Locality via Partially Lifted Codes*

S. Luna Frank-Fischer[1], Venkatesan Guruswami[†2], and Mary Wootters[‡3]

1 **Computer Science Department, Stanford University, Stanford, CA, USA**
`luna16@stanford.edu`
2 **Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, USA**
`venkatg@cs.cmu.edu`
3 **Computer Science Department, Stanford University, Stanford, CA, USA**
`marykw@stanford.edu`

Abstract

In error-correcting codes, *locality* refers to several different ways of quantifying how easily a small amount of information can be recovered from encoded data. In this work, we study a notion of locality called the s-Disjoint-Repair-Group Property (s-DRGP). This notion can interpolate between two very different settings in coding theory: that of Locally Correctable Codes (LCCs) when s is large – a very strong guarantee – and Locally Recoverable Codes (LRCs) when s is small – a relatively weaker guarantee. This motivates the study of the s-DRGP for intermediate s, which is the focus of our paper. We construct codes in this parameter regime which have a higher rate than previously known codes. Our construction is based on a novel variant of the *lifted codes* of Guo, Kopparty and Sudan. Beyond the results on the s-DRGP, we hope that our construction is of independent interest, and will find uses elsewhere.

1998 ACM Subject Classification E.4 Error Control Codes

Keywords and phrases Error correcting codes, locality, lifted codes

Digital Object Identifier 10.4230/LIPIcs.APPROX/RANDOM.2017.43

1 Introduction

In the theory of error correcting codes, *locality* refers to several different ways of quantifying how easily a small amount of information can be recovered from encoded data. Slightly more formally, suppose that $\mathcal{C} \subset \Sigma^N$ is a *code* over an alphabet Σ; that is, $\mathcal{C}$ is any subset of Σ^N. Suppose that $c \in \mathcal{C}$, and that we have query access to a noisy version $\tilde{c}$ of c. We are tasked with finding $c_i \in \Sigma$ for some $i \in [N]$. Informally, we say that the code $\mathcal{C}$ exhibits good *locality* if we may recover c_i using very few queries to $\tilde{c}$. Of course, the formal definition of locality in this set-up depends on the nature of the noise, and the question is interesting for a wide variety of noise models.

One (extremely strong) model of noise is that handled by *Locally Correctable Codes* (LCCs), which have been extensively studied in theoretical computer science for over 15 years. This model is motivated by a variety of applications in theoretical computer science and cryptography, including probabilistically checkable proofs (PCPs), derandomization, and private information retrieval (PIR); we refer the reader to [30] for an excellent survey

* Full version available at `https://arxiv.org/abs/1704.08627`.
† Research supported in part by NSF grants CCF-1563742 and CCF-1422045.
‡ Research supported in part by NSF grants DMS-1400558 and CCF-1657049.

on LCCs. In the LCC setting, $\tilde{c} \in \Sigma^N$ has a constant fraction of errors: that is, we are guaranteed that the Hamming distance between $\tilde{c}$ and c is no more than δN, for some small constant $\delta > 0$. The goal is to recover c_i with high probability from $Q = o(N)$ randomized queries to $\tilde{c}$.

Another (much weaker) model of noise is that handled by *Locally Recoverable Codes* (LRCs) and related notions, which have been increasingly studied recently motivated by applications in distributed storage [14, 10, 23]. In this model, $\tilde{c} \in (\Sigma \cup \{\perp\})^N$ has a constant *number* of *erasures*: that is, we are guaranteed that the number of $\perp$ symbols in $\tilde{c}$ is at most some constant $e = O(1)$, and further that $c_i = \tilde{c}_i$ whenever $\tilde{c}_i \neq \perp$. As before, the goal is to recover c_i using as few queries as possible to $\tilde{c}$. Batch codes [15, 7] and PIR codes [8, 6] are other variants that are interesting in this parameter regime.

A key question in both of these lines of work is how to achieve these recovery guarantees with as high a *rate* as possible. The rate of a code $\mathcal{C} \in \Sigma^N$ is defined to be the ratio $\log_{|\Sigma|}(|\mathcal{C}|)/N$; it captures how much information can be transmitted using such a code. In other words, given N, we seek to find a $\mathcal{C} \subseteq \Sigma^N$ with good locality properties, so that $|\mathcal{C}|$ is as large as possible.

In the context of the second line of work above, recent work [28, 21, 25, 27, 1, 8] has studied (both implicitly and explicitly) the trade-off between rate and something called the s-Disjoint-Repair-Group-Property (s-DRGP) for small s. Informally, $\mathcal{C}$ has the s-DRGP if any symbol c_i can be obtained from s disjoint query sets $c|_{S_1}, c|_{S_2}, \ldots, c|_{S_s}$ for $S_i \subseteq [N]$. (Notice that there is no explicit bound on the size of these query sets, just that they must be disjoint).

One observation which we will make below is that the s-DRGP provides a natural way to interpolate between the first (LCC) setting and the second (LRC) setting above. More precisely, while the LRC setting corresponds to small s (usually, $s = O(1)$), the LCC setting is in fact equivalent to the case when $s = \Omega(N)$. This observation motivates the study of *intermediate* s, which is the goal in this paper.

Contributions

Before we give a more detailed overview of previous work, we outline the main contributions of this paper.

1. **Constructions of codes with the s-DRGP for intermediate s.** We give a construction of a family of codes which have the s-DRGP for $s \sim N^{1/4}$. Our construction can achieve a higher rate than previous constructions with the same property.
2. **A general framework, based on partially lifted codes.** Our codes are based on a novel variant of the *lifted codes* of Guo, Kopparty and Sudan [12]. In that work, with the goal of obtaining LCCs, the authors showed how to construct affine-invariant codes by a "lifting" operation. In a bit more detail, their codes are multivariate polynomial codes, whose entries are indexed by $\mathbb{F}_q^m$ (so $N = q^m$). These codes have the property that, the restriction of each codeword to every line in $\mathbb{F}_q^m$ is a codeword of a suitable univariate polynomial code. (For example, a Reed-Muller code is a *subset* of a lift of a Reed-Solomon code; the beautiful insight of [12] is that in fact the lifted code may be much larger.)
   In our work, we introduce a version of the lifting operation where we only require that the restriction to *some* lines lie in the smaller code, rather than the restriction to *all* lines; we call such codes "partially lifted codes." This partial lifting operation potentially allows for higher-rate codes, and, as we will see, it naturally gives rise to codes with the s-DRGP. One of our main contributions is the introduction of these codes, as well as some machinery which allows us to control their rate. We instantiate this machinery with a particular

example, in order to obtain the construction advertised above. We can also recover previous results in the context of this machinery.

3. **Putting the study of the s-DRGP in the context of LRCs and LCCs.** While the s-DRGP has been studied before, to the best of our knowledge, it is not widely viewed as a way to interpolate between the two settings described above. One of the goals of this paper is to highlight this property and its potential importance to our understanding of locality, both from the LRC/batch code/PIR code side of things, and from the LCC side.

1.1 Background and related work

As mentioned above, in this work we study the s-Disjoint-Repair-Group Property (s-DRGP). We begin our discussion of the s-DRGP with some motivation from the LRC end of the spectrum, from applications in distributed storage. The following model is common in distributed storage: imagine that each server or node in a distributed storage system is holding a single symbol of a codeword $c \in \mathcal{C}$. Over time, nodes fail, usually one at a time, and we wish to repair them (formally, recovering c_i for some i). Moreover, when they fail, it is clear that they have failed. This naturally gives rise to the second parameter regime described above, where $\tilde{c}$ has a constant number of erasures.

Locally recoverable (or repairable) codes (LRCs) [14, 10, 23] were introduced to deal with this setting. The guarantee of an LRC[1] with locality Q is that for any $i \in \{1, \ldots, n\}$, the i'th symbol of the codeword can be determined from a set of at most Q other symbols. There has been a great deal of work recently aimed at pinning down the trade-offs between rate, distance, and the locality parameter Q in LRCs. At this point, we have constructions which have optimal trade-offs between these parameters, as well as reasonably small alphabet sizes [26]. However, there are still many open questions; a major question is how to handle a small number of erasures, rather than a single erasure. This may result from either multiple node failures, or from "hot" data being overloaded with requests. There are several approaches in the literature, but the approach relevant to this work is the study of *multiple disjoint repair groups*.

▶ **Definition 1.** Given a code $\mathcal{C} \subset \Sigma^N$, we say that a set $S \subset \{1, \ldots, N\}$ is a *repair group* for $i \in \{1, \ldots, N\}$ if $i \notin S$, and if there is some function $g : \Sigma^{|S|} \to \Sigma$ so that $g(c|_S) = c_i$ for all $c \in \mathcal{C}$. That is, the codeword symbols indexed by S uniquely determine the symbol indexed by i.

▶ **Definition 2.** We say that $\mathcal{C}$ has the s-Disjoint-Repair-Group Property (s-DRGP) if for every $i \in \{1, \ldots, N\}$, there are s disjoint repair groups $S_1^{(i)}, \ldots, S_s^{(i)}$ for i.

In the context of LRCs, the parameter s is called the *availability* of the code. An LRC with availability s is not exactly the same as a code with the s-DRGP (the difference is that, in Definition 2, there is no mention of the size Q of the repair groups), but it turns out to be deeply related; it is also directly related to other notions of locality in distributed storage (like batch codes), as well as in cryptography (like PIR codes). We will review some of this work below, and we point the reader to [24] for a survey of batch codes, PIR codes, and their connections to LRCs and the s-DRGP.

While originally motivated for small s, as we will see below, the s-DRGP is interesting (and has already been implicitly studied) for a wide range of s, from $O(1)$ to $\Omega(N)$. For

[1] In some works, the guarantee holds for information symbols only, rather than for all codeword symbols; we stick with all symbols here for simplicity of exposition.

$s = o(N)$, we can hope for codes with very high rate, approaching 1; the question is how fast we can hope for this rate to approach 1. More formally, if $K = \log_{|\Sigma|} |\mathcal{C}|$, then the rate is K/N, and we are interested in how the gap $N - K$ behaves with N and s. We will refer to the quantity $N - K$ as the *co-dimension* of the code; when $\mathcal{C}$ is linear (that is, when $\Sigma = \mathbb{F}$ is a finite field and $\mathcal{C} \subseteq \mathbb{F}^N$ is a linear subspace), then this is indeed the co-dimension of $\mathcal{C}$ in $\mathbb{F}^N$. The main question we seek to address in this paper is the following.

▶ **Question 1.** *For a given s and N, what is the smallest codimension $N - K$ of any code with the s-DGRP? In particular, how does this quantity depend on s and N?*

We know a few things about Question 1, which we survey below. However, there are many things about this question which we still do not understand. In particular, the dependence on s is wide open, and this dependence on s is the focus of the current work. Below, we survey the state of Question 1 both from the LRC end (when s is small) and the LCC end (when s is large).

The s-DRGP when s is small

In [28, 21, 25, 27], the s-DRGP was explicitly considered, with a focus on small s ($s = 2$ is of particular interest). In those works, some bounds on the rate and distance of codes with the s-DRGP were derived (some of them in terms of the locality Q). However, for larger s, these bounds degrade. More precisely, [28, 21] establish bounds on $N - K$ in terms of Q, s, and the distance of the code, but as s grows these are not much stronger than the Singleton bound. The results of [25, 27] give an upper bound on the rate of a code in terms of Q and s. One corollary is that the rate satisfies $K/N \leq (s+1)^{-1/Q}$; if we are after high-rate codes, this implies that we must take $Q = \Omega(\ln(s+1))$, and this implies that the codimension $N - K$ must be at least $\Omega(N \ln(s)/Q)$.

A similar notion to the s-DRGP was introduced in [8], with the application of *Private Information Retrieval* (PIR). PIR schemes are an important primitive in cryptography, and they have long been linked to constant-query LCCs. In [8], PIR was also shown to be related to the s-DRGP. The work [8] introduces *PIR codes*, which enable PIR schemes with much less storage overhead. It turns out that the requirement for PIR codes is very similar to the s-DRGP.[2]

In the context of PIR codes [8, 6], there are constructions of s-DRGP codes with $N - K \leq O(s\sqrt{N})$. For $s = 2$, this is known to be tight, and there is a matching lower bound [20]. However, it seems difficult to use this lower bound technique to prove a stronger lower bound when s is larger (possibly growing with N).

The s-DRGP when s is large

As we saw above, when s is small then the s-DRGP is intimately related to LRCs, PIR codes and batch codes. On the other end of the spectrum, when s is large (say, $\Omega(N)$ or $\Omega(N^{1-\varepsilon})$) then it is related to LCCs.

When $s = \Omega(N)$, then the s-DRGP is in fact *equivalent* to a constant-query LCC (that is, an LCC as described above, where the number of queries to $\tilde{c}$ is $O(1)$). The fact that the $\Omega(N)$-DRGP implies a constant-query LCC is straightforward: the correction algorithm to recover c_i is to choose a random j in $\{1, \ldots, s\}$ and use the repair group $S_j^{(i)}$ to recover c_i.

[2] The only difference is that PIR codes only need to recover information symbols, but possibly with non-systematic encoding.

Since in expectation the size of $S_j^{(i)}$ is constant, we can restrict our attention only to the constant-sized repair groups. Then, with some constant probability none of the indices in $S_j^{(i)}$ will be corrupted, and this success probability can be amplified by independent repetitions. The converse is also true [16, 29], and any constant-query LCC has the s-DRGP for $s = \Omega(n)$; in fact, this connection is one of the few ways we know how to get lower bounds on LCCs.

When s is large, but not as large as $\Omega(N)$, there is still a tight relationship with LCCs. By now we know of several high-rate ($(1-\alpha)$, for any constant α) LCCs with query complexity $Q = N^{\varepsilon}$ for any $\varepsilon > 0$ [17, 12, 13] or even $Q = N^{o(1)}$ [18]. It is easy to see[3] that any LCC with query complexity Q has the s-DGRP for $s = \Omega(N/Q)$. Thus, these codes immediately imply high-rate s-DRGP codes with $s = \Omega(N^{1-\varepsilon})$ or even larger. (See also [1]). Conversely, the techniques of [13, 18] show how to take high-rate linear codes with the s-DGRP for $s = \Omega(N^{1-\varepsilon})$ and produce high-rate LCCs with query complexity $O(N^{\varepsilon'})$ (for a different constant ε').

These relationships provide some bounds on the codimension $N - K$ in terms of s: from existing lower bounds on constant-query LCCs [29], we know that any code with the s-DGRP and $s = \Omega(N)$ must have vanishing rate. On the other hand from high-rate LCCs, there exist s-DGRP codes with $s = \Omega(N^{1-\varepsilon})$ and with high rate. However, these techniques do not immediately given anything better than high (constant) rate, while in Question 1 we are interested in precisely controlling the co-dimension $N - K$.

The s-DRGP when s is intermediate

The fact that the s-DRGP interpolates between the LRC setting for small s and the LCC setting for large s motivates the question of the s-DGRP for intemediate s, say $s = \log(N)$ or $s = N^c$ for $c < 1/2$. Our goal is to understand the answer to Question 1 for intermediate s.

We have only a few data points to answer this question. As mentioned above, the constructions of [8, 6] show that there are codes with $N - K \leq s\sqrt{N}$ for $s \leq \sqrt{N}$. However, the best general lower bounds known [20, 27] can only establish $N - K \geq \max\left\{\sqrt{2N}, N - \frac{N}{(s+1)^{1/Q}}\right\}$. Above, we recall that Q is a parameter bounding the size of the repair groups; in order for the second term above (from [27]) to be $o(N)$, we require $Q \gg \ln(s+1)$; in this case, the second bound on the codimension reads $N - K \geq \Omega(N \ln(s)/Q)$. As the size of the repair groups Q may in general be as large as N/s, in our setting this second bound gives better dependence on s, but worse dependence on N.

The upper bound of $s\sqrt{N}$ is not tight, at least for large s. For $s = \sqrt{N}$, there are several classical constructions which have the s-DRGP and with $N - K = \Theta(N^{\log_4(3)})$; for example, this includes affine geometry codes and/or codes constructed from difference sets (see [2], [19], or [12] – we will also recover these in Corollary 15). Notice that this is much better than the upper bound of $N - K \leq s\sqrt{N}$, which for $s = \sqrt{N}$ would be trivial.

However, other than these codes, before this work we did not know of any constructions for

[3] Indeed, suppose that $\mathcal{C}$ is an LCC with query complexity Q and error tolerance δ, and let $s = \delta N/Q$. In order to obtain s disjoint repair groups for a symbol c_i from the LCC guarantee, we proceed as follows. First, we make one (randomized) set of queries to c; this gives the first repair group. Continuing inductively, assume we have found $t \leq s$ disjoint repair groups already, covering a total of at most $tQ < \delta N$ symbols. To get the $t+1$'st set of queries, we again choose at random as per the LCC requirement. These queries may not be disjoint from the previous queries, but the LCC guarantee can handle errors (and hence erasures) in up to δN positions, so it suffices to query the points which have not been already queried, and treat the already-queried points as unavailable. We repeat this process until t reaches $s = \delta N/Q$.

$s \ll \sqrt{N}$ which beat the bounds in [8, 6] of $N - K \leq s\sqrt{N}$.[4] One of the main contributions of this work is to give a construction with $s = N^{1/4}$, which achieves codimension $N - K = N^{0.714}$. Notice that the bound of $s\sqrt{N}$ would be $N^{0.75}$ in this case, so this is a substantial improvement. We remark that we do not believe that our construction is optimal, and unfortunately we don't have any deep insight about the constant 0.714. Rather, we stress that the point of this work is to (a) highlight the fact that the $s\sqrt{N}$ bound can be beaten for $s \ll \sqrt{N}$, and (b) highlight our techniques, which we believe may be of independent interest.

1.2 Lifted codes, and our construction

Our construction is based on the *lifted codes* of Guo, Kopparty and Sudan [12]. The original motivation for lifted codes was to construct high-rate LCCs, as described above. However, since then they have found several other uses, for example list-decoding and local-list-decoding [11]. The codes are based on multivariate polynomials, and we describe them below. Suppose that $\mathcal{F} \subseteq \mathbb{F}_q[X, Y]$ is a collection of bivariate polynomials over a finite field $\mathbb{F}_q$ of order q. This collection naturally gives rise to a code $\mathcal{C} \subseteq \mathbb{F}^{q^2}$:

$$\mathcal{C} = \left\{ \langle P(x,y) \rangle_{(x,y) \in \mathbb{F}_q^2} \,:\, P \in \mathcal{F} \right\}. \tag{1}$$

Above, we assume some fixed order on the elements of $\mathbb{F}_q^2$, and by $\langle P(x,y) \rangle_{(x,y) \in \mathbb{F}_q^2}$, we mean the vector in $\mathbb{F}_q^{q^2}$ whose entries are the evaluations of P in this prescribed order. For example, a bivariate Reed-Muller code is formed by taking $\mathcal{F}$ to be the set of all polynomials of total degree at most d.

One nice property of Reed-Muller codes is their locality. More precisely, suppose that $P(X, Y)$ is a bivariate polynomial over $\mathbb{F}_q$ of total degree at most d. For an affine line in $\mathbb{F}_q^2$, parameterized as $L(T) = (\alpha T + \beta, \gamma T + \delta)$, we can consider the *restriction* $P|_L$ of P to L, given by

$$P|_L(T) := P(\alpha T + \beta, \gamma T + \delta) \mod T^q - T,$$

where we think of the above as a polynomial of degree at most $q - 1$. It is not hard to see that if P has total degree at most d, then $P|_L(T)$ also has degree at most d; in other words, it is a univariate Reed-Solomon codeword. This property – that the restriction of any codeword to a line is itself a codeword of another code – is extremely useful, and has been exploited in coding theory since Reed's majority logic decoder in the 1950's [22]. A natural question is whether or not there exist any bivariate polynomials $P(X, Y)$ *other* than those of total degree at most d which have this property. That is, are there polynomials which have high degree, but whose restrictions to lines are always low-degree? In many settings (for example, over the reals, or over prime fields) the answer is no. However, the insight of [12] is that there are settings – high degree polynomials over small-characteristic fields – for which the answer is yes.

This motivates the definition of *lifted codes*, which are multivariate polynomial evaluation codes, all of whose restrictions to lines lie in some other base code. Guo, Kopparty and

[4] We note that there have been some works in the intermediate-s parameter regime which can obtain excellent locality Q but are not directly relevant for Question 1. In particular, the work of [21] gives a construction of s-DRGP codes with $s = \Theta(K^{1/3-\varepsilon})$ and $Q = \Theta(K^{1/3})$ for arbitarily small constant ε; while this work obtains a smaller Q than we will eventially obtain (our results will have $Q \sim \sqrt{N}$), they are only able to establish high (constant) rate codes, and thus do not yield tight bounds on the co-dimension. The work of [3] gives constructions of high-rate fountain codes which have $s, Q = \Theta(\log(N))$. As these are rateless codes, again they are not directly relevant to Question 1.

Sudan showed that, in the case above, not only do these codes exist, but in fact they may have rate much higher than the corresponding Reed-Muller code.

Lifted codes very naturally give rise to codes with the s-DRGP. Indeed, consider the bivariate example above, with $d = q - 2$. That is, $\mathcal{C}$ is the set of codewords arising from evaluations of functions P that have the property that for all lines $L : \mathbb{F}_q \to \mathbb{F}_q^2$, $\deg(P|_L) \leq q - 2$. The restrictions then lie in the parity-check code: we always have $\sum_{t \in \mathbb{F}_q} P|_L(t) = 0$. Thus, for every coordinate of a codeword in $\mathcal{C}$ – which corresponds to an evaluation point $(x, y) \in \mathbb{F}_q^2$ – there are q disjoint repair groups for this symbol, corresponding to the q affine lines through (x, y).

However, it's not obvious how to use these codes to obtain the s-DRGP for $s \ll \sqrt{N}$; increasing the number of variables causes s to grow, and this is the approach taken in [12] to obtain high-rate LCCs. Since we are after smaller s, we take a different approach. We stick with bivariate codes, but instead of requiring that the functions $P \in \mathcal{F}$ restrict to low-degree polynomials on *all* affine lines L, we make this requirement only for *some* lines. This allows us to achieve the s-DRGP (if there are s lines through each point), while still being able to control the rate.

While special cases of this idea – notably tensor codes – have been considered before, allowing more complicated sets of lines requires some new machinery, and we hope that this machinery may be useful more generally. In the next section, we will set up our notation and give an outline of this approach, after a brief review of the notation we will use throughout the paper.

1.3 Outline

Next, in Section 2, we define *partially lifted codes*, and give a technical overview of our approach. This approach consists of two parts. The first is a general framework for understanding the dimension of partially lifted codes of a certain form, which we then discuss more in Section 3. The second part is to instantiate this framework, which we do in Section 4. This gives rise to the s-DRGP code with $s = N^{1/4}$ described above. Due to space constraints, we omit many details from this extended abstract, and refer the reader to the full version of the paper [9].

2 Technical Overview

In this section, we give a high-level overview of our construction and approach. We begin with some basic definitions and notation.

2.1 Notation and basic definitions

We study linear codes $\mathcal{C} \subseteq \mathbb{F}_q^N$ of block length N over an alphabet of size q. We will always assume that $\mathbb{F}_q$ has characteristic 2, and write $q = 2^\ell$. (We note that this is not strictly necessary for our techniques to apply – the important thing is only that the field is of relatively small characteristic – but it simplifies the analysis, and so we work in this special case).

The specific codes $\mathcal{C}$ that we consider are *polynomial evaluation codes*. Formally, let $\mathcal{F}$ be a collection of m-variate polynomials over $\mathbb{F}_q$. Letting $N = q^m$, we may identify $\mathcal{F}$ with a code $\mathcal{C} \subseteq \mathbb{F}_q^N$ as in (1); we assume that there is some fixed ordering on the elements of $\mathbb{F}_q^m$ to make this well-defined. For a polynomial $P \in \mathbb{F}_q[X_1, \ldots, X_m]$, we write its corresponding

codeword as

$$\mathsf{eval}(P) = \langle P(x_1, \ldots, x_m) \rangle_{(x_1,\ldots,x_m)\in\mathbb{F}_q^m} \in \mathcal{C}.$$

We will only focus on $m = 1, 2$, as we consider the restriction of bivariate polynomial codes to lines, which results in univariate polynomial codes. Formally, a (parameterization of an) *affine line* is a map $L : \mathbb{F}_q \to \mathbb{F}_q^2$, of the form $L(T) = (\alpha T + \beta, \gamma T + \delta)$ for $\alpha, \beta, \gamma, \delta \in \mathbb{F}_q$. We say that two parameterizations L, L' are *equivalent* if the result in the same line as a set: $\{L(t) : t \in \mathbb{F}_q\} = \{L'(t) : t \in \mathbb{F}_q\}$. We denote the restriction of a polynomial $P \in \mathbb{F}_q[X, Y]$ to L by $P|_L$:

▶ **Definition 3.** For a line $L : \mathbb{F}_q \to \mathbb{F}_q^2$ with $L(T) = (L_1(T), L_2(T))$, and a polynomial $P : \mathbb{F}_q^2 \to \mathbb{F}_q$, we define the *restriction* of P on L, denoted $P|_L : \mathbb{F}_q \to \mathbb{F}_q$, to be the unique polynomial of degree at most $q - 1$ so that $P|_L(T) = P(L_1(T), L_2(T))$.

We note that the definition above makes sense, because all functions $f : \mathbb{F}_q \to \mathbb{F}_q$ can be written as polynomials of degree at most $q - 1$ over $\mathbb{F}_q$; in this case, we have $P|_L(T) = P(L_1(T), L_2(T)) \mod (T^q - T)$.

▶ Remark 1. Throughout this paper, all polynomials will be considered mod $T^q - T$, although we will frequently drop this notation for ease of reading.

Finally, we'll need some tools for reasoning about integers and their binary expansions.

▶ **Definition 4.** Let $m < q$ be a positive integer. If $m = \sum_{i=0}^{\ell-1} m_i 2^i$, where $m_i \in \{0, 1\}$, then we let $B(m) = \{i \in \{0, ..., \ell - 1\} \mid m_i = 1\}$. That is, $B(m)$ is the set of indices where the binary expansion of m has a 1.

▶ **Definition 5.** For any two integers $m, n < q$, we say that m lies in the 2-shadow of n, denoted $m \leq_2 n$, if $B(m) \subseteq B(n)$. Equivalently, letting $m = \sum_{i=0}^{\ell-1} m_i 2^i$ and $n = \sum_{i=0}^{\ell-1} n_i 2^i$, we write $m \leq_2 n$ if for all $i \in \{0, ..., \ell - 1\}$, whenever $m_i = 1$ then also $n_i = 1$.

The reason that we are interested in 2-shadows is because of Lucas' Theorem.

▶ **Theorem 6** (Lucas' Theorem). *For any $m, n \in \mathbb{Z}$, $\binom{m}{n} \equiv 0 \mod 2$ exactly when $m \not\leq_2 n$.*

Finally, for integers a, b, s, we will say $a \equiv_s b$ if a is equal to b modulo s. For a positive integer n, we use $[n]$ to denote the set $[n] = \{0, \ldots, n - 1\}$.

2.2 Partially lifted codes

With the preliminaries out of the way, we proceed with a description of our construction and techniques. As alluded to above, our codes will be bivariate polynomial codes, which are "partial lifts" of parity check codes.

▶ **Definition 7.** Let $\mathcal{F}_0 \subseteq \mathbb{F}_q[T]$ be a collection of univariate polynomials, and let $\mathcal{L}$ be a collection of parameterizations of affine lines $L : \mathbb{F}_q \to \mathbb{F}_q^2$. We define the *partial lift* of $\mathcal{F}_0$ with respect to $\mathcal{L}$ to be the set

$$\mathcal{F} = \{P \in \mathbb{F}_q[X, Y] : \forall P \in \mathcal{F}, \forall L \in \mathcal{L}, P|_L \in \mathcal{F}_0\}.$$

We make a few remarks about Definition 7 before proceeding.

▶ Remark 2 (Equivalent lines). We remark that the definition above allows $\mathcal{L}$ to be a collection of *parameterizations* of lines. A priori, it is possible that equivalent parameterizations may behave very differently with respect to $\mathcal{F}_0$, and it is also possible to include several equivalent parameterizations in $\mathcal{L}$. In this work, $\mathcal{F}_0$ will always be affine-invariant (in particular, it will just be the set of polynomials of degree strictly less than $q-1$), and so if L and L' equivalent, then $P|_L \in \mathcal{F}_0$ if and only if $P|_{L'} \in \mathcal{F}_0$. Thus, these issues won't be important for this work.

▶ Remark 3 (Why only bivariate lifts?). This definition works just as well for m-variate partial lifts, and we hope that further study will explore this direction. However, as all of our results are for bivariate codes, we will stick to the bivariate case to avoid having to introduce another parameter.

Let $\mathcal{F}_0 := \{P \in \mathbb{F}_q[X], \deg(P) < q-1\}$. Then it is not hard to see that the code $\mathcal{C}_0 = \{\mathsf{eval}(P) : P \in \mathcal{F}_0\}$ is just the parity-check code, $\mathcal{C}_0 = \left\{c \in \mathbb{F}_q^q : \sum_{i=1}^q c_i = 0\right\}$. Indeed, for any $d < q-1$, we have $\sum_{x \in \mathbb{F}_q} x^d = 0$.

We will construct codes with the s-DRGP by considering codes that are partial lifts of $\mathcal{F}_0$. We first observe that such codes, with an appropriate set of lines $\mathcal{L}$, will have the s-DRGP. Indeed, suppose we wish to recover a particular symbol, given by $P(x,y)$ for $(x,y) \in \mathbb{F}_q^2$. Let $L^{(1)}, \ldots, L^{(s)} \in \mathcal{L}$ be s distinct (non-equivalent) lines that pass through (x,y); say they are parameterized so that $L^{(j)}(0) = (x,y)$. Then the s disjoint repair groups are the sets indices corresponding to $S_j := \{L^{(j)}(t) : t \in \mathbb{F}_q \setminus \{0\}\}$. For any P in the partial lift of $\mathcal{F}_0$, we have $P|_L(0) = \sum_{t \in \mathbb{F}_q \setminus \{0\}} P|_L(t)$, which means that $P(x,y) = \sum_{(a,b) \in S_j} P(a,b)$. That is, $P(x,y)$ can be recovered from the coordinates of $\mathsf{eval}(P)$ indexed by S_j, as desired. Finally we observe that the S_j are all disjoint, as the lines are all distinct, and intersect only at (x,y). We summarize the above discussion in the following observation.

▶ **Observation 8.** *Suppose that $\mathcal{F}_0 = \{P \in \mathbb{F}_q[T] : \deg(P) < q-1\}$, and let $\mathcal{L}$ be any collection of parameterizations of affine lines so that every point in $\mathbb{F}_q^2$ is contained in at least s non-equivalent elements of $\mathcal{L}$. Let $\mathcal{F}$ be the bivariate partial lift of $\mathcal{F}_0$ with respect to $\mathcal{L}$. Then the code $\mathcal{C} \subseteq \mathbb{F}_q^{q^2}$ corresponding to $\mathcal{F}$ is a linear code with the s-DRGP.*

To save on notation later, we say that a polynomial $P : \mathbb{F}_q^2 \to \mathbb{F}_q$ *restricts nicely* on a line $L : \mathbb{F}_q \to \mathbb{F}_q^2$ if $P|_L$ has degree strictly less than $q-1$. Thus, to define our construction, we have to define the collection $\mathcal{L}$ of lines used in Definition 7. We will actually develop a framework that can handle a family of such collections, but for intuition in this section, let us just consider lines $L(T) = (T, \alpha T + \beta)$ where α lives in a multiplicative subgroup G_s of $\mathbb{F}_q^*$ of size s, and $\beta \in \mathbb{F}_q$. That is, we are essentially restricting the slope of the lines to lie in a multiplicative subgroup. It is not hard to see that every point $(x,y) \in \mathbb{F}_q^2$ has s non-equivalent lines in $\mathcal{L}$ that pass through it.

Following Observation 8, the resulting code will immediately have the s-DRGP. The only question is, what is the rate of this code? Equivalently, we want to know:

▶ **Question 2.** *How many polynomials $P \in \mathbb{F}_q[X,Y]$ have $\deg(P|_L) < q-1$ for all $L \in \mathcal{L}$, where $\mathcal{L}$ is as described above?*

In [12], Guo, Kopparty and Sudan develop some machinery for answering this question when $\mathcal{L}$ is the set of all affine lines. What they show in that work is that in fact the (fully) lifted code is affine-invariant, and is equal to the span of the monomials $P(X,Y) = X^a Y^b$ so that $\deg(P|_L) < q-1$ for all affine lines L. We might first hope that this is the case for partial lifts – but then upon reflection we would immediately retract this hope, because it turns out that we do not get any more monomials this way: Theorem 13 establishes that if a monomial restricts nicely on even one line of the form $(T, \alpha T + \beta)$ (for nonzero α, β), then in fact it

restricts nicely on *all* such lines. In fact, the partial lift is not in general affine-invariant, and this is precisely where we are able to make progress. More precisely, there may be polynomials $P(X, Y)$ of the form

$$P(X, Y) = X^{a_1}Y^{b_1} + X^{a_2}Y^{b_2} \tag{2}$$

which are contained in the partial lift $\mathcal{F}$, but so that $X^{a_1}Y^{b_1}, X^{a_2}Y^{b_2} \notin \mathcal{F}$. This gives us many more polynomials to use in a basis for $\mathcal{F}$ than just the relevant monomials, and allows us to construct families $\mathcal{F}$ of larger dimension.

▶ Remark 4 (Breaking affine invariance). We emphasize that breaking affine-invariance is a key departure from [12]. In some sense, it is not surprising that we are able to make progress by doing this: the assumption of affine-invariance is one way to prove *lower bounds* on locality [4, 5]. This is also where our techniques diverge from those of [12]. Because of their characterization of affine-invariant codes, that work focused on understanding the dimension of the relevant set of monomials. This is not sufficient for us, and so to get a handle on the dimension of our constructions, we must study more complicated polynomials. This may seem daunting, but we show – perhaps surprisingly – that one can make a great deal of progress by considering only the additional "more complicated" polynomials of the form (2), which are arguably the simplest of the "more complicated" polynomials.

In order to obtain a lower bound on the dimension of $\mathcal{F}$, our strategy get a handle on the dimension of the space of these binomials (2). If we can show that there are many linearly independent such binomials, then the answer to Question 2 must be "lots."

Following this strategy, we examine binomials of the form (2), and we ask, for which a_1, b_1, a_2, b_2 and which $L(T) = (T, \alpha T + \beta)$ does $P(X, Y)$ restrict nicely? Our main tool is Lucas's Theorem (Theorem 6), which was also used in [12]. To see why this is useful, consider the restriction of a monomial $P(X, Y) = X^aY^b$ to a line $L(T) = (T, \alpha T + \beta)$. We obtain

$$P|_L(T) = T^a (\alpha T + \beta)^b = \sum_{i \le b} \binom{b}{i} \alpha^i \beta^{b-i} T^{a+i}.$$

Above, the binomial coefficient $\binom{b}{j}$ is shorthand for the sum of 1 with itself $\binom{b}{j}$ times. Thus, in a field of characteristic 2, this is either equal to 1 or equal to 0; Lucas's theorem tells us which it is. This means that our question reduces to asking, when does the coefficient of T^{q-1} vanish? The above gives us an expression for this coefficient, and allows us to compute an answer, in terms of the binary expansions of a and b.

So far, this is precisely the approach of [12]. From here, we turn to the binomials of the form (2). When do these restrict nicely? As above, we may compute the coefficient of the T^{q-1} term and examine it. Fortunately, when the set of lines $\mathcal{L}$ is chosen as above, the number of linearly independent binomials that restrict nicely ends up having a nice expression, in terms of the number of non-empty equivalence classes of a particular relation defined by the binary expansion of the numbers $1, \ldots, q-1$; this is our main technical theorem (Theorem 12, which is proved in Section 3.2).

The approach of Section 3.2 holds for more general families than the $\mathcal{L}$ described above; instead of taking α in a multiplicative subgroup of $\mathbb{F}_q^*$, we may alternately restrict β, or restrict both. However, numerical calculations indicated that the choice above (where α is in a multiplicative subgroup of order s) is a good one, so for our construction we make this choice and we focus on that for our formal analysis in Section 4.

In order to get our final construction and obtain the results advertised above, it suffices to count these equivalence classes. For the result advertised in the introduction, we choose

the order of the multiplicative subgroup to be $s = 2^{\ell/2} - 1 = \sqrt{q} - 1$. Then, we use an inductive argument in Section 4 to count the resulting equivalence classes, obtaining the bounds advertised above. More precisely, we obtain the following theorem.

▶ **Theorem 9.** *Suppose that $q = 2^\ell$ for even ℓ, and let $N = q^2 - 1$. There is a linear code $\mathcal{C}$ over $\mathbb{F}_q$ of length N and dimension*

$$K \geq N - O(N^{.714})$$

which has the s-DRGP for $s = \sqrt{q} - 2 = (N+1)^{1/4} - 1$.

▶ Remark 5 (Puncturing at the origin). We note that the statement of the theorem differs slightly from the informal description above; in our analysis, we will puncture the origin, and ignore lines that go through the origin; that is, our codes will have length $q^2 - 1$, rather than q^2, and the number of lines through every point will be $s - 1$, rather than s, as it makes the calculations somewhat easier and does not substantially change the results.

2.3 Discussion and open questions

Before we dive into the technical details in Section 3, we close the front matter with some discussion of open questions left by our work and our approach. We view the study of the s-DRGP for intermediate s to be an important step in understanding locality in general, since the s-DRGP nicely interpolates between the two extremes of LRCs and LCCs. When $s = 2$, we completely understand the answer to Question 1. However, by the time s reaches $\Omega(N)$, this becomes a question about the best rate of constant-query LCCs, which is a notoriously hard open problem. It is our hope that by better understanding the s-DRGP, we can make progress on these very difficult questions.

The main question left by our work is Question 1, which we do not answer. What is the correct dependence on s in the codimension of codes with the s-DRGP? We have shown that it is not $s\sqrt{N}$, even for $s \ll \sqrt{N}$. However, we have no reason to believe that our construction is optimal.

Our work also raises questions about partially lifted codes. These do not seem to have been studied before. The most immediate question arising from our work is to improve or generalize our approach; in particular, is our analysis tight? Our approach proceeds by counting the binomials of the form (2). This is in principle lossy, but empirical simulations suggest that at least in the setting of Theorem 9, this approach is basically tight. Are there situations in which this is not tight? Or can we prove that it is tight in any situation? Finally, are there other uses of partially lifted codes? As with lifted codes, we hope that these prove useful in a variety of settings.

3 Framework

As discussed in the previous section, the proof of Theorem 9 is based on the partially lifted codes of Definition 7. In this section, we lay out the partially lifted codes we consider, as well as the basic tools we need to analyze them. As before, we say that a polynomial $P : \mathbb{F}_q^2 \to \mathbb{F}_q$ *restricts nicely* to a line $L : \mathbb{F}_q \to \mathbb{F}_q^2$ if $P|_L$ has degree strictly less than $q - 1$. We will consider partial lifts of the parity-check code with respect to a collection of affine lines $\mathcal{L}$; reasoning about the rate of this code will amount to reasoning about the polynomials which restrict nicely to lines in $\mathcal{L}$. To ease the computations, we will form our family $\mathcal{L}$ out of lines that have a simple parameterization:

▶ **Definition 10.** We say a line $L : \mathbb{F} \to \mathbb{F}^2$ is *simple* if it can be written in the form $L(T) = (T, \alpha T + \beta)$, with $\alpha, \beta \neq 0$.

Notice that this rules out lines through the origin. At the end of the day, we will pucture our code at the origin to achieve our final result. Note also that no two simple parameterizations of lines are equivalent to each other (that is, they form distinct lines as sets), so as we go forward, we may apply Observation 8 without worry of the repair groups coinciding.

We consider a family of constructions, indexed by parameters s and t, so that $s, t \mid q-1$. This family will be the partial lift with respect to the following set of simple lines.

▶ **Definition 11.** Let $s, t \mid q-1$, and let $G_s, G_t \subseteq \mathbb{F}_q^*$ be multiplicative subgroups of $\mathbb{F}_q^*$ of orders s and t, respectively. That is, $G_s = \left\{x \in \mathbb{F}_q^* : x^s = 1\right\}$ and $G_t = \left\{x \in \mathbb{F}_q^* : x^t = 1\right\}$. Then we define $\mathcal{L}_{s,t}$ to be the family of simple lines

$$\mathcal{L}_{s,t} = \{L(T) = (T, \alpha T + \beta) : \alpha \in G_s, \beta \in G_t\}.$$

For the rest of the paper, we will study the following construction, for various choices of s and t.

▶ **Construction 1.** *Let $\mathcal{L}_{s,t}$ be as in Definition 11 for $s, t \mid q-1$, and let $\mathcal{F}_0$ be the set of univariate polynomials of degree strictly less than $q-1$. Define $\mathcal{F}_{s,t}$ to be the partial lift of $\mathcal{F}_0$ with respect to $\mathcal{L}_{s,t}$.*

Our main theorem, which we will prove in the rest of this section, is a characterization of the dimension of $\mathcal{F}_{s,t}$ as in Construction 1. (We recall the definition of $\leq_2$ from Definition 5 above).

▶ **Theorem 12.** *Suppose that $s, t \mid q-1$. For nonnegative integers $i < s, j < t$, define*

$$e(s,t) = \Big|\{(i,j) : i < s, \text{ and } j < t, \\ \text{so that there is some } m, n \in [q]^2 \text{ with } m \equiv_s i, n \equiv_t j, \text{ and } n \leq_2 m\}\Big|.$$

Then the dimension of $\mathcal{F}_{s,t} \subseteq \mathbb{F}_q[X, Y]$ is at least

$$\dim(\mathcal{F}_{s,t}) \geq q^2 - e(s,t).$$

Theorem 12 may seem rather mysterious. The expression $e(s,t)$ comes up in counting the number of binomials of the form (2) the restrict nicely on lines in $\mathcal{L}_{s,t}$. We omit the full proof of Theorem 12 in this extended abstract, but we will sketch the outline in Section 3.2.

The reason that Theorem 12 is useful is that for some s and t, it turns out to be possible to get a very tight handle on $e(s,t)$, which leads to the quantitative result in Thorem 9. For now, we focus on proving Theorem 12. Our starting point is the work of [12]; we summarize the relevant points below in Section 3.1.

3.1 Basic Setup: Lucas' Theorem and Monomials

In [12], Guo, Kopparty and Sudan give a characterization of lifted codes. In our setting, their work shows that when the set $\mathcal{L}$ is the set of *all* affine lines, then the lifted code $\mathcal{F}$ is affine invariant and in fact is equal to the span of the *monomials* which restrict nicely. In the case where the number of variables is large, or the base code $\mathcal{F}_0$ is more complicated than a parity-check code, [12] provides some bounds, but it seems quite difficult to get a tight characterization of these monomials. However, for bivariate lifts of the parity-check

code, it is actually possible to completely understand the situation, and this was essentially done in [12]. We review their approach here.

First, we use Lucas' Theorem (Theorem 6) to characterize which monomials X^aY^b restrict nicely to simple lines. Theorem 13 follows from the analysis in [12]; we refer the reader to the full version of this paper [9] for a direct proof.

▶ **Theorem 13.** *Suppose $a + b < 2(q-1)$ and let $P(X,Y) = X^aY^b$. Then for all simple lines $L(T) = (T, \alpha T + \beta)$, $P|_L$ has degree $< q-1$ if and only if $q-1-a \not\leq_2 b$. Further, if $q-1-a \leq_2 b$, then $P|_L$ is a degree $q-1$ polynomial with leading coefficient $\alpha^{-a}\beta^{b+a}$*

Theorem 13 implies that whether a monomial $P(X,Y) = X^aY^b$ restricts nicely to a simple line L is independent of the choice of L. Thus it makes sense to consider this a property of the monomial itself.

▶ **Definition 14.** We say that a monomial $P(X,Y) = X^aY^b$ with $0 \leq a, b \leq q-1$ is *good* if it restricts nicely on all simple lines.

▶ Remark 6 (The special case of $X^{q-1}Y^{q-1}$). In Theorem 13, we required $a + b < 2(q-1)$, which does not cover the monomial $P_*(X,Y) = X^{q-1}Y^{q-1}$. However, in Definition 14, we allow $a = b = q-1$, and in fact according to this definition $P_*(X,Y)$ is good; we will treat it that way in this work, even though it would not be considered good in the analysis of [12]. (In their language, P_* does not live in the lift of the degree set $\{0, \ldots, q-2\}$).

Theorem 13 implies (see [9]) that there are $q^2 - 3^\ell + 1$ good monomials. This allows us to recover the codes of Theorem 1.2 in [12] up to the technicalities about simple lines vs. all lines. Following Observation 8, these codes have the s-DRGP for $s = q-1$; indeed, there are $q-1$ simple lines through every non-zero point of $\mathbb{F}_q^2$. The dimension of these codes is at least the number of monomials that they contain (indeed, all monomials are linearly independent), which by the above is at least $q^2 - 3^\ell + 1 = (N+1) - (N+1)^{\log_4(3)} + 1$.

▶ **Corollary 15** (Implicit in [12]). *There are codes linear $\mathcal{C}$ over $\mathbb{F}_q$ of length $N = q^2 - 1$ with dimension $K \geq N + 2 - (N+1)^{\log_4(3)}$ which have the s-DRGP for $s = q - 1 = \sqrt{N+1} - 1$.*

We note that this recovers the results of one of the classical constructions of the s-DRGP for $s = \sqrt{N}$ mentioned in the introduction (and this is not an accident: these codes are in fact the same as affine geometry codes). In the next section, we show how to use the relaxation to partial lifts in order to create codes with the s-DRGP for $s \ll \sqrt{N}$.

3.2 Partially lifted codes

In this section we extend the analysis above to partial lifts. The work of [12] characterizes the polynomials which restrict nicely on all lines $L : \mathbb{F}_q \to \mathbb{F}_q^2$: they show that this is exactly the span of the good monomials (except the special monomial P_* of Remark 6, which restricts to degree lower than $q-1$ only on *simple* lines). However, since our goal is to obtain codes with the s-DRGP for $s \ll \sqrt{N}$, increasing the dimension while decreasing s, we would like to allow for more polynomials.

Thus, as in Definition 7, we will consider polynomials which restrict nicely only on some particular subset $\mathcal{L}$ of simple lines. We would like to find a subset $\mathcal{L}$ such that the space of polynomials which restrict nicely on all lines in $\mathcal{L}$ has large degree. Additionally, we would like to guarantee the s-DRGP by ensuring that, for every point (x,y), there are many lines in $\mathcal{L}$ that pass through (x,y). Relaxing requirements in this manner will allow us to get codes with good rate and locality trade-offs.

Theorem 13 shows that if a monomial restricts nicely on one simple line, it will restrict nicely on all simple lines. This means that in order to find a larger space of polynomials, we cannot only consider monomials. Towards this end, we will consider *binomials* of the form

$$P(X,Y) = X^{a_1}Y^{b_1} + X^{a_2}Y^{b_2}. \tag{3}$$

That is, we will look only at binomials with both coefficients equal to 1.

We note that this ability to extend beyond monomials is possible crucially because our partially lifted codes are not affine-invariant. While affine-invariance allowed [12] to get a beautiful characterization of (fully) lifted codes, it also greatly restricts the flexibility of these codes. By breaking affine-invariance, we also break some of the rigidity of these constructions. This is in some sense not surprising: affine invariance is often exploited in order to prove *lower bounds* on locality [4, 5].

3.2.1 Which binomials play nice with which lines?

We would like to characterize which binomials of the form (3) restrict nicely on which lines. Unlike the case with monomials, now this will depend on the line as well as on the binomial. When both individual terms in the binomial are good monomials, the binomial will certainly restrict nicely. However, if this is not the case, then the binomial could still restrict nicely, if the contributions to the leading coefficient of $P|_L$ from the two terms cancel with each other. Using Theorem 13, we may write down these contributions and characterize when the cancel; we omit the details due to space constraints, but (see [9]) this approach can establish the following Corollary.

▶ **Corollary 16.** *Let s and t divide $q-1$, and let $G_s = \{x \in \mathbb{F}_q : x^s = 1\}$ and $G_t = \{x \in F_q : x^t = 1\}$. Let*

$$\mathcal{L}_{s,t} = \{(T, \alpha T + \beta) : \alpha \in G_s, \beta \in G_t\}$$

as in Definition 11. Suppose that $P(X,Y) = X^{a_1}Y^{b_1} + X^{a_2}Y^{b_2}$ is a binomial so that neither term is good. Suppose that $a_1 \equiv a_2 \mod s$ and $a_1 + b_1 \equiv a_2 + b_2 \mod t$. Then for all $L \in \mathcal{L}_{s,t}$, P restricts nicely to L.

Thus, a choice of s and t dividing $q-1$ produces a code by using $\mathcal{L}_{s,t}$ in Construction 1. Each choice of s and t produces a different code, and by varying s and t we can vary the parameters of this code. This is the general framework for our construction, but we still must explore the dimension and the number of disjoint repair groups produced by different choices of s and t.

3.2.2 Dimension

Given some choice of s and t, we would like to understand dimension of the space of polynomials $\mathcal{F}_{s,t}$ which restrict nicely on all lines in $\mathcal{L}_{s,t}$. We will lower bound this dimension by building a linearly independent set $S \subseteq \mathcal{F}_{s,t}$ comprised of monomials and binomials. In order to construct S and understand its size, we will need some more notation.

Let $i < s$ and $j < t$ be nonnegative integers. Define

$$E_{i,j} = \{(m,n) \in [q]^2 : m \equiv_s i, n \equiv_t j, n \leq_2 m\}.$$

Thus, the term $e(s,t)$ from Theorem 12 is the number of (i,j) so that $E_{i,j}$ is not empty. It turns out, that $E_{i,j}$ is (up to a ± 1 term that we are careful about in the full version) in

bijection with the set $\hat{M}_{i,j} = \left\{X^a Y^b \text{ not good } : a \equiv_s i, b+a \equiv_t j\right\}$. Notice that the sum of two monomials in $\hat{M}_{i,j}$ meets the hypotheses of Corollary 16.

This observation is at the heart of the proof of Theorem 12. In slightly more detail, we want to establish a lower bound on the dimension of polynomials which restrict nicely; to do this we will exhibit a large linearly independent set of such polynomials. We will start with all of the good monomials, and add to them a collection of binomials that satisfy Corollary 16. We can do this as follows. First, from each $\hat{M}_{i,j}$, we fix one monomial, call it $X^{a^*}Y^{b^*}$. Then, we include into our large linearly independent set all the binomails of the form $X^{a^*}Y^{b^*} + X^a Y^b$ for $X^a Y^b \in \hat{M}_{i,j} \setminus X^{a^*}Y^{b^*}$. Doing this for all i, j results in a collection of linearly independent binomials of size at least (ignoring some details about ± 1 terms)

$$\sum_{|E_{i,j} \neq 0} (|E_{i,j} - 1) - 1 = \left(\sum_{|E_{i,j}| \neq 0} |E_{i,j}| - 1 \right) - e(s,t).$$

However, the first term, which is equal to $\sum_{i,j} |E_{i,j}| - 1$, is exactly the number of not-good monomials. So our count of good monomials, plus these binomials that restrict nicely, is precisely equal to the number of all monomials, minus $e(s,t)$. This establishes Theorem 12; we refer the reader to [9] for more details.

This theorem does give us a lower bound on the dimension of the code, but the expression depends on $e(s,t)$. We would like to know that $e(s,t)$ is not too big. It is easy to see that $e(s,t) \leq st$, because there are only st choices for (i,j). Moreover, we know that $e(s,t) \leq q^2 - g = 3^\ell - 1$, the total number of not-good monomials. As we will see in Section 4, this first bound $e(s,t) \leq st$ is nontrivial, and can in fact recover the result of $N - K = s\sqrt{N}$ of [8]. However, the point of all this work is that in fact we will be able to choose s and t so that we can get a much tighter bound on $e(s,t)$, establishing Theorem 9.

4 Instantiations

Finally, we choose t and s. One of the simplest choices we can make within our framework is to set $t = q-1$, while $s|q-1$ is any divisor. That is, we consider all simple lines $L(T) = (T, \alpha T + \beta)$ where β may vary over all of $\mathbb{F}_q^*$, and where $\alpha \in G_s$ lives in a multiplicative subgroup of $\mathbb{F}_q^*$. One reason that this choice is convenient is that it is easy to understand the number of disjoint repair groups: there are $s-1$ lines of $\mathcal{L}_{s,q-1}$ through any nonzero point.

Thus, Theorem 12, along with the observation of the previous section that $e(s, q-1) \leq s(q-1)$ trivially, immediately implies DRGP codes that match the results of [8], with dimension $K \geq N - O(s\sqrt{N})$. However, by choosing s carefully we can actually get a tighter bound on $e(s,t)$:

▶ **Theorem 17.** *Let $q = 2^\ell$ be an even power of 2. Then*

$$e(\sqrt{q} - 1, q - 1) = O\left((5 + \sqrt{5})^{\ell/2}\right).$$

Theorem 9 follows straightforwardly from Theorem 17 and Theorem 12. We omit the proof of Theorem 17 here, and refer the reader to the full version [9] for details.

5 Conclusion

We have studied the s-DRGP for intermediate values of s. As s grows, the study of the s-DRGP interpolates between the study of LRCs and LCCs, and our hope is that by

understanding intermediate s, we will improve our understanding on either end of this spectrum. Using a new construction that we term a "partially lifted code," we showed how to obtain codes of length N with the s-DRGP for $s = \Theta(N^{1/4})$, that have dimension $K \geq N - N^{.714}$. This is an improvement over previous results of $N - N^{3/4}$ in this parameter regime. We stress that the main point of interest of this result is not the exponent 0.714, which we do not believe is tight for Question 1; rather, we think that our results are interesting because (a) they show that one can in fact beat $N - O(s\sqrt{N})$ for $s = N^{1/4} \ll \sqrt{N}$, and (b) they highlight the class of partially lifted codes, which we hope will be of independent interest.

Acknowledgements. We thank Alex Vardy and Eitan Yaakobi for helpful exchanges. We also thank the anonymous reviewers for suggestions which improved the paper.

References

1 Hilal Asi and Eitan Yaakobi. Nearly optimal constructions of PIR and batch codes. *CoRR*, abs/1701.07206, 2017. URL: http://arxiv.org/abs/1701.07206.

2 E. F. Assmus and J. D. Key. Polynomial codes and finite geometries. *Handbook of coding theory*, 2(part 2):1269–1343, 1998.

3 Megasthenis Asteris and Alexandros G. Dimakis. Repairable fountain codes. *IEEE Journal on Selected Areas in Communications*, 32(5):1037–1047, 2014.

4 Eli Ben-Sasson and Madhu Sudan. Limits on the rate of locally testable affine-invariant codes. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 412–423. Springer, 2011.

5 Arnab Bhattacharyya and Sivakanth Gopi. Lower bounds for constant query affine-invariant LCCs and LTCs. In *Proceedings of the 31st Conference on Computational Complexity*, volume 50 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 12:1–12:17. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2016. doi:10.4230/LIPIcs.CCC.2016.12.

6 S. Blackburn and T. Etzion. PIR Array Codes with Optimal PIR Rate. *CoRR*, abs/1607.00235, 2016. URL: http://arxiv.org/abs/1607.00235.

7 Alexandros G Dimakis, Anna Gál, Ankit Singh Rawat, and Zhao Song. Batch codes through dense graphs without short cycles. *arXiv preprint arXiv:1410.2920*, 2014.

8 Arman Fazeli, Alexander Vardy, and Eitan Yaakobi. Codes for distributed PIR with low storage overhead. In *2015 IEEE International Symposium on Information Theory (ISIT)*, pages 2852–2856. IEEE, 2015.

9 S Luna Frank-Fischer, Venkatesan Guruswami, and Mary Wootters. Locality via partially lifted codes. *arXiv preprint arXiv:1704.08627*, 2017.

10 Parikshit Gopalan, Cheng Huang, Huseyin Simitci, and Sergey Yekhanin. On the locality of codeword symbols. *IEEE Transactions on Information Theory*, 58(11):6925–6934, 2012.

11 Alan Guo and Swastik Kopparty. List-decoding algorithms for lifted codes. *CoRR*, abs/1412.0305, 2014. URL: http://arxiv.org/abs/1412.0305.

12 Alan Guo, Swastik Kopparty, and Madhu Sudan. New affine-invariant codes from lifting. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, ITCS'13, pages 529–540, New York, NY, USA, 2013. ACM. URL: http://arxiv.org/abs/1208.5413, arXiv:1208.5413, doi:10.1145/2422436.2422494.

13 Brett Hemenway, Rafail Ostrovsky, and Mary Wootters. Local Correctability of Expander Codes. In *ICALP*, LNCS. Springer, April 2013. arXiv:1304.8129.

14 Cheng Huang, Minghua Chen, and Jin Li. Pyramid codes: Flexible schemes to trade space for access efficiency in reliable data storage systems. *ACM Transactions on Storage (TOS)*, 9(1):3, 2013.

15 Yuval Ishai, Eyal Kushilevitz, Rafail Ostrovsky, and Amit Sahai. Batch codes and their applications. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 262–271. ACM, 2004.

16 Jonathan Katz and Luca Trevisan. On the efficiency of local decoding procedures for error-correcting codes. In *STOC'00: Proceedings of the 32nd Annual Symposium on the Theory of Computing*, pages 80–86, 2000.

17 S. Kopparty, S. Saraf, and S. Yekhanin. High-rate codes with sublinear-time decoding. In *Proceedings of the 43rd annual ACM symposium on Theory of computing*, pages 167–176. ACM, 2011.

18 Swastik Kopparty, Or Meir, Noga Ron-Zewi, and Shubhangi Saraf. High-rate locally-correctable and locally-testable codes with sub-polynomial query complexity. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*, pages 202–215. ACM, 2016.

19 Shu Lin and Daniel J Costello. *Error control coding*. Pearson Education India, 2004.

20 Sankeerth Rao and Alexander Vardy. Lower bound on the redundancy of PIR codes. *CoRR*, abs/1605.01869, 2016. URL: `http://arxiv.org/abs/1605.01869`.

21 Ankit Singh Rawat, Dimitris S. Papailiopoulos, Alexandros G. Dimakis, and Sriram Vishwanath. Locality and availability in distributed storage. In *2014 IEEE International Symposium on Information Theory*, pages 681–685. IEEE, 2014.

22 I. Reed. A class of multiple-error-correcting codes and the decoding scheme. *Information Theory, Transactions of the IRE Professional Group on*, 4(4):38–49, September 1954.

23 Maheswaran Sathiamoorthy, Megasthenis Asteris, Dimitris Papailiopoulos, Alexandros G Dimakis, Ramkumar Vadali, Scott Chen, and Dhruba Borthakur. Xoring elephants: Novel erasure codes for big data. In *Proceedings of the VLDB Endowment*, volume 6, pages 325–336. VLDB Endowment, 2013.

24 Vitaly Skachek. Batch and PIR codes and their connections to locally-repairable codes. *CoRR*, abs/1611.09914, 2016. URL: `http://arxiv.org/abs/1611.09914`.

25 Itzhak Tamo and Alexander Barg. Bounds on locally recoverable codes with multiple recovering sets. In *2014 IEEE International Symposium on Information Theory*, pages 691–695. IEEE, 2014.

26 Itzhak Tamo and Alexander Barg. A family of optimal locally recoverable codes. *IEEE Transactions on Information Theory*, 60(8):4661–4676, 2014.

27 Itzhak Tamo, Alexander Barg, and Alexey Frolov. Bounds on the parameters of locally recoverable codes. *IEEE Transactions on Information Theory*, 62(6):3070–3083, 2016.

28 Anyu Wang and Zhifang Zhang. Repair locality with multiple erasure tolerance. *IEEE Transactions on Information Theory*, 60(11):6979–6987, 2014.

29 David P. Woodruff. *A Quadratic Lower Bound for Three-Query Linear Locally Decodable Codes over Any Field*, pages 766–779. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

30 Sergey Yekhanin. Locally Decodable Codes. *Foundations and Trends in Theoretical Computer Science*, 2010.