Capacity-achieving Polar-based LDGM Codes with Crowdsourcing Applications

James (Chin-Jen) Pang, Hessam Mahdavifar, and S. Sandeep Pradhan
Department of Electrical Engineering and Computer Science, University of Michgan, Ann Arbor, MI 48109, USA
Email: cjpang, hessam, pradhanv@umich.edu

***Abstract*—In this paper we study codes with sparse generator matrices. More specifically, codes with a certain constraint on the weight of all the columns in the generator matrix are considered. The end result is the following. For any binary-input memoryless symmetric (BMS) channel and any $\epsilon > 2\epsilon^*$, where $\epsilon^* = \frac{1}{6} - \frac{5}{3}\log\frac{4}{3} \approx 0.085$, we show an explicit sequence of capacity-achieving codes with all the column weights of the generator matrix upper bounded by $(\log N)^{1+\epsilon}$, where N is the code block length. The constructions are based on polar codes. Applications to crowdsourcing are also shown.**

I. Introduction

Capacity-approaching error-correcting codes such as low-density parity-check (LDPC) codes [1] and polar codes [2] have been extensively studied for applications in wireless and storage systems. Besides conventional applications of codes for error correction, a surge of new applications has also emerged in the past decade including crowdsourcing [3], [4], distributed storage [5], and speeding up distributed machine learning [6]. To this end, new motivations have arisen to study codes with sparsity constraints in their encoding and/or decoding processes. For instance, the stored data in a failed server needs to be recovered by downloading data from a few servers only, due to bandwidth constraints, imposing sparsity constraints in the decoding process in a distributed storage system. In crowdsourcing applications, e.g., when workers are asked to label items in a dataset, each worker can be assigned a few items only due to capability limitations imposing sparsity constraints in the encoding process. More specifically, low-density generator matrix (LDGM) codes become relevant for such applications [7], [8].

A. LDGM and Related Works

LDGM codes, often regarded as the dual of LDPC codes, are associated with sparse factor graphs. The sparsity of the generator matrices of LDGM codes implies low encoding complexity. However, unlike LDPC and polar codes, LDGM code has not received significant attention. In [9], [10] it is pointed out that certain constructions of LDGM codes are not asymptotically *good*, a behavior which is also studied by an error floor analysis in [11], [12]. Several prior works, e.g., [11]–[13], adopt concatenation of two LDGM codes to construct systematic capacity-approaching LDGM codes with significantly lower error floors in simulations. As a sub-class of LDPC codes, the systematic LDGM codes are advantageous for their low encoding and decoding complexity.

In terms of the sparsity of the generator matrices, [14] showed the existence of capacity achieving codes over binary symmetric channels (BSC) using random linear coding arguments when the column weights of the generator matrix are upper bounded by ϵN, for any fixed $\epsilon > 0$, where N is the code block length. Also, it is conjectured in [14] that column weights polynomially sublinear in N suffice to achieve the capacity. For binary erasure channels (BEC), column weights being $O(\log N)$ suffice for capacity achieving, again using random linear coding arguments [14]. Furthermore, the scaling exponent of such random linear codes are studied in [15]. Later, in [16], the existence of capacity achieving systematic LDGM ensembles over any BMS channel with the expected value of the weight of the entire generator matrix upper bounded by ϵN^2, for any $\epsilon > 0$, is shown.

In [8], we formulated the problem of label learning through asking queries from crowd workers as a coding theory problem. Due to practical constraints in such crowdsourcing scenarios, each query can only contain a small number of items. When some workers do not respond, resembling a binary erasure channel, we showed that a combination of LDPC codes and LDGM codes gives a query scheme where the number of queries approaches information theoretic lower bound [8].

B. Our Contributions

In this paper, we focus on studying capacity achieving LDGM codes over BMS channels with sparsity constraints on column weights. Leveraging polar codes, invented by Arıkan [2], and their extensions to large kernels, with errors exponents studied in [17], we show that capacity-achieving polar codes with column weights bounded by any polynomial of N exist. However, a similar result can not be obtained with any polynomial of $\log N$ as the constraint on column weights. A new construction for LDGM codes is proposed so that most of the column weights can be bounded by a degree $1 + \delta''$ polynomial of $\log N$, where $\delta'' > 0$ can be chosen arbitrarily small. One issue of the new construction is the existence of, though only a few, *heavy* columns in the generator matrix. In order to resolve this, we propose a splitting algorithm which, roughly speaking, splits *heavy* columns into several *light* columns, a process which will be

This work was supported by the National Science Foundation under grants CCF–1717299, CCF–1763348, and CCF–1909771.

clarified in the paper. The rate loss due to this modification is characterized and is shown to approach zero as N grows large. Hence, the proposed modification leads to capacity achieving constructions with column wights of the generator matrix upper bounded by $(\log N)^{1+\epsilon}$, for any $\epsilon > 2\epsilon^*$, where $\epsilon^* = \frac{1}{6} - \frac{5}{3}\log\frac{4}{3} \approx 0.085$.

In crowdsourcing applications, building upon the model in [8], we consider a scenario where some workers are not reliable, i.e., their reply to the query is not correct, each with a certain probability independent of others. We show that the LDGM codes presented in this paper in concatenation with LDPC codes can be used as query schemes where the number of queries approaches information theoretic lower bound and the number of items in each query is polylogarithmic in the number of items.

The organization of this paper is as follows. Section II provides the necessary background. Section III contains the sparsity results for both polar codes with general kernels and the proposed new construction of LDGM codes. Section IV considers the query schemes for crowdsourced labelling based on the concatenation of the LDGM codes with LDPC codes.

II. Preliminaries

A. Channel Polarization and Polar Codes

The *channel polarization* phenomenon was discovered by Arıkan [2] and is based on a 2×2 polarization transform as the building block. For $N = 2^n$, the polarization transform is obtained from the $N \times N$ matrix $G_2^{\otimes n}$, where $G_2 = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$ [2]. Polar codes of length N are constructed by selecting certain rows of $G_2^{\otimes n}$. More specifically, let K denote the code dimension. Then sort all the N bit-channels, resulting from the polarization transform, with respect to their probability of error, select the best K of them with the lowest probability of error, and then select the corresponding rows from $G_2^{\otimes n}$. In other words, the generator matrix of an (N, K) polar code is a $K \times N$ submatrix of $G_2^{\otimes n}$. The probability of error of this code, under successive cancellation decoding, is upper bounded by the sum of probabilities of error of the selected K best bit-channels [2]. Polar codes and polarization phenomenon have been successfully applied to a wide range of problems including data compression [18], [19], broadcast channels [20], [21], multiple access channels [22], [23], physical layer security [24], [25], and coded modulations [26].

B. General Kernels and Error Exponent

It is shown in [17] that if G_2 is replaced by an $l \times l$ polarization kernel G, then polarization still occurs if and only if G is an invertible matrix in $\mathbb{F}_2$ and none of its column permutations is upper triangular. Furthermore, the authors of [17] provided a general formula for the error exponent of polar codes constructed based on an arbitrary $l \times l$ polarization matrix G. More specifically, let $N = l^n$ denote the block length and C denote the capacity of the channel. For any $\beta < E(G)$, specified next, the rate $\frac{K}{N}$ of the polar code with probability of error P_e upper bounded by

$$P_e(n) \leqslant 2^{-N^\beta}$$

approaches C as n grows large. The rate of polarization (defined in [17, Definition 7]), $E(G)$, is given by

$$E(G) = \frac{1}{l}\sum_{i=1}^{l} \log_l D_i, \tag{1}$$

where $\{D_i\}_{i=1}^{l}$ are the partial distances of G. Formally, for $G = [g_1^T, g_2^T, \ldots, g_l^T]^T$, the partial distances D_i are defined as follows:

$$D_i \stackrel{\text{def}}{=} d_H(g_i, \text{span}(g_{i+1}, \ldots, g_l)), \qquad i = 1, 2, \ldots, l-1 \tag{2}$$

$$D_l \stackrel{\text{def}}{=} d_H(g_l, 0) = w_H(g_l), \tag{3}$$

where $d_H(a, b)$ is the Hamming distance between two vectors a and b, and $d_H(a, U)$ is the minimum distance between a vector a and a subspace U, i.e., $d_H(a, U) = \min_{u\in U} d_H(a, u)$.

III. Constructions and Main Results

The main results of this paper are stated in this section. We refer to [27] for a longer version of this paper with all the proofs.

A. Sparsity Study

Leveraging results in polar coding theory, we first show the existence of capacity achieving polar codes with generator matrices of which all column weights are polynomial in the block length N, hence validating the conjecture in [14]. Second, we show that, for any polar code, *almost* all of the column weights of the generator matrix are larger than polylogarithmic in N.

Proposition 1. *For any fixed $s > 0$, there are capacity-achieving polar codes with generator matrices having column weights upper bounded by N^s.*

Proposition 2. *Given $l \geqslant 2$ and an $l \times l$ polarizing kernel G, the ratio of columns in $G^{\otimes n}$ with $O((\log N)^r)$ Hamming weight vanishes for any $r > 0$ as n grows large.*

B. New Approach: Construction

We propose a new construction of codes with even sparser generator matrices than those given in section III-A. In particular, *almost all* the column weights of the generator matrices of such codes are logarithmic in the code block length, and there is an upper bound $w_{u.b.}$, polynomial in the logarithm of the block length, on *all* the column weights.

Formally, let $G = G_l^{\otimes n} \otimes I_{n'}$, where G_l is an $l \times l$ polarization kernel and $I_{n'}$ is the $n' \times n'$ identity matrix. The matrix has the following form:

$$G = \begin{bmatrix} G_l^{\otimes n} & \mathbf{0}_{l^n} & \mathbf{0}_{l^n} & \ldots & \mathbf{0}_{l^n} \\ \mathbf{0}_{l^n} & G_l^{\otimes n} & \mathbf{0}_{l^n} & \ldots & \mathbf{0}_{l^n} \\ \mathbf{0}_{l^n} & \mathbf{0}_{l^n} & G_l^{\otimes n} & \ldots & \mathbf{0}_{l^n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{l^n} & \mathbf{0}_{l^n} & \mathbf{0}_{l^n} & \ldots & G_l^{\otimes n} \end{bmatrix}. \tag{4}$$

Let $N = l^n$, $N' = N \times n'$ be the block length, and $K' = n'K$ be the code dimension. Then $\frac{K'}{N'} = \frac{K}{N}$ is the code rate. To construct the polar-based code, we use the K' bit-channels with the lowest probability of error and the generator matrix of an (N', K') code *based on* G is a $K' \times N'$ sub-matrix of G.

When all columns are required to have low Hamming weights, a *splitting algorithm* is applied. Given a column weight threshold $w_{u.b.}$, the splitting algorithm splits any column in G with weight exceeding $w_{u.b.}$ into columns that sum to the original column both in $\mathbb{F}_2$ and in $\mathbb{R}$, and that have weights no larger than $w_{u.b.}$. That is, for a column in G with weight W, if $W \leqslant w_{u.b.}$, keep the column as it is. If $W = m \cdot w_{u.b.} + r$ for some $m \in \mathbb{N}$ and some $0 \leqslant r < w_{u.b.}$, replace the column with $m + 1$ columns, such that each column has no more than $w_{u.b.}$ ones. Denote the resulting $N' \times N'(1 + R)$ matrix by G'. A new code *based on* G' selects the same K' rows as the code based on G to form the generator matrix, whose column weights are uniformly bounded by $w_{u.b.}$.

We give a toy example for the splitting algorithm: assume the threshold $w_{u.b.}$ is chosen to be 1, and the first column of an N-column matrix G is $(1, 1, 0, \ldots, 0)^T$. Then this column will be split into two new columns, $(1, 0, 0, \ldots, 0)^T$ and $(0, 1, 0, \ldots, 0)^T$, called v_1' and v_1'' here. If all the other columns of G have weights 0 or 1, then resulting G' will be

$$G' = [v_1', v_1'', v_2, \ldots, v_N],$$

where v_i denotes the ith column of G.

C. New Approach: Analysis of Error Probability

First, we show that, for an appropriate choice of n', codes based on G have vanishing probability of error as n grows large. Let $\beta < E(G_l)$ be given, there are polar codes with kernel G_l such that the probability of error is bounded by 2^{-N^β}. For the code based on G, the probability of error is bounded above, through union bound, by $n' \cdot 2^{-N^\beta}$. Throughout this paper, we choose

$$n' = 2^{N^{(1-\delta)E(G_l)}}, \tag{5}$$

for an arbitrarily small constant $\delta > 0$. We then have the following lemma.

Lemma 3. *Let G be as in* (4) *and n' be as in* (5)*. Then for any $\beta < E(G_l)$, the rate of the code based on G with the probability of error upper bounded by 2^{-N^β} approaches C as n grows large.*

When the splitting algorithm is applied, we show in the following proposition that the probabilities of error of the code based on G' and G can be bounded in the same way.

Proposition 4. *For any $\beta < E(G_l)$, there is a decoding scheme based on successive cancellation(SC) decoding such that the probability of error of the code based on G', with dimension $K' < N'C$, can be bounded by 2^{-N^β} for sufficiently large n.*

The block length of the code based on G is

$$N' = n'N = 2^{N^{(1-\delta)E(G_l)}} N. \tag{6}$$

We use $log(N')$ as *sparsity benchmark* in this paper, which can be bounded by

$$\begin{aligned} N^{E(G_l)} \geqslant log(N') &= N^{(1-\delta)E(G_l)} + \log N \\ &= N^{(1-\delta)E(G_l)+o(1)} \geqslant N^{(1-\delta)E(G_l)}, \end{aligned} \tag{7}$$

for sufficiently large n.

D. Geometric Mean and Maximum Column Weight

The column weights of G compared to $log(N')$ can be analyzed in two scenarios: (1) geometric mean column weight, and (2) maximum column weight.

Definition 1. *For a binary matrix G with m columns, whose weights are denoted by $w_1, w_2, \ldots, w_m$, the geometric mean column weight $w_{GM}(G)$ and the maximum column weight $w_{max}(G)$ are defined as follows:*

$$w_{GM}(G) \stackrel{\text{def}}{=} (w_1 \times w_2 \times \ldots \times w_l)^{\frac{1}{l}}, \tag{8}$$

$$w_{max}(G) \stackrel{\text{def}}{=} \max_i w_i. \tag{9}$$

Let $w_1, w_2, \ldots, w_l$ denote the column weights of the $l \times l$ binary matrix G_l. The geometric mean column weight of $G = G_l^{\otimes n} \otimes I_{n'}$ equals to that of $G_l^{\otimes n}$, which is denoted by $w_{GM}(n, G_l)$ and defined as follows:

$$w_{GM}(n, G_l) \stackrel{\text{def}}{=} w_{GM}(G_l^{\otimes n} \otimes I_{n'}). \tag{10}$$

The maximum column weight of G is the same as that of $G_l^{\otimes n}$, which is denoted by $w_{max}(n, G_l)$ and defined as follows:

$$w_{max}(n, G_l) \stackrel{\text{def}}{=} w_{max}(G). \tag{11}$$

Note that $w_{GM}(n, G_l) = [(w_1 \times w_2 \times \ldots \times w_l)^{\frac{1}{l}}]^n = w_{GM}(G_l)^n$. Also, $w_{max}(n, G_l) = (max_i(w_i))^n \leqslant l^n$.

E. Sparsity with Kernel G_2

Let $G = G_2^{\otimes n} \otimes I_{n'}$ with n' chosen as in (5). We show two things in this subsection: $w_{GM}(n, G_2) \approx \log N'$ and, after careful splitting we get a matrix G' such that $w_{max}(G') \leqslant (\log N')^{1+2\epsilon^*}$ for a constant $\epsilon^* \approx 0.085$ with vanishing loss of rate compared to G.

Proposition 5. *There is a sequence of capacity achieving codes over any BMS channel with the geometric mean column weight almost logarithmic in the block length. More specifically, for any fixed $\delta' > 0$, n' in* (5) *can be chosen such that*

$$w_{GM}(n, G_2) = [\log(N')]^{1+\delta'+o(1)} \tag{12}$$

for sufficiently large n.

By the central limit theorem, the column weights concentrate around the geometric mean column weight, the ratio of columns with weights exceeding $[\log(N')]^{1+\delta''+o(1)}$ is vanishing as n grows large for any $\delta'' > \delta'$.

Although the geometric mean column weight of G and the weights of most columns are almost logarithmic in N', the maximum column weight is $w_{max}(G) = 2^n = [w_{GM}(G)]^2$ and is approximately $(\log N')^2$. However, we show next that a matrix G' can be obtained from the splitting algorithm such that all column weights are below some threshold $w_{u.b.}$ which would be much smaller than $w_{max}(G)$.

Since polar codes and the code based on G are capacity-achieving, as shown in lemma 3, and that the code rates of the codes based on G and G' differ by a ratio $1 + R$, the latter is capacity achieving if R vanishes as n grows large. In the following, we will explore appropriate choices of the column weight threshold for the splitting algorithm that allow the value R goes to 0 exponentially fast.

Let $\epsilon > 0$ be given and

$$w_{u.b.} = (\log N')^{1+\epsilon} = N'^{\frac{1}{2}+\epsilon'}, \tag{13}$$

be the upper bound for the column weights, where

$$\epsilon' = (1+\epsilon)(\frac{1-\delta}{2} + o(1)) - \frac{1}{2}, \tag{14}$$

for large n. To estimate the multiplicative rate loss of $1 + R$, we may study the effect on $G_2^{\otimes n}$, since that is equivalent to the overall effect on G.

First note that R is the ratio of the number of extra columns resulting from the splitting algorithm to the number of columns N of $G_2^{\otimes n}$. Let $w_1, w_2, \ldots, w_N$ denote the column weights of $G_2^{\otimes n}$. R can be characterized as follows:

$$R = \frac{1}{N}\sum_{k=1}^{k_{max}} |\{i : kw_{u.b.} \leqslant w_i < (k+1)w_{u.b.}\}| \times k, \tag{15}$$

where $k_{max} = \lfloor 2^n / w_{u.b.} \rfloor$.

In fact, with G_2 as the polarization kernel, each w_i is an integer power of 2. By grouping the k_{max} terms in (15), the ratio R can be expressed as a sum of $\log k_{max}$ terms. Let $\lambda(x,y) \stackrel{\text{def}}{=} -D(\frac{1}{2}+x+y||\frac{1}{2}) + y$ for $x, y \geqslant 0$ and $x + y \leqslant \frac{1}{2}$, where $D(p_1||p_2)$ is the Kullback–Leibler divergence between two distributions Ber(p_1) and Ber(p_2). We characterize the asymptotic behaviour of the rate R in the following theorem.

Theorem 6. *For $G = G_2^{\otimes n} \otimes I_{n'}$, where $n', N', w_{u.b.}$, and ϵ' are given by* (5)*,* (6)*,* (13)*, and* (14)*, apply the splitting algorithm to form a matrix $G' \in \{0,1\}^{N' \times N'(1+R)}$ such that $w_{max}(G') \leqslant w_{u.b.}$. Then R has the following asymptotic expression:*

$$R \doteq \begin{cases} 2^{n(\epsilon^* - \epsilon')} & \to 0, \quad \text{if } \epsilon' > \epsilon^* \\ 2^{\lambda(\epsilon', \alpha_{max})} & \to \infty, \quad \text{if } \epsilon' < \epsilon^* \end{cases}, \tag{16}$$

where $\epsilon^* \stackrel{\text{def}}{=} \frac{1}{6} - \frac{5}{3}\log\frac{4}{3} \approx 0.085$, $\alpha_{max} = \max_i \alpha_i$, and $a_n \doteq b_n$ means that $\frac{1}{n}\log\frac{a_n}{b_n} \to 0$ as $n \to \infty$.

We can express the conditions in (16) in terms of the relation between ϵ and ϵ^* leading to the following corollary.

Corollary 7. *Let $n', N', \epsilon', \epsilon^*$, $w_{u.b.}$, and α_{max} be as in theorem 6. Then*

$$R \doteq \begin{cases} 2^{n(\epsilon^* - \epsilon')} & \to 0, \quad \text{if } \epsilon > 2\epsilon^* \\ 2^{\lambda(\epsilon', \alpha_{max})} & \to \infty, \quad \text{if } \epsilon < 2\epsilon^* \end{cases}.$$

The rate loss $1 + R$ of the code based on G' to the code based on G can thus be made arbitrarily close to 1 when the column weight upper bound $w_{u.b.}$ is properly chosen. Combining results in subsection III-C and the corollary 7, we have the following corollary:

Corollary 8. *Let $\beta < E(G_2) = 0.5$ and $\epsilon > 2\epsilon^*$ be given. Then there exists a sequence of codes based on G', generated by applying the splitting algorithm to $G = G_2^{\otimes n} \otimes I_{n'}$, with the following properties:*

1) *The error probability is upper bounded by 2^{-N^β}.*
2) *The Hamming weight of each column of the generator matrix is upper bounded by $w_{u.b.} = (\log N')^{1+\epsilon}$.*
3) *The rate approaches C as n grows large.*

F. Sparsity with General Kernels

In this subsection we consider $l \times l$ kernels G_l with $l > 2$ and show the existence of G_l with $w_{GM}(n, G_l) = O((\log N')^\lambda)$ for some $\lambda < 1$. However, we do not bound $w_{max}(.,.)$ as in the case with the G_2 kernel. To characterize the geometric mean column weight and the maximum column weight, the *sparsity order* is defined as follows:

Definition 2. *The sparsity order of the geometric mean column weight is*

$$\lambda_{GM}(n, G_l) \stackrel{\text{def}}{=} \log_{\log(N')} w_{GM}(n, G_l) = \frac{\log w_{GM}(n, G_l)}{\log\log(N')}, \tag{17}$$

where n' and N' are defined in (5) *and* (6)*, respectively.*

Definition 3. *the sparsity order of the maximum column weight*

$$\lambda_{max}(n, G_l) \stackrel{\text{def}}{=} \log_{\log(N')} w_{max}(n, G_l) = \frac{\log w_{max}(n, G_l)}{\log\log(N')}. \tag{18}$$

For example, if $w_{GM}(n, G_l)$ (or $w_{max}(n, G_l)$) can be expressed in the Landau notations as $\Theta([\log N']^r)$, then $\lambda_{GM}(n, G_l)$ (or $\lambda_{max}(n, G_l)$) goes to r as n grows large.

We give Table I[1] for

$$G_3^* = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}, G_4^* = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix},$$

and G_{16}^* (the smallest l with $E_l > 0.5$; see [17] for explicit construction), the kernels achieving E_3, E_4 and E_{16}, the maximal error exponents for $l = 3, 4, 16$, respectively.

[1]The limits of the sparsity orders when $n \to \infty$ are shown, hence $o(1)$ terms are neglected.

Table I
λ_{GM} AND λ_{max} FOR G_2, G_3^*, G_4^* AND G_{16}^* AS $n \to \infty$

	$E(G_l)$	$\lambda_{GM}(n, G_l)$	$\lambda_{max}(n, G_l)$
G_2	0.5	$1+\delta'$	$2(1+\delta')$
G_3^*	$\frac{2}{3}\log_3 2 \approx 0.42$	$1+\delta'$	$1.5(1+\delta')$
G_4^*	0.5	$\approx 1.15(1+\delta')$	$\log 3(1+\delta')$
G_{16}^*	≈ 0.5183	$\approx 1.443(1+\delta')$	omitted

Table II
λ_{GM} AND λ_{max} FOR G_3' AND G_4' AS $n \to \infty$

	$E(G_l)$	$\lambda_{GM}(n, G_l)$	$\lambda_{max}(n, G_l)$
G_3'	$\frac{2}{3}\log_3 2 \approx 0.42$	$\approx 0.79(1+\delta')$	$\approx 2.38(1+\delta')$
G_4'	$\frac{3}{8} = 0.375$	$\frac{2}{3}(1+\delta')$	$\frac{8}{3}(1+\delta')$

However, the error exponent is not the only factor that determines the sparsity orders. For example, for $l = 3$ and $l = 4$, the matrices

$$G_3' = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}, G_4' = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix},$$

instead of G_3^* and G_4^*, have the smallest sparsity orders of the geometric mean column weight (found through exhaustive search), as shown in table II. By central limit theorem, most column weights have similar orders over the logarithm of the block length. Therefore, if sparsity constraint is only required for almost all of the columns of the generator matrix, G_3' and G_4' are the more preferable polarization kernels over G_3^* and G_4^*, respectively.

For a given G_l, we may relate the two terms $E(G_l)$ and $w_{GM}(n)$, or, more specifically, the partial distances $D_1, \ldots, D_l$ and the column weights $w_1, \ldots, w_l$ as follows.

Lemma 9. *The ratio of $\lambda_{GM}(n, G_l)$ to $\frac{\sum_{i=1}^{l} \log_l w_i}{\sum_{i=1}^{l} \log_l D_i}$ lies between 1 and $\frac{1}{1-\delta}$ for sufficiently large n.*

The following theorem shows that an arbitrarily small order can be achieved with a large l and some G_l.

Theorem 10. *For any fixed constant $0 < r \leqslant 1$, there exist an $l \times l$ polarizing kernel G_l, where $l = l(r, \delta)$, such that $\lambda_{GM}(n, G_l) < r$ for sufficiently large n.*

Let $r < 1$ and $\eta > 0$ be fixed. For a proper choice of G_l with $\lambda_{GM}(n, G_l) < r$, concentration of the column weights, i.e., the central limit theorem, implies only vanishing fraction of columns in G have weight larger than $[\log N']^{(1+\eta)r}$.

IV. Application to Crowdsourcing

A. Recap of Coding for Crowdsourced Label Learning

The problem model considered in [8] is the following. There are n items, each of which is associated with a binary label X_i unknown to a taskmaster and X_i is i.i.d. $\sim \text{Ber}(p), \forall i$.

Let $H_b(\cdot)$ denote the binary entropy function. From [8], when workers in the crowd are perfect, there exists a XOR-querying scheme using

$$m = n[H_b(p) + \zeta(1 - H_b(p))]$$

queries, each involving no more than $(H_b(p)^{-1} - 1)\frac{K_1 - K_2 ln(\zeta)}{1-\zeta}$ items for some $\zeta \in (0, 1)$, that achieves perfect recovery.

In the case where queries are not responded, each with a probability r independent of others, the number of queries is lower bounded by $m_{BER} = n(H_b(p))/(1 - r)$ [8]. Also, existence of a XOR-querying scheme with

$$m = n[H_b(p) + \zeta(1 - H_b(p))]/(1 - r)$$

queries, each with $O(\log \frac{1}{\zeta} \log n)$ items, that guarantees perfect recovery of the labels as n grows large is shown in [8].

B. BSC scenario

The case when some queries are answered incorrectly is widely observed in crowdsourced label learning in the real world [4], [28]. When the queries are answered correctly with probability $1 - q$ for some $q \in [0, 0.5)$, referred to here as the $BSC(q)$ model, the information-theoretic lower bound on the number of queries is

$$m_{BSC}(n, p, q) = \frac{nH_b(p)}{1 - H_b(q)}.$$

We can apply corollary 8 to design a query scheme with number of queries, m', arbitrarily close to m_{BSC} and small number of items in each query.

Theorem 11. *For the $BSC(q)$ model, for any $\zeta \in (0, 1)$ and $\epsilon > 2\epsilon^*$, there is a query scheme using*

$$m' = (1 + o(1))\frac{H_b(p) + \zeta(1 - H_b(p))}{1 - H_b(q)} \tag{19}$$

queries, each involving no more than $O(\log \frac{1}{\zeta}[\log n]^{1+\epsilon})$ items, that achieves perfect recovery.

References

[1] R. Gallager, "Low-density parity-check codes," *IRE Transactions on information theory*, vol. 8, no. 1, pp. 21–28, 1962.

[2] E. Arikan, "Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels," *IEEE Trans. Inf. Theory*, vol. 55, no. 7, pp. 3051–3073, 2009.

[3] D. R. Karger, S. Oh, and D. Shah, "Iterative learning for reliable crowdsourcing systems," in *Advances in neural information processing systems*, 2011, pp. 1953–1961.

[4] A. Vempaty, L. R. Varshney, and P. K. Varshney, "Reliable crowdsourcing for multi-class labeling using coding theory," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 4, pp. 667–679, 2014.

[5] A. G. Dimakis, P. B. Godfrey, Y. Wu, M. J. Wainwright, and K. Ramchandran, "Network coding for distributed storage systems," *IEEE transactions on information theory*, vol. 56, no. 9, pp. 4539–4551, 2010.

[6] K. Lee, M. Lam, R. Pedarsani, D. Papailiopoulos, and K. Ramchandran, "Speeding up distributed machine learning using codes," *IEEE Transactions on Information Theory*, vol. 64, no. 3, pp. 1514–1529, 2018.

[7] A. Mazumdar and S. Pal, "Semisupervised clustering, and-queries and locally encodable source coding," in *Advances in Neural Information Processing Systems*, 2017, pp. 6489–6499.

[8] C.-J. Pang, H. Mahdavifar, and S. S. Pradhan, "Coding for crowdsourced classification with xor queries," *Proceedings of IEEE Information Theory Workshop (ITW)*, 2019.

[9] D. J. MacKay, "Good error-correcting codes based on very sparse matrices," *IEEE transactions on Information Theory*, vol. 45, no. 2, pp. 399–431, 1999.
[10] D. J. MacKay and R. M. Neal, "Good codes based on very sparse matrices," in *IMA International Conference on Cryptography and Coding*. Springer, 1995, pp. 100–111.
[11] W. Zhong, H. Chai, and J. Garcia-Frias, "Approaching the shannon limit through parallel concatenation of regular LDGM codes," in *Proceedings. International Symposium on Information Theory, 2005. ISIT 2005.* IEEE, 2005, pp. 1753–1757.
[12] J. Garcia-Frias and W. Zhong, "Approaching shannon performance by iterative decoding of linear codes with low-density generator matrix," *IEEE Communications Letters*, vol. 7, no. 6, pp. 266–268, 2003.
[13] W. Zhong and J. Garcia-Frias, "LDGM codes for channel coding and joint source-channel coding of correlated sources," *EURASIP Journal on Applied Signal Processing*, vol. 2005, pp. 942–953, 2005.
[14] A. M. Kakhaki, H. K. Abadi, P. Pad, H. Saeedi, K. Alishahi, and F. Marvasti, "Capacity achieving random sparse linear codes," *Preprint*, 2011.
[15] H. Mahdavifar, "Scaling exponent of sparse random linear codes over binary erasure channels," in *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2017, pp. 689–693.
[16] W. Lin, S. Cai, B. Wei, and X. Ma, "Coding theorem for systematic LDGM codes under list decoding," in *2018 IEEE Information Theory Workshop (ITW)*. IEEE, 2018, pp. 1–5.
[17] S. B. Korada, E. Sasoglu, and R. Urbanke, "Polar codes: Characterization of exponent, bounds, and constructions," *IEEE Transactions on Information Theory*, vol. 56, no. 12, pp. 6253–6264, 2010.
[18] E. Arıkan, "Source polarization," *Proceedings of IEEE International Symposium on Information Theory (ISIT)*, pp. 899–903, 2010.
[19] E. Abbe, "Polarization and randomness extraction," *Proceedings of IEEE International Symposium on Information Theory (ISIT)*, pp. 184–188, 2011.
[20] M. Mondelli, S. H. Hassani, I. Sason, and R. L. Urbanke, "Achieving Marton's region for broadcast channels using polar codes," *IEEE Transactions on Information Theory*, vol. 61, no. 2, pp. 783–800, 2015.
[21] N. Goela, E. Abbe, and M. Gastpar, "Polar codes for broadcast channels," *IEEE Transactions on Information Theory*, vol. 61, no. 2, pp. 758–782, 2015.
[22] E. Şaşoğlu, E. Telatar, and E. Yeh, "Polar codes for the two-user binary-input multiple-access channel," *IEEE Transactions on Information Theory*, vol. 59, no. 10, pp. 6583–6592, 2013.
[23] H. Mahdavifar, M. El-Khamy, J. Lee, and I. Kang, "Achieving the uniform rate region of general multiple access channels by polar coding," *IEEE Transactions on Communications*, vol. 64, no. 2, pp. 467–478, 2016.
[24] H. Mahdavifar and A. Vardy, "Achieving the secrecy capacity of wiretap channels using polar codes," *IEEE Transactions on Information Theory*, vol. 57, no. 10, pp. 6428–6443, 2011.
[25] M. Andersson, V. Rathi, R. Thobaben, J. Kliewer, and M. Skoglund, "Nested polar codes for wiretap and relay channels," *IEEE Communications Letters*, vol. 14, no. 8, pp. 752–754, 2010.
[26] H. Mahdavifar, M. El-Khamy, J. Lee, and I. Kang, "Polar coding for bit-interleaved coded modulation," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 5, pp. 3115–3127, 2015.
[27] J. C.-J. Pang, H. Mahdavifar, and S. S. Pradhan, "Capacity-achieving polar-based ldgm codes with crowdsourcing applications," 2020.
[28] D. R. Karger, S. Oh, and D. Shah, "Budget-optimal task allocation for reliable crowdsourcing systems," *Operations Research*, vol. 62, no. 1, pp. 1–24, 2014.