exists an $(m-1)$-dimensional affine subspace $H \subset \mathbb{F}_2^m$ such that $0 < \delta(f|_H, 0) \leq \delta(f, 0) < \delta$, contradicting the fact that $f|_H \in \mathcal{F}$. □

A similar approach works for $q = 3$, thus we have the following.

Lemma 5.5. *If $\mathcal{F} \subseteq \{\mathbb{F}_3^t \to \mathbb{F}_3\}$ has distance $\delta(\mathcal{F}) > \frac{1}{3^t}$, then $\delta(\mathrm{Lift}_m(\mathcal{F})) \geq \delta(\mathcal{F})$ for all $m \geq t$.*

Again, we prove this by stating and proving the following analogue of Lemma 5.4.

Lemma 5.6. *For all $m \geq 2$, if $f : \mathbb{F}_3^m \to \mathbb{F}_3$ such that $\delta(f, 0) \geq \frac{1}{3^{m-1}}$, then there exists an $(m-1)$-dimensional affine subspace $H \subset \mathbb{F}_3^m$ such that $0 < \delta(f|_H, 0) \leq \delta(f, 0)$.*

Proof. Let $\delta = \delta(f, 0)$. We proceed by induction on m. For the base case $m = 2$, $\delta \geq \frac{1}{3}$. Suppose $f = f(x, y)$ and consider $f|_{y=i}$ for $i \in \mathbb{F}_3$. If $f|_{y=i}$ is not identically zero for all $i \in \mathbb{F}_3$, then by averaging there is some $i \in \mathbb{F}_3$ for which $0 < \Pr_{x \in \mathbb{F}_3}[f(x, i) \neq 0] \leq \delta$. Otherwise, w.l.o.g. suppose $f|_{y=2} \equiv 0$. Further, w.l.o.g. suppose $f|_{y=0} \not\equiv 0$ and $f(0, 0) \neq 0$. Now, if $\delta \geq \frac{2}{3}$, then the line $H = \{(x, y) \in \mathbb{F}_3^2 \mid x = 0\}$ does the job, since $0 < \Pr_{y \in \mathbb{F}_3}[f(0, y) \neq 0] \leq \frac{2}{3} \leq \delta$. If $\delta < \frac{2}{3}$, then there must exist some $a, b \in \mathbb{F}_3$ and $c \in \{0, 1\}$ such that $f(a, c) \neq 0$ and $f(b, 1 - c) = 0$. Then the line $H = \{(a, c), (b, 1 - c), (2b - a, 2)\}$ does the job, since $0 < \Pr_{(x,y) \in H}[f(x, y) \neq 0] = \frac{1}{3} \leq \delta$.

Now suppose $m > 2$ and the assertion holds for $m - 1$. For $i \in \mathbb{F}_3$, let H_i be the hyperplane cut out by $x_m = i$ and let $\delta_i = \delta(f|_{H_i}, 0)$. Then $\delta_1 + \delta_2 + \delta_3 = 3\delta$. If $\delta_i > 0$ for all $i \in \mathbb{F}_3$, then by simple averaging for some $i \in \mathbb{F}_3$ we have $0 < \delta_i \leq \delta$, so assume w.l.o.g. $\delta_2 = 0$ and $\delta_0 \geq \delta_1$. First suppose $\delta_0 \geq \frac{1}{3^{m-2}}$. Then, by the inductive hypothesis, there exists an $(m-2)$-dimensional affine subspace $H \subset H_0$ such that $0 < \delta(f|_H, 0) \leq \delta_1$. Let $H^{(0)}$ be defined by the linear equations $\sum_{i=1}^m a_i x_i - a_0 = 0$ and $x_m = 0$ for some $\langle a_0, \ldots, a_m \rangle \in \mathbb{F}_3^{m+1}$. For each $i, j \in \mathbb{F}_3$, let $H^{(i)} + j \subset H_1$ denote the affine subspace defined by $\sum_{i=1}^m a_i x_i - a_0 = j$ and $x_m = i$. By averaging, for some $i \in \mathbb{F}_3$, $\delta(f|_{H^{(1)}+i}, 0) \leq \delta_2$. Take $H = H^{(0)} \cup (H^{(1)} + i) \cup (H^{(2)} + 2i)$. Then $0 < \delta(f|_H, 0) \leq \delta$. Otherwise, suppose $\frac{1}{3^{m-2}} > \delta_0$, so $\delta_0, \delta_1 \leq \frac{2}{3^{m-1}}$. There exists $H^{(0)} \subset H_0$ be an $(m-2)$-dimensional

affine subspace such that $\delta(f|_{H^{(0)}}, 0) = \frac{1}{3^{m-1}}$. To see this, let $a, b \in H_0$ such that $f(a), f(b)$ are nonzero, and suppose a and b differ in the k-th coordinate. Then take $H^{(0)}$ defined by $x_k = a_k$ and $x_m = 0$. Again, for $i, j \in \mathbb{F}_3$ let $H^{(j)} + i$ be the $(m-2)$-dimensional affine subspace defined by $x_k = a_k + i$ and $x_m = j$. Since $\delta_2 \leq \frac{2}{3^{m-2}}$, there is $i \in \mathbb{F}_3$ such that $f|_{H^{(1)}+i} \equiv 0$. Then, taking $H = H^{(0)} \cup (H^{(1)} + i) \cup (H^{(2)} + 2i)$, we have $0 < \delta(f|_H, 0) = \frac{1}{3^{m-1}} \leq \delta$. □

Proof of Lemma 5.5. We prove the lemma by induction on $m - t$. The inductive step is straightforward since $\mathrm{Lift}_m(\mathcal{F}) = \mathrm{Lift}_m(\mathrm{Lift}_{m-1}(\mathcal{F}))$ and by induction both lifts on the RHS have smaller value of $m - t$ and so the distance does not reduce in either step. The main case is thus the base case with $m = t + 1$.

Suppose $f \in \mathrm{Lift}_m(\mathcal{F}) \subsetneq \{\mathbb{F}_3^m \to \mathbb{F}_3\}$ such that $0 < \delta(f, 0) < \delta(\mathcal{F})$. If $\delta(f, 0) \geq \frac{1}{3^{m-1}}$, then, by Lemma 5.6, there exists an $(m-1)$-dimensional affine subspace $H \subset \mathbb{F}_3^m$ such that $0 < \delta(f|_H, 0) \leq \delta(f, 0) < \delta(\mathcal{F})$, contradicting the fact that $f|_H \in \mathcal{F}$. If $\delta(f, 0) < \frac{1}{3^{m-1}}$, then there are at most two points $a, b \in \mathbb{F}_3^m$ such that $f(a), f(b)$ are nonzero. Let $i \in [m]$ such that $a_i \neq b_i$ and let H be the hyperplane defined by $x_i = a_i$. Then $f|_H$ is nonzero only on a, so $0 < \delta(f|_H) = \frac{1}{3^{m-1}} < \delta(\mathcal{F})$, again contradicting the fact that $f|_H \in \mathcal{F}$. □

For general $q > 3$, we have the following.

Lemma 5.7. *If $\mathcal{F} \subseteq \{\mathbb{F}_Q^t \to \mathbb{F}_q\}$ has distance $\delta(\mathcal{F}) = \delta$, then $\delta(\mathrm{Lift}_m(\mathcal{F})) > \delta - \frac{1-\delta}{Q^t - 1}$.*

Proof. Fix a non-zero $f \in \mathrm{Lift}_m(\mathcal{F})$ and let $\tau = \delta(f, 0)$. Fix $a \in \mathbb{F}_Q^m$ such that $f(a) \neq 0$. Now let A be a t-dimensional affine subspace containing a chosen uniformly at random from all such subspaces. Let $X(A) = |\{x \in A \mid f(x) \neq 0\}|$ be the random variable denoting the number of non-zero points of f on A. Since A samples every point of $\mathbb{F}_Q^n - \{a\}$ uniformly, we have

$$\mathbb{E}_A[X(A)] = 1 + \frac{\tau Q^m - 1}{Q^m - 1}(Q^t - 1) < 1 + \tau(Q^t - 1).$$

Therefore there must exist a t-dimensional subspace A containing a with $X(A) < \tau(Q^t - 1) + 1$. Since $f|_A$ is a non-zero function in $\mathcal{F}$, we have $\tau(Q^t - 1) + 1 \geq \delta Q^t$ and thus we conclude that $\tau \geq \delta - \frac{1-\delta}{Q^t - 1}$. In other words every non-zero function in $\mathcal{F}$ is non-zero on $\delta - \frac{1-\delta}{Q^t-1}$ fraction of the points, as asserted. □

Finally we mention examples which show that, in some senses the gaps in Theorem 5.1, Parts 2 and 3 are inherent.

First note that if $\mathcal{F} = \{F_Q^t \to \mathbb{F}_q\}$ then $\mathrm{Lift}_m(\mathcal{F}) = \{\mathbb{F}_Q^m \to \mathbb{F}_q\}$ whose distance is Q^{-m}, and so some loss in the distance is inherent in Part 2 of Theorem 5.1. However, one could hope that if $\mathcal{F} \subsetneq \{F_Q^t \to \mathbb{F}_q\}$ then its distance is preserved by lifting (as in Part 3 of Theorem 5.1). Unfortunately (actually fortunately, since this is where the rate improvement of codes in Theorem 1.2 comes from) even this hope is not true. If one takes $\mathcal{F}$ to be the binary code with degree set being all weight one integers, then its lift contains all the weight one integers as well as some integers of weight greater than one. The code consisting of only weight one integers in its degree set has distance exactly $1/2$ while codes that have rate greater than these must have distance strictly smaller than $1/2$ (by the Plotkin bound). This suggests that distances can reduce under lifts. A search reveals that the code $\mathcal{F} \subseteq \{\mathbb{F}_4 \to \mathbb{F}_2\}$ with degree set $\mathrm{Deg}(\mathcal{F}) = \{0, 1, 2\}$ has distance $1/2$ while its lift $\mathcal{L} = \mathrm{Lift}_2(\mathcal{F})$ has distance $3/8$.

Acknowledgments

We would like to thank Sergey Yekhanin for introducing us to the projective space codes which led to the parameter settings of Section 3.2. We would like to thank Elad Haramaty for clarifying discussions on the relationship between the definition of lifting in [6] and in this work.

References

[1] Noga Alon, Tali Kaufman, Michael Krivelevich, Simon Litsyn, and Dana Ron. Testing Reed-Muller codes. *IEEE Transactions on Information Theory*, 51(11):4032–4039, 2005.

[2] Sanjeev Arora and Boaz Barak. *Computational Complexity: A Modern Approach*. Cambridge, 2009.

[3] Sanjeev Arora, Carsten Lund, Rajeev Motwani, Madhu Sudan, and Mario Szegedy. Proof verification and the hardness of approximation problems. *Journal of the ACM*, 45(3):501–555, May 1998.

[4] Sanjeev Arora and Madhu Sudan. Improved low degree testing and its applications. *Combinatorica*, 23(3):365–426, 2003. Preliminary version in Proceedings of ACM STOC 1997.

[5] Boaz Barak, Parikshit Gopalan, Johan Håstad, Raghu Meka, Prasad Raghavendra, and David Steurer. Making the long code shorter, with applications to the unique games conjecture. *CoRR*, abs/1111.0405, 2011.

[6] Eli Ben-Sasson, Elena Grigorescu, Ghid Maatouk, Amir Shpilka, and Madhu Sudan. On sums of locally testable affine invariant properties. *Electronic Colloquium on Computational Complexity (ECCC)*, 18:79, 2011.

[7] Eli Ben-Sasson, Ghid Maatouk, Amir Shpilka, and Madhu Sudan. Symmetric LDPC codes are not necessarily locally testable. In *IEEE Conference on Computational Complexity*, pages 55–65. IEEE Computer Society, 2011.

[8] Eli Ben-Sasson, Noga Ron-Zewi, and Madhu Sudan. Sparse affine-invariant linear codes are locally testable. *Electronic Colloquium on Computational Complexity (ECCC)*, 19:49, 2012.

[9] Eli Ben-Sasson and Madhu Sudan. Limits on the rate of locally testable affine-invariant codes. *Electronic Colloquium on Computational Complexity (ECCC)*, 17:108, 2010.

[10] Arnab Bhattacharyya, Swastik Kopparty, Grant Schoenebeck, Madhu Sudan, and David Zuckerman. Optimal testing of Reed-Muller codes. In *FOCS*, pages 488–497. IEEE Computer Society, 2010.

[11] Pier Vittorio Ceccherini and J. W. P. Hirschfeld. The dimension of projective geometry codes. *Discrete Mathematics*, 106-107:117–126, 1992.

[12] Zeev Dvir. On the size of Kakeya sets in finite fields. *Journal of the American Mathematical Society*, (to appear), 2008. Article electronically published on June 23, 2008.

[13] Zeev Dvir, Swastik Kopparty, Shubhangi Saraf, and Madhu Sudan. Extensions to the method of multiplicities, with applications to kakeya sets and mergers. In *FOCS*, pages 181–190. IEEE Computer Society, 2009.

[14] Zeev Dvir and Amir Shpilka. An improved analysis of linear mergers. *Computational Complexity*, 16(1):34–59, 2007.

[15] Zeev Dvir and Avi Wigderson. Kakeya sets, new mergers, and old extractors. *SIAM J. Comput.*, 40(3):778–792, 2011.

[16] Chunrong Feng, Liangpan Li, and Jian Shen. Some inequalities in functional analysis, combinatorics, and probability theory. *Electr. J. Comb.*, 17(1), 2010.

[17] Katalin Friedl and Madhu Sudan. Some improvements to total degree tests. In *Proceedings of the 3rd Annual Israel Symposium on Theory of Computing and Systems*, pages 190–198, Washington, DC, USA, 4-6 January 1995. IEEE Computer Society. Corrected version available online at `http://people.csail.mit.edu/madhu/papers/friedl.ps`.

[18] Peter Gemmell, Richard Lipton, Ronitt Rubinfeld, Madhu Sudan, and Avi Wigderson. Self-testing/correcting for polynomials and for approximate functions. In *Proceedings of the Twenty Third Annual ACM Symposium on Theory of Computing*, pages 32–42, New Orleans, Louisiana, 6-8 May 1991.

[19] Peter Gemmell and Madhu Sudan. Highly resilient correctors for multivariate polynomials. *Information Processing Letters*, 43(4):169–174, September 1992.

[20] Oded Goldreich and Tali Kaufman. Proximity oblivious testing and the role of invariances. *Electronic Colloquium on Computational Complexity (ECCC)*, 17:58, 2010.

[21] Elena Grigorescu, Tali Kaufman, and Madhu Sudan. 2-transitivity is insufficient for local testability. In *IEEE Conference on Computational Complexity*, pages 259–267, 2008.

[22] Elena Grigorescu, Tali Kaufman, and Madhu Sudan. Succinct representation of codes with applications to testing. In *Proceedings of RANDOM-APPROX 2009*, volume 5687 of *Lecture Notes in Computer Science*, pages 534–547. Springer, 2009.

[23] Alan Guo and Madhu Sudan. New affine-invariant codes from lifting. *CoRR*, abs/1208.5413, 2012. Also appears as *ECCC TR 12-106*.

[24] Alan Guo and Madhu Sudan. Some closure features of locally testable affine-invariant properties. *Electronic Colloquium on Computational Complexity (ECCC)*, 19:48, 2012.

[25] Elad Haramaty, Noga Ron-Zewi, and Madhu Sudan. Absolutely sound testing of lifted codes. Manuscript, November 2012.

[26] Elad Haramaty, Amir Shpilka, and Madhu Sudan. Optimal testing of multivariate polynomials over small prime fields. In Rafail Ostrovsky, editor, *IEEE 52nd Annual Symposium on Foundations of Computer Science, FOCS 2011, Palm Springs, CA, USA, October 22-25, 2011*, pages 629–637. IEEE, 2011.

[27] Charanjit S. Jutla, Anindya C. Patthak, Atri Rudra, and David Zuckerman. Testing low-degree polynomials over prime fields. *Random Struct. Algorithms*, 35(2):163–193, 2009.

[28] Tali Kaufman and Shachar Lovett. Testing of exponentially large codes, by a new extension to weil bound for character sums. *Electronic Colloquium on Computational Complexity (ECCC)*, 17:65, 2010.

[29] Tali Kaufman and Alexander Lubotzky. Edge transitive ramanujan graphs and symmetric ldpc good codes. In Howard J. Karloff and Toniann Pitassi, editors, *STOC*, pages 359–366. ACM, 2012.

[30] Tali Kaufman and Dana Ron. Testing polynomials over general fields. *SIAM Journal of Computing*, 36(3):779–802, 2006.

[31] Tali Kaufman and Madhu Sudan. Algebraic property testing: The role of invariance. *Electronic Colloquium on Computational Complexity (ECCC)*, 14(111), 2007.

[32] Tali Kaufman and Avi Wigderson. Symmetric LDPC codes and local testing. In Andrew Chi-Chih Yao, editor, *ICS*, pages 406–421. Tsinghua University Press, 2010.

[33] Swastik Kopparty, Shubhangi Saraf, and Sergey Yekhanin. High-rate codes with sublinear-time decoding. In Lance Fortnow and Salil P. Vadhan, editors, *STOC*, pages 167–176. ACM, 2011.

[34] Liangpan Li. On the size of Nikodym sets in finite fields. *ArXiv e-prints*, March 2008.

[35] Ran Raz and Shmuel Safra. A sub-constant error-probability low-degree test, and a sub-constant error-probability PCP characterization of NP. In *Proceedings of the Twenty-Ninth Annual ACM Symposium on Theory of Computing*, pages 475–484, New York, NY, 1997. ACM Press.

[36] Ronitt Rubinfeld and Madhu Sudan. Robust characterizations of polynomials with applications to program testing. *SIAM Journal on Computing*, 25(2):252–271, April 1996.

[37] Shubhangi Saraf and Madhu Sudan. Improved lower bound on the size of Kakeya sets over finite fields. *ArXiv e-prints*, August 2008.

[38] K.J.C. Smith. On the p-rank of the incidence matrix of points and hyperplanes in a finite projective geometry. *Journal of Combinatorial Theory*, 7(2):122–129, 1969.

[39] Sergey Yekhanin. Personal communication, April 2011.

A Equivalence of invariance under affine transformations and permutations

In their work initiating the study of the testability of affine-invariant properties (codes), Kaufman and Sudan [31] studied properties closed under general affine transformations and not just permutations. While affine transformations are nicer to work with when available, they are not mathematical elegant (they don't form a group under composition). Furthermore in the case of

codes they also do not preserve the code - they only show that every codeword stays in the code after the transformation. Among other negative features affine transformations do not even preserve the weight of non-zero codewords, which can lead to some rude surprises. Here we patch the gap by showing that families closed under affine permutations are also closed under affine transformations. So one can assume the latter, without restricting the class of properties under consideration. We note that such a statement was proved in [6] for the case of univariate functions. Unfortunately their proof does not extend to the multivariate setting and forces us to rework many steps from [31].

Theorem A.1. *If $\mathcal{F} \subseteq \{\mathbb{F}_Q^m \to \mathbb{F}_q\}$ is an $\mathbb{F}_q$-linear code invariant under affine permutations, then $\mathcal{F}$ is invariant under all affine transformations.*

The central lemma (Lemma A.2) that we prove is that every non-trivial function can be split into more basic ones. This leads to a proof of Theorem A.1 fairly easily.

We first start with the notion of a basic function. For $Q = q^n$, let $\mathrm{Tr} : \mathbb{F}_Q \to \mathbb{F}_q$ denote the *trace* function $\mathrm{Tr}(x) = x + x^q + \cdots + x^{q^{n-1}}$. We say that $f : \mathbb{F}_Q^m \to \mathbb{F}_q$ is a *basic* function if $f(\mathbf{x}) = \mathrm{Tr}(\lambda \mathbf{x}^{\mathbf{d}})$ for some $\mathbf{d} \in \{0, \ldots, Q-1\}^m$. For $\mathcal{F} \subseteq \{\mathbb{F}_Q^m \to \mathbb{F}_q\}$ and $f \in \mathcal{F}$ we say f can be *split* (in $\mathcal{F}$) if there exist functions g and h such that $f = g + h$ and $\mathrm{supp}(g), \mathrm{supp}(h) \subsetneq \mathrm{supp}(f)$.

Lemma A.2. *If $\mathcal{F} \subseteq \{\mathbb{F}_Q^m \to \mathbb{F}_q\}$ is an $\mathbb{F}_q$-linear code invariant under affine permutations, then for every function $f \in \mathcal{F}$, f is either basic or f can be split.*

We first prove Theorem A.1 from Lemma A.2.

Proof of Theorem A.1. First we assert that it suffices to prove that for every function $f \in \mathcal{F}$ the function $\tilde{f} = f(x_1, \ldots, x_{m-1}, 0)$ is also in $\mathcal{F}$. To see this, consider $f \in \mathcal{F}$ and $A : \mathbb{F}_Q^m \to \mathbb{F}_Q^m$ which is not a permutation. Then there exists affine permutations $B, C : \mathbb{F}_Q^m \to \mathbb{F}_Q^m$ such that $A(\mathbf{x}) = B(C(\mathbf{x})_1, \ldots, C(\mathbf{x})_r, 0, \ldots, 0)$ where $r < m$ is the dimension of the image of A. By closure under affine permutations, it follows $f \circ C \in \mathcal{F}$. Applying the assertion above $m - r$ times we have that $f'(\mathbf{x}) = f \circ C(x_1, \ldots, x_r, 0, \ldots, 0)$ is also in $\mathcal{F}$. Finally $f \circ A = f' \circ B$ is also in $\mathcal{F}$. So we turn to proving that for every $f \in \mathcal{F}$ the function $\tilde{f} = f(x_1, \ldots, x_{m-1}, 0)$ is also in $\mathcal{F}$.

Let $f(\mathbf{x}) = \sum_{\mathbf{d}} c_{\mathbf{d}} \mathbf{x}^{\mathbf{d}}$. Notice $\tilde{f}(\mathbf{x}) = \sum_{\mathbf{d} | d_m = 0} c_{\mathbf{d}} \mathbf{x}^{\mathbf{d}}$. Writing $f = \tilde{f} + f_1$, we use Lemma A.2 to split f till we express it as a sum of basic functions $f = \sum_{i=1}^{N} b_i$, where each b_i is a basic function in $\mathcal{F}$. Note that for every b_i, we have $\mathrm{supp}(b_i) \subseteq \mathrm{supp}(\tilde{f})$ or $\mathrm{supp}(b_i) \subseteq \mathrm{supp}(f_1)$ (since the trace preserves $d_m = 0$). By reordering the b_i's assume the first M b_i's have their support in the support of $\tilde{f}$. Then we have $\tilde{f} = \sum_{i=1}^{M} b_i \in \mathcal{F}$. □

We thus turn to the proof of Lemma A.2. We prove the lemma in a sequence of cases, based on the kind of monomials that f has in its support.

We say that $\mathbf{d}$ and $\mathbf{e}$ are equivalent (modulo q), denoted $\mathbf{d} \equiv_q \mathbf{e}$ if there exists a j such that for every i, $d_i = q^j e_i \pmod^* Q$. The following proposition is immediate from previous works (see, for example, [6]). We include a proof for completeness.

Proposition A.3. *If every pair $\mathbf{d}, \mathbf{e}$ in the support of $f : \mathbb{F}_Q^m \to \mathbb{F}_q$ are equivalent, then f is a basic function.*

Proof. We first note that since the $\mathrm{Tr} : \mathbb{F}_Q \to \mathbb{F}_q$ is a (Q/q)-to-one function, we have in particular that for every $\beta \in \mathbb{F}_q$ there is an $\alpha \in \mathbb{F}_Q$ such that $\mathrm{Tr}(\alpha) = \beta$. As an immediate consequence we

have that every function $f : \mathbb{F}_Q^m \to \mathbb{F}_q$ can be expressed $\mathrm{Tr} \circ g$ where $g : \mathbb{F}_Q^m \to \mathbb{F}_Q$. Finally we note that we can view g as an element of $\mathbb{F}_Q[\mathbf{x}]$, to conclude that $f = \mathrm{Tr} \circ g$ for some polynomial g.

Now fix $f : \mathbb{F}_Q^m \to \mathbb{F}_q$ all of whose monomials are equivalent. By the above we can express $f = \mathrm{Tr} \circ g$ for some polynomial g. By inspection we can conclude that all monomials in the support of g are equivalent to the monomials in the support of f. Finally, using the fact that $\mathrm{Tr}(\alpha \mathbf{x}^{\mathbf{d}}) = \mathrm{Tr}(\alpha^q \mathbf{x}^{q\mathbf{d} \pmod^* Q})$ we can assume that g is supported on a single monomial and so $f = \mathrm{Tr}(\lambda \mathbf{x}^{\mathbf{d}}$ for some $\lambda \in \mathbb{F}_Q$. □

So it suffices to show that every function that contains non-equivalent degrees in its support can be split. We first prove that functions with "non-weakly-equivalent" monomials can be split.

We say that $\mathbf{d}$ and $\mathbf{e}$ are weakly equivalent if there exists a j such that for every i, $d_i = q^j e_i (\mod Q-1)$.

Lemma A.4. *If $\mathcal{F} \subseteq \{\mathbb{F}_Q^m \to \mathbb{F}_q\}$ is an $\mathbb{F}_q$-linear code invariant under affine permutations and $f \in \mathcal{F}$ contains a pair of non-weakly equivalent monomials in its support, then f can be split.*

Proof. Let $\mathbf{d}$ and $\mathbf{e}$ be two non weakly-equivalent monomials in the support of f. Fix j and consider the function $f_j(\mathbf{x}) = \sum_{\mathbf{a} \in (\mathbb{F}_Q^*)^m} \prod a_i^{-q^j d_i} f(a_1 x_1, \ldots, a_m x_m)$. We claim that (1) the support of f_j is a subset of the support of f, (2) $q^j \mathbf{d}$ is in the support of f_j, (3) $\mathbf{f}$ is in the support of f_j only if for every i $f_i = q^j d_i (\mod Q-1)$ and in particular (4) $\mathbf{e}$ is not in the support of f_j.

Now let $b = b(\mathbf{d})$ be the smallest positive integer such that $q^b d_i = d_i \pmod^* Q$ for every i. Now consider the function $g = \sum_{j=0}^{b-1} f_j$. We have that $g \in \mathcal{F}$ since it is an $\mathbb{F}_q$-linear combination of linear transforms of functions in $\mathcal{F}$. By the claims about the f_j's we also have that $\mathbf{d}$ is in the support of g, the support of g is contained in the support of f and $\mathbf{e}$ is not in the support of f. Expressing $f = g + (f - g)$ we now have that f can be split. □

The remaining cases are those where some of coordinates of $\mathbf{d}$ are zero or $Q-1$ for every $\mathbf{d}$ in the support of f. We deal with a special case of such functions next.

Lemma A.5. *Let $\mathcal{F}$ be a linear affine-invariant code. Let $f \in \mathcal{F}$ be given by $f(\mathbf{x}, \mathbf{y}) = \mathrm{Tr}(\mathbf{y}^{\mathbf{d}} p(\mathbf{x}))$ where every variable in $p(\mathbf{x})$ has degree in $\{0, Q-1\}$ in every monomial, and $\mathbf{d}$ is arbitrary. Further, let degree of $p(\mathbf{x})$ be $a(Q-1)$. Then for every $0 \le b \le a$ and for every $\lambda \in \mathbb{F}_Q$, the function $(x_1 \cdots x_b)^{Q-1} \mathrm{Tr}(\lambda \mathbf{y}^{\mathbf{d}}) \in \mathcal{F}$.*

Note that in particular the lemma above implies that such f's can be split into basic functions.

Proof. We prove the lemma by a triple induction, first on a, then on b, and then on the number of monomials in p. The base case is $a = 0$ and that is trivial. So we consider general $a > 0$.

First we consider the case $b < a$. Assume w.l.o.g. that the monomial $(x_1 \cdots x_a)^{Q-1}$ is in the support of p and write $p = p_0 + x_1^{Q-1} p_1$ where p_0, p_1 do not depend on x_1. Note that $p_1 \neq 0$ and $\deg(p_1) = (a-1)(Q-1)$. We will prove that $-\mathrm{Tr}(\mathbf{y}^{\mathbf{d}} p_1(\mathbf{x})) \in \mathcal{F}$ and this will enable us to apply the inductive hypothesis to p_1. Let $g(\mathbf{x}, \mathbf{y}) = \sum_{\beta \in \mathbb{F}_Q} f(x_1 + \beta, x_2, \ldots, x_m, \mathbf{y})$. By construction $g \in \mathcal{F}$. By linearity of the Trace we have

$$g = \mathrm{Tr}\left(\mathbf{y}^{\mathbf{d}} \left(\sum_{\beta \in \mathbb{F}_Q} p_0 + (x_1 + \beta)^{Q-1} p_1\right)\right) = \mathrm{Tr}(\mathbf{y}^{\mathbf{d}}(-p_1(\mathbf{x}))),$$

where the second equality follows from the fact that $\sum_{\beta\in\mathbb{F}_Q}(z+\beta)^{Q-1} = -1$. Thus we can now use induction to claim $(x_1 \dots x_b)^{Q-1}\operatorname{Tr}(\lambda \mathbf{y}^{\mathbf{d}}) \in \mathcal{F}$.

Finally we consider the case $b = a$. Now note that since the case $b < a$ is known, we can assume w.l.o.g that p is homogenous (else we can subtract off the lower degree terms). Now if $a = m$ there is nothing to be proved since p is just a single monomial. So assume $a < m$. Also if p has only one monomial then there is nothing to be proved, so assume p has at least two monomials. In particular assume p is supported on some monomial that depends on x_1 and some monomial that does not depend on x_1. Furthermore, assume w.l.o.g. that a monomial depending on x_1 does not depend on x_2. Write $p = x_1^{Q-1}p_1 + x_2^{Q-1}p_2 + (x_1x_2)^{Q-1}p_3 + p_4$ where the p_i's don't depend on x_1 or x_2. By assumption on the monomials of p we have that $p_1 \neq 0$ and at least one of $p_2, p_3, p_4 \neq 0$. Now consider the affine transform A that sends x_1 to $x_1 + x_2$ and preserves all other x_i's. We have $g = f \circ A = \operatorname{Tr}\left(\mathbf{y}^{\mathbf{d}}(x_1^{Q-1}p_1 + x_2^{Q-1}(p_1+p_2) + (x_1x_2)^{Q-1}p_3 + p_4 + r)\right)$ where the x_1-degree of every monomial in r is in $\{1, \dots, Q-2\}$. Now consider $g'(\mathbf{x}, \mathbf{y}) = \sum_{\alpha\in\mathbb{F}_Q^*} g(\alpha x_1, x_2, \dots, x_m, \mathbf{y})$. The terms of r vanish in g' leaving $g' = -(f \circ A - r) = \operatorname{Tr}\left(\mathbf{y}^{\mathbf{d}}\left(-x_1^{Q-1}p_1 - x_2^{Q-1}(p_1+p_2) - (x_1x_2)^{Q-1}p_3 - p_4\right)\right)$. Finally we consider the function $\tilde{g} = f + g' = \operatorname{Tr}(\mathbf{y}^{\mathbf{d}}(-x_2^{Q-1}p_1))$ which is a function in $\mathcal{F}$ of degree $a(Q-1)$ supported on a smaller number of monomials than f, so by applying the inductive hypothesis to $\tilde{g}$ we have that $\mathcal{F}$ contains the monomial $(x_1 \cdots x_a)^{Q-1}$. □

The following lemma converts the above into the final piece needed to prove Lemma A.2.

Lemma A.6. *If $\mathcal{F} \subseteq \{\mathbb{F}_Q^m \to \mathbb{F}_q\}$ is an $\mathbb{F}_q$-linear code invariant under affine permutations and all monomials in $f \in \mathcal{F}$ are weakly equivalent, then f can be split.*

Proof. First we describe the structure of a function $f : \mathbb{F}_Q^m \to \mathbb{F}_q$ that consists only of weakly equivalent monomials. First we note that the m variables can be separated into those in which every monomial has degree in $\{1, \dots, Q-2\}$ and those in which every monomial has degree in $\{0, Q-1\}$ (since every monomial is weakly equivalent). Let us denote by $\mathbf{x}$ the variables in which the monomials of f have degree in $\{0, Q-1\}$ and $\mathbf{y}$ be the remaining monomials. Now consider some monomial of the form $M = c\mathbf{x}^{\mathbf{e}}\mathbf{y}^{\mathbf{d}}$ in f. Since f maps to $\mathbb{F}_q$ we must have that the coefficient of $(\mathbf{x}^{\mathbf{e}}\mathbf{y}^{\mathbf{d}})^{q^j}$ is c^{q^j}. Furthermore, we have every other monomial M' in the support of f is of the form $c'\mathbf{y}^{q^j\mathbf{d}}\mathbf{x}^{\mathbf{e}'}$. Thus f can be written as $\operatorname{Tr}(\mathbf{y}^{\mathbf{d}}p(\mathbf{x}))$ where $p(x_1, \dots, x_m) = \tilde{p}(x_1^{Q-1}, \dots, x_m^{Q-1})$. But, by Lemma A.5, such an f can be split. □

Proof of Lemma A.2. If f contains a pair of non-weakly equivalent monomials then f can be split by Lemma A.4. If not, then f is either basic or, by Lemma A.6 is can be split. □

We also prove an easy consequence of Lemma A.2.

Lemma A.7. *Let $\mathcal{F} \subseteq \{\mathbb{F}_Q^m \to \mathbb{F}_q\}$ be affine invariant. If $\mathbf{d} \in \operatorname{Deg}(\mathcal{F})$, then $\operatorname{Tr}(\lambda\mathbf{x}^{\mathbf{d}}) \in \mathcal{F}$ for all $\lambda \in \mathbb{F}_Q$.*

Proof. We first claim that Lemma A.2 implies that there exists $\beta \in \mathbb{F}_Q$ such that $\operatorname{Tr}(\beta\mathbf{x}^{\mathbf{d}})$ is a non-zero function in $\mathcal{F}$. To verify this, consider a "minimal" function (supported on fewest monomials) $f \in \mathcal{F}$ with $\mathbf{d} \in \operatorname{supp}(f)$. Since f can't be split in $\mathcal{F}$ (by minimality), by Lemma A.2 f must be basic and so equals (by definition of being basic) $\operatorname{Tr}(\beta\mathbf{x}^{\mathbf{d}})$.

Now let $b = b(\mathbf{d})$ be the smallest positive integer such that $q^b\mathbf{d} \pmod^* Q) = \mathbf{d}$. If $Q = q^n$, note that b divides n and so one can write $\operatorname{Tr} : \mathbb{F}_Q \to \mathbb{F}_q$ as $\operatorname{Tr}_1 \circ \operatorname{Tr}_2$ where $\operatorname{Tr}_1 : \mathbb{F}_{q^b} \to \mathbb{F}_q$ is the function

$\mathrm{Tr}_1(z) = z + z^q + \cdots + z^{q^{b-1}}$ and $\mathrm{Tr}_2 : \mathbb{F}_Q \to \mathbb{F}_{q^b}$ is the function $\mathrm{Tr}_2(z) = z + z^{q^b} + \cdots + z^{Q/q^b}$. (Both Tr_1 and Tr_2 are trace functions mapping the domain to the range.) It follows that $\mathrm{Tr}(\beta \mathbf{x}^{\mathbf{d}}) = \mathrm{Tr}_1(\mathrm{Tr}_2(\beta)\mathbf{x}^{\mathbf{d}})$.

We first claim that $\mathrm{Tr}_1(\tau \mathbf{x}^{\mathbf{d}}) \in \mathcal{F}$ for every $\tau \in \mathbb{F}_{q^b}$. Let $S = \{\sum_{\alpha \in (\mathbb{F}_Q^*)^m} a_\alpha \cdot \alpha^{\mathbf{d}} \mid a_\alpha \in \mathbb{F}_q\}$. We note that by linearity and affine-invariance of $\mathcal{F}$, we have that $\mathrm{Tr}_1(\mathrm{Tr}_2(\beta) \cdot \eta \mathbf{x}^{\mathbf{d}}) \in \mathcal{F}$ for every $\eta \in S$. By definition S is closed under addition and multiplication and so is a subfield of $\mathbb{F}_Q$. In fact, since every $\eta \in S$ satisfies $\eta^{q^b} = \eta$ (which follows from the fact that $\alpha^{\mathbf{d}} = \alpha^{q^b \mathbf{d}}$), we have that $S \subseteq \mathbb{F}_{q^b}$. It remains to show $S = \mathbb{F}_{q^b}$. Suppose it is a strict subfield of size q^c for $c < b$. Consider γ^{d_i} for $\gamma \in \mathbb{F}_Q$ and $i \in [m]$. Since $\gamma^{d_i} \in S$, we have that $\gamma^{d_i q^c} = \gamma^{d_i}$ for every $\gamma \in \mathbb{F}_Q$ and so we get $x_i^{q^c d_i} = x_i \pmod{x_i^Q - x_i}$. We conclude that $\mathbf{x}^{q^c \mathbf{d}} = \mathbf{x}^{\mathbf{d}} \pmod{\mathbf{x}^Q - \mathbf{x}}$ which contradicts the minimality of $b = b(\mathbf{d})$. We conclude that $S = \mathbb{F}_{q^b}$. Since $\mathrm{Tr}_2(\beta) \in \mathbb{F}_{q^b}^*$, we conclude that the set of coefficients τ such that $\mathrm{Tr}_1(\tau \mathbf{x}^{\mathbf{d}}) \in \mathcal{F}$ is all of $\mathbb{F}_{q^b}$ as desired.

Finally consider any $\lambda \in \mathbb{F}_Q$. since $\mathrm{Tr}_2(\lambda) \in \mathbb{F}_{q^b}$, we have that $\mathrm{Tr}_1(\mathrm{Tr}_2(\lambda)\mathbf{x}^{\mathbf{d}}) \in \mathcal{F}$ (from the previous paragraph), and so $\mathrm{Tr}(\lambda \mathbf{x}^{\mathbf{d}}) = \mathrm{Tr}_1(\mathrm{Tr}_2(\lambda)\mathbf{x}^{\mathbf{d}}) \in \mathcal{F}$ □

B Coefficients of multinomial expansions modulo a prime

For an integer d, let $d^{(i)}$ denote the ith digit in the p-ary expansion of d, so that $d = \sum_{i=1}^{\infty} d^{(i)} p^i$. Let $\equiv_p$ denote equivalence modulo p. The following is a well known theorem of Lucas.

Theorem B.1 (Lucas' theorem)**.** *If $d, e \in \mathbb{Z}$, then $\binom{d}{e} \equiv_p \prod_i \binom{d_i}{e_i}$.*

In particular, $\binom{d}{e} \not\equiv 0 \pmod{p}$ if and only if $e \leq_p d$, so we have $(x+y)^d \equiv_p \sum_{e \leq_p d} x^e y^{d-e}$. More generally, we would like to know when $\binom{\mathbf{d}}{\mathbf{E}}$ vanishes modulo p. To this end, we use the following claim.

Lemma B.2. *If $d \in \mathbb{Z}$ and $\mathbf{e} \in \mathbb{Z}^t$, then $\binom{d}{\mathbf{e}} \not\equiv_p 0$ only if $\mathbf{e} \leq_p d$. More generally, if $\mathbf{d} \in \mathbb{Z}^m$ and $\mathbf{E} \in \mathbb{Z}^{m \times t}$, then $\binom{\mathbf{d}}{\mathbf{E}} \not\equiv_p 0$ only if $\mathbf{E} \leq_p \mathbf{d}$.*

Proof. We have $\binom{d}{\mathbf{e}} = \prod_{i=1}^{t-1} \binom{d - \sum_{j=1}^{i-1} e_j}{e_i}$. For this to be nonzero modulo p, by Lucas' theorem we have $e_i \leq_p d - \sum_{j=1}^{i-1} e_j$, from which it follows that $\mathbf{e} \leq_p d$. The more general statement then follows immediately from definition. □

Lemma B.3. *Let $\mathbf{A} \in \mathbb{Z}^{m \times t}$ and let $\mathbf{d}, \mathbf{b} \in \mathbb{Z}^m$ and $\mathbf{x} \in \mathbb{Z}^t$. Then*

$$(\mathbf{A}\mathbf{x} + \mathbf{b})^{\mathbf{d}} = \sum_{\mathbf{E}} \binom{\mathbf{d}}{\mathbf{E}} \mathbf{A}^{\mathbf{E}} \mathbf{x}^{\Sigma(\mathbf{E})} \mathbf{b}^{\mathbf{d} - \Sigma(\mathbf{E}^\top)}.$$

Proof. For matrices $\mathbf{A}, \mathbf{E}$, let a_{ij}, e_{ij} denote their entries respectively. The lemma follows by

straightforward calculation. We have

$$\begin{aligned}
(\mathbf{Ax}+\mathbf{b})^{\mathbf{d}} &= \prod_{i=1}^{m}\left(\sum_{j=1}^{t} a_{ij}x_j + b_i\right)^{d_i} \\
&= \prod_{i=1}^{m}\left(\sum_{e_{i1},\ldots,e_{it}} \binom{d_i}{\langle e_{i1},\ldots,e_{it}\rangle}\left(\prod_{j=1}^{t} a_{ij}^{e_{ij}}x_j^{e_{ij}}\right) b_i^{d_i-\sum_{j=1}^{t} e_{ij}}\right) \\
&= \sum_{\mathbf{E}}\prod_{i=1}^{m}\left(\binom{d_i}{\langle e_{i1},\ldots,e_{it}\rangle}\left(\prod_{j=1}^{t} a_{ij}^{e_{ij}}x_j^{e_{ij}}\right) b_i^{d_i-\sum_{j=1}^{t} e_{ij}}\right) \\
&= \sum_{\mathbf{E}}\binom{\mathbf{d}}{\mathbf{E}}\left(\prod_{i=1}^{m}\prod_{j=1}^{t} a_{ij}^{e_{ij}}\right)\left(\prod_{j=1}^{t} x_j^{\sum_{i=1}^{m} e_{ij}}\right)\left(\prod_{i=1}^{m} b_i^{d_i-\sum_{j=1}^{t} e_{ij}}\right) \\
&= \sum_{\mathbf{E}}\binom{\mathbf{d}}{\mathbf{E}}\mathbf{A}^{\mathbf{E}}\mathbf{x}^{\Sigma(\mathbf{E})}\mathbf{b}^{\mathbf{d}-\Sigma(\mathbf{E}^{\top})}.
\end{aligned}$$

□