Coding Sets with Asymmetric Information

Alexandr Andoni[1], Javad Ghaderi[2], Daniel Hsu[1], Dan Rubenstein[1], Omri Weinstein[1]

[1]Department of Computer Science, Columbia University

[2]Department of Electrical Engineering, Columbia University

Abstract

We study the following one-way asymmetric transmission problem, also a variant of model-based compressed sensing: A resource-limited encoder has to report a small set S from a universe of N items to a more powerful decoder (server). The distinguishing feature is *asymmetric information*: the subset S is comprised of i.i.d. samples from a prior distribution μ, and μ is only known to the *decoder*. The goal for the encoder is to encode S *obliviously*, while achieving the information-theoretic bound of $\approx |S| \cdot H(\mu)$, i.e., the Shannon entropy bound.

We first show that any such compression scheme must be *randomized*, if it gains non-trivially from the prior μ. This stands in contrast to the symmetric case (when both the encoder and decoder know μ), where the Huffman code provides a near-optimal *deterministic* solution. On the other hand, a rather simple argument shows that, when $|S| = k$, a *random* linear code achieves near-optimal communication rate of about $k \cdot H(\mu)$ bits. Alas, the resulting scheme has prohibitive *decoding time*: about $\binom{N}{k} \approx (N/k)^k$.

Our main result is a *computationally efficient* and *linear* coding scheme, which achieves an $O(\lg \lg N)$-competitive communication ratio compared to the optimal benchmark, and runs in $\mathrm{poly}(N, k)$ time. Our "multi-level" coding scheme uses a combination of hashing and syndrome-decoding of Reed-Solomon codes, and relies on viewing the (unknown) prior μ as a rather small convex combination of uniform ("flat") distributions.

Index Terms

Coding Theory, Algorithmic Information Theory, Compression, IoT

I. Introduction

We study the problem of coding a set with *asymmetric* information, defined as follows. There is a universe $[N] := \{1, 2, \ldots, N\}$ of N items, and the encoder's task is to transmit a subset $S \subset [N]$ using an m-bit message so that a decoder can reconstruct the set S efficiently. In our setup, the decoder has a *prior distribution* σ over the sets S that may be sent, which is not available to the encoder. The main goal is to design compression schemes that (1) obtain communication rate as close as possible to the information-theoretic minimum, namely the (Shannon) entropy bound with respect to the distribution σ, and (2) are computationally efficient.

This problem is the one-way communication version of the asymmetric transmission problem [1], as well as a type of model-based compressed sensing. While we expand on these a little below, for now we note that the standard asymmetric transmission problem is two-way, with the decoder sending much more information to the

Emails: {andoni@cs,jghaderi@ee,djhsu@cs,danr@cs,omri@cs}.columbia.edu

encoder. Here we seek to eliminate this inefficiency, in the setting of communicating a set S. One can envision many scenarios where it is imperative to eliminate an expensive down-link from decoder to encoder; we give one such scenario for designing very light communication protocols for tracking ultra-low-power devices in Internet-of-Things environments. Here, a common task is for a set of such devices to communicate their identities to a router (e.g., an entry point of a physical region) [2], [3], [4]. Since the devices are low power, the main goal is to minimize their total communication costs. The communication can be further improved using some side information, in particular a prior distribution on which devices are more likely to be present (i.e., which sets are more likely to be sent). However, the side information is typically asymmetric: the prior is specific to the decoding router, or uses statistics that are not known to or are too expensive to maintain by the devices (see the discussion in [1] or [5]).

In addition to the natural goal of communication efficiency, a common requirement for such coding schemes is also to have a *computationally efficient* decoding procedure. Our goal here is for the decoding time to be polynomial/linear in N (which is the best we can hope for without further assumptions — the input to the decoder is the distribution σ, of potentially $\Omega(N)$ description size)[1].

Without further assumptions on the distribution σ, this problem does not admit any viable solutions: both communication and computation are essentially doomed. Indeed, [1], [5] show that the trivial bound of $\sim N$ communication is required, even when the entropy of σ is much smaller. We note that [1] circumvented this barrier by allowing two-way communication where the decoder can send much larger messages back to encoder, whereas we focus on purely *one-way* protocols only. As for the distributional setting, a generic (non-product) prior distribution σ has a high description complexity (exponential in N, or max set size), thus dooming the time-efficiency of any decoding scheme.

In this paper, we consider the most natural class of priors σ of i.i.d. items: the sets $S \sim \mu^k$ are comprised of k items, each drawn independently from some distribution μ over $[N]$. We note that this a common assumption, implicitly assumed in (vanilla) compressed sensing, as well as classic (symmetric information) source-coding problems.

For this setting, we develop protocols that achieve efficient decoding time, and competitive communication costs. Our coding scheme is *linear*—the encoding is $C \cdot \mathbb{1}_S$ where C is the coding matrix and $\mathbb{1}_S$ is the indicator vector of the set S—which is a further desirable property of coding scheme. This property is similar to the one imposed in compressed sensing. Linearity facilitates quick and simple updates to the message in streaming/dynamic environments (e.g., in the IoT application above) as the message can be simply updated as items are added one by one to the set S.

A. Relation to Problems in Prior Literature

Our problem relates to many other problems studied previously, but, surprisingly, has not been explicitly studied. When there is *no side information*, the problem is the classic problem of coding a set S. Without requiring linearity,

[1]With further assumptions—e.g., preprocessing—one may ask for sublinear runtime, of the order of $\mathrm{poly}(|S|, \lg N)$, as was accomplished in some compressed sensing literature; see, e.g., [6], [7].

a trivial solution is to append the indices of items in S, yielding communication $k \lg N$ for sets S of size k.[2] If we further require linearity, then the problem becomes a variant of compressed sensing. A slight caveat is that the compressed sensing schemes usually work over reals [8], [9], and the vector $C \cdot \mathbb{1}_S$ is a real vector, which raises the issue of rounding and real number representation. Nevertheless, it is possible to do compressed sensing over the $\mathbb{F}_2$ field; see, e.g., [10], [11], [12], [13].

Another related model is source coding, where both the encoder and the decoder have access to some prior distribution μ, and the set S is composed of k items i.i.d. items drawn from μ. Then a (near-)optimal solution can be obtained via, say, *Huffman coding* [14]. The length of the compression of a set S is $\sum_{i \in S} \lceil \lg 1/\mu(i) \rceil$, which, in expectation, is upper bounded by $k \cdot H(\mu) + k$, close to the information-theoretic optimum of $k \cdot H(\mu)$ (up to the rounding issues).

When the side information is not known to the encoder (as it is in our case), the problem becomes the classic asymmetric transmission problem [1], [15], [16], [17], [5] (see also [18]). In this problem, the encoder generates an item from a probability distribution μ and needs to communicate its identity to the router/server (decoder). The goal is again to reach the information capacity of $\approx H(\mu)$. While there are protocols that achieve such capacity, the protocols require *two-way* communication—the backchannel from the decoder to the encoder is on the order of $\Omega(\lg N)$ bits. Furthermore, this is necessary: [1] shows that either the encoder or decoder has to communicate the trivial $\Omega(\lg N)$ bits [1] (see also the follow-up work of [5] for a lower bound on the number of interactive rounds required).

In contrast, our protocols use one-way communication only. We circumvent the above lower bound by exploiting the fact that the encoder sends a *set* S of items, instead of a single one, with a randomized protocol. In particular, we can amortize the lower bound of $\Omega(\lg N)$ against $|S|$ items. In other words, in our setting, we encode a set S using $m \geq \lg N$ bits, with the goal of achieving $m \ll O(|S| \cdot \lg N)$ where possible.

Finally, we remark that the problem also falls under the umbrella of model-based compressed sensing, where one generally assumes some prior knowledge on *the possible structure* (model) of the set S (beyond, say, an upper bound on its size); see, e.g., [19]. While the asymmetry is typically not an explicit goal, the encoding schemes are usually agnostic to this prior knowledge (e.g., the coding uses the usual matrix with random Gaussian entries), and hence, in fact, constitute an asymmetric coding scheme.

B. Formal Problem Setup

There are a few ways to formalize our problem, and hence we introduce three related definitions below, of growing generality. As before, there is a universe $[N] := \{1, 2, \ldots, N\}$ of items. For a given set $S \subseteq [N]$, the encoder $\mathsf{Enc}\colon 2^{[N]} \to \{0,1\}^m$ must construct a (possibly randomized) message $y := \mathsf{Enc}(S)$ of at most m bits, where m is the allowed message length, fixed in advance. The decoder $\mathsf{Dec}_\star\colon \{0,1\}^m \to 2^{[N]}$, for some side-information $\star$, must produce a set $\hat{S} := \mathsf{Dec}_\star(y)$ from the message y such that $\hat{S} = S$ with, say, at least $1 - \delta$ probability, where δ is the error probability parameter (think $\delta = 0.1$). Note that, when the side information $\star$ is null, this

[2]We use lg to denote base-2 logarithm.

task is generally impossible unless $m \geq \lg 2^N = N$. Note that the encoder's message does not depend on the side information, i.e., the encoding function $\mathsf{Enc}(S)$ is *oblivious* (in the information theory literature this is referred to as *universal compression* [20], [21]; see also Section I-E).

To measure the optimality of a coding scheme, we compare our message lengths to the information-theoretic minimum, which we denote by the parameter m^* (which is a function of $\star$). In particular, for $\alpha \geq 1$, a coding scheme is called α-*competitive* if it uses m bits while the "information-theoretic optimal" is $m^* \geq m/\alpha$ bits. Note that the value of "information-theoretic optimal" is not obvious, and in fact will differ between different definition.

There are also a few ways to measure the success of a scheme. We now introduce a few related definitions of asymmetric coding in the order of generality.

Following the discussion from before, one natural way to model the side information is via a prior distribution σ on subsets of $[N]$. In particular, we assume σ is a distribution on k items, each drawn from a distribution μ on $[N]$.

Definition 1. *For $N, m, \alpha \geq 1$, a (randomized) scheme $\mathcal{A} = (\mathsf{Enc}, \mathsf{Dec})$ is* entropy-asymmetric-coding *α-competitive scheme if: for any integer k, and prior μ on $[N]$ such that $k \cdot H(\mu) \leq m/\alpha$, we have the following where the prior σ generates a set of k items drawn iid from μ:*

$$\Pr_{\mathcal{A}, S \sim \sigma}[\mathsf{Dec}_\sigma(\mathsf{Enc}(S)) = S] \geq 1 - \delta.$$

We clarify that the randomness of the encoder and decoder is via a shared random string, which is an (auxiliary) input to both Enc *and* Dec*.*

Note that $m^* = k \cdot H(\mu)$ is the lower bound on communication necessary to transmit a set S of k items drawn iid from μ. The trivial scheme would achieve a bound[3] of $k \lg N$, which can be much higher than $kH(\mu)$.

We now consider a slightly more general definition, where we do not need to fix the size k of S, but rather be "adaptive" to the number of items in the set S, in the analogy to what the Huffman coding achieves in the symmetric case.

Definition 2. *For $N, m, \alpha \geq 1$, a (randomized) scheme $\mathcal{A} = (\mathsf{Enc}, \mathsf{Dec})$ is said to be a* Huffman-asymmetric-coding *α-competitive scheme if: for any distribution μ over $[N]$, if the set S satisfies*

$$\sum_{i \in S} \lg 1/\mu(i) \leq m^*, \tag{1}$$

where $m^* = m/\alpha$, then

$$\Pr_{\mathcal{A}}[\mathsf{Dec}_\mu(\mathsf{Enc}(S)) = S] \geq 1 - \delta.$$

In particular, a Huffman-asymmetric-coding 1-competitive scheme matches the performance of the aforementioned Huffman coding (where the encoder knows the prior μ), for $\delta = 0$ (deterministically). We also note that Eqn. (1) (with $\alpha = 1$) is the tightest condition we can require in order for a set S to be decodable with a classic Huffman

[3]The more precise bound is $\lg \binom{N}{k} \approx k \lg N/k$, but since we think of $k \ll N$, this amounts to a negligible difference.

code. Hence, the above definition asks to match the efficiency of the Huffman code (symmetric information setting) in the *asymmetric* setting, up to α-factor loss in communication.

It is not hard to note that Huffman-asymmetric-coding scheme is more general than the entropy-asymmetric-coding scheme: if we pick a random set S as in Def. 1, then it satisfies Eqn. (1) (up to a small loss in communication efficiency). See Claim 5 in Appendix A.

Finally, we give the most general definition, which is the most natural from an algorithmic perspective, but is less operational than the two above. It stems from the observation than any desirable encoding/decoding scheme is (implicitly) specifying a *list* (ordered set) $L \subseteq 2^{[N]}$ of subsets $S \subseteq [N]$ that are decoded correctly. It is immediate to see that any such list L can have at most 2^m such sets. In the presence of a prior distribution σ, one could take these sets to be the "most likely" in σ (with ties broken arbitrarily).

Definition 3. *For $N, m, \alpha \geq 1$, a (randomized) scheme $\mathcal{A} = (\mathsf{Enc}, \mathsf{Dec})$ is said to be a* list-asymmetric-coding *α-competitive scheme if: for any list L of sets $S \subseteq [N]$, where $|L| \leq 2^{m/\alpha}$, and any $S \in L$, we have that:*

$$\Pr_{\mathcal{A}}[\mathsf{Dec}_L(\mathsf{Enc}(S)) = S] \geq 1 - \delta.$$

Again, the latter definition is more general than both the definitions. In particular, a list-asymmetric-coding scheme is also a Huffman-asymmetric-coding scheme: given a prior μ, just fix the list L to be the sets satisfying condition (1). It is easy to see that the size of the list will be $\leq e2^{m/\alpha}$ (which results in just an additive $\lg e$ additive loss in communication); see details in Claim 2 in Appendix A.

The last definition has the major downside that one has to specify a list L to the decoder, which is exponential in m, thus affecting the computational efficiency of a coding scheme. Therefore, for algorithmic efficiency, it is more natural to work with the Huffman-asymmetric-coding definition, which is the focus here.

C. Our Results

First, we establish that any asymmetric-coding scheme must be randomized if it is to non-trivially exploit the prior μ or list L. In particular, if $\delta = 0$ (i.e., no randomization), then, there exists some priors where the optimal communication in the symmetric case is $m^* = O(|S| \cdot \lg |S|)$, but any asymmetric-coding scheme must have $m \approx \Theta(|S| \cdot \lg N)$. See details in Section IV.

Second, as a warm-up, we show a simple scheme that solves the most general definition, of list-asymmetric-coding scheme, but which is not computationally efficient.

Theorem 1 (Information-theoretic; see Section II)**.** *Fix error probability $\delta > 0$. There is an α-competitive list-asymmetric-coding scheme with $\alpha = \frac{m}{m - \lg 1/\delta} = 1 + o(1)$, while achieving error probability of δ.*

The scheme is a standard one: a random linear code. In particular, pick a random $C \in M_{m \times N}(\mathbb{F}_2)$, and set $\mathsf{Enc}(S) = C \cdot \mathbb{1}_S$ (all computations are done in $\mathbb{F}_2$). The decoder $\mathsf{Dec}(y)$ is the "maximum likelihood" decoder: for a given list L, go over the list in order and output the first set $\hat{S} \in L$ such that $C\mathbb{1}_{\hat{S}} = y$. See Section II for further details and proofs.

While the above scheme achieves the information-theoretic bound (up to additive $\lg 1/\delta$), it is *not computationally-efficient* and requires runtime of about $\Omega(2^m)$. Even when the list L is somehow more efficiently represented (e.g., all sets S that satisfy the Huffman condition Eqn. (1)), the problem appears computationally hard. In particular, it is a variant of the classic problem of decoding random linear codes. Obtaining a coding scheme with faster decoding is precisely the focal point of our work:

> **Main goal:** *Develop computationally efficient oblivious compression schemes, that have only poly(N) encoding/decoding time, at the expense of a (mild, multiplicative) overhead in communication cost compared to random codes (α-competitive).*

Our main result is the design of a *computationally-efficient*, Huffman-asymmetric-coding scheme which is optimal up to a $O(\log\log N)$-factor loss in the message length.

Theorem 2 (Main; see Section III)**.** *Fix target message length $m > \lg N + 4$, and error probability $\delta \geq 1/\lg N$. There is a linear Huffman-asymmetric-coding scheme, which is $O(\log\log N)$-competitive, and has poly(N) decoding time and error probability of δ.*

D. Technical Overview of Theorem 2

The proof of Theorem 2 is based on a "multi-level" coding scheme. The basic building block of our "multi-level" coding scheme is the *uniform* compressed sensing scheme of [13], which is the finite-alphabet equivalent of standard compressed sensing schemes (with a "uniform" prior). In particular, their scheme is a computationally efficient linear sparse recovery scheme for k-sparse vectors in $\mathbb{F}_2^N$, using $O(k \log N)$ bits. Their (deterministic) scheme relies on *syndrome decoding* of linear codes, which allows to decode in polynomial time any k-sparse vector $x \in \mathbb{F}_2^N$, using the *parity check* matrix C_{RS} of Reed-Solomon codes with the appropriate rate/dimension generated by a binary symmetric (BSC) channel (see Section III-A for details).

Recall that in our setup, the prior μ is nonuniform and *unknown* to the encoder. We view the ground set of $[N]$ items as being partitioned into T buckets of doubly-exponentially decaying probabilities w.r.t. μ, where bucket B_i contains all elements with probability between $2^{2^{-i}}$ and $2^{-2^{i+1}}$ w.r.t. μ. This allows us to set T to be *doubly-logarithmic*, i.e., $T = O(\lg\lg N)$.

The encoder sends T concatenated messages, where the goal of the i^{th} message is to allow the decoder to decode the subset $S \cap B_i$, where $S \sim \mu^k$ is the input set at the encoder. For each "level" i, the encoder uses an appropriately-sized sensing matrix $C_{RS}^{(i)}$, whose dimensions are determined by the (worst-case) number of elements that could be encoded from B_i (here we implicitly assume that μ is uniform on B_i, which may lose a factor of ≤ 2 w.r.t the optimal message size per item, since the *encoding lengths* of items in B_i are within a factor 2). Since in the ith step we only need to distinguish items in B_i, the encoder first *hashes* the set S to the minimal universe $N_i \ll N$ that still ensures collision-freeness in B_i (using a *public* hash function shared by the encoder and the decoder), and $C_{RS}^{(i)}$ is applied to the *hashed* vector in the reduced universe. This carefully-chosen universe-reduction "preprocessing" step is essential to save on communication—e.g., using [13] on k items will cost us only $\sim k \log N_i \ll k \log N$. Note that, the encoder doesn't actually know the items B_i, and hence we don't know the

items $S \cap B_i$ to be encoded in the level i either. Instead, the level i encoding will contain all items S (this is precisely where we lose the $O(\log \log N)$-factor in communication overall), and the identification of the set $S \cap B_i$ is done at decoding time only, as described next.

Our decoding procedure is *adaptive* and runs in T successive steps. In the i^{th} step, we assume we've already successfully decoded items $S \cap B_{<i} = S \cap (B_1 \cup B_2 \cup \ldots B_{i-1})$. The decoder then "peels off" the encoding of $S \cap B_{<i}$ from the original message that it has received. This step crucially uses the *linearity* of the encoding scheme. The remaining i^{th} level message now encodes items $S \cap (B_i \cup B_{i+1} \cup \ldots B_T)$, which allows us to decode $S \cap B_i$. Note that, in addition to the aforementioned required property of no collisions inside B_i, we also need universe $[N_i]$ to be sufficiently large so that there are no collisions between items B_i and in $S \cap B_{>i}$ — otherwise we may misidentify an item from $S \cap B_{>i}$ as being in B_i. Luckily, as $|S \cap B_{>i}| \leq |S|$ is generally much smaller than $|B_i|$, this new condition on N_i does not ultimately influence the communication bound. Note that, at level i, the decoder will decode any item in B_i, and potentially identify that there exist items $S \cap B_{>i}$ (which will be left for the subsequent steps).

We present the full details of our coding scheme and its analysis in Section III.

E. Connection to Universal Compression

Finally, on a somewhat different note, noiseless compression in asymmetric scenarios was also previously studied in the information theory literature, in the context of *universal compression* (see e.g., [20], [21], [13] and references therein). This line of work exploits an elegant connection between channel coding and source coding, via *syndrome-decoding*, a connection that also plays an important role as a sub-procedure in our main result (Theorem 2, see also the discussion in Section III-A). These works exhibit (fixed-length) codes with efficient encoding and decoding procedures against a subclass of discrete *memoryless* channels (DMCs), e.g., via belief-propagation for LDPC codes [20] and Turbo codes [22]. A main technical difference of our model is that the aforementioned line of work relies on an interpretation of the set to be encoded (S) as a (sparse) additive noise vector generated by a discrete *memoryless* channel (or even further restricted symmetric channels such as the BSC), where each coordinate in $[N]$ is corrupted by the channel *independently with identical* probability. (Indeed, decoding procedures such as belief-propagation algorithms are only guaranteed to converge under specific DMC channels such as the BSC). By contrast, in our setting each coordinate in $[N]$ has a different (arbitrary and unknown to the decoder) corruption probability, hence the underlying channel is *not* memoryless.

F. Organization of the rest of the paper

The rest of the paper is organized as follows. Section II is devoted to the proof of Theorem 1 via random linear codes. In Section III, we describe the "multi-level" coding scheme and prove Theorem 2. In Section IV, we show that any asymmetric coding scheme needs to be randomized in order to gain advantage from using the side information. We end the paper with some conclusions and open problems. The connection between different notions of asymmetric-coding schemes is presented in Appendix.

II. A Basic Scheme: Random Linear Codes

We establish Theorem 1 by designing a list-asymmetric-coding scheme via a *random linear* code. It achieves essentially optimal communication (up to additive $O(1)$ bits), nearly matching the performance of the symmetric-information schemes. The runtime of this scheme is exponential in m.

Consider a randomized linear scheme where C is a uniformly random matrix $C \in \mathbb{F}_2^{m \times N}$, and $\mathsf{Enc}(S) = C \cdot \mathbb{1}_S$. The decoder for a list $L = (S_1, S_2, \ldots, S_{|L|})$ is the "maximum likelihood" decoder: given the message y, the decoder returns the *first* set S in the list L such that $\mathsf{Enc}(S) = y$:

$$\mathsf{Dec}_L^{\mathsf{ML}}(y) \;:=\; S_{\min\left\{t \in [\![|L|]\!] : \mathsf{Enc}(S_t) = y\right\}}.$$

(The random matrix C is determined using the public random bits). For brevity, we call this the *random linear scheme*.

The next lemma establishes that the random linear scheme is a list-asymmetric-coding scheme for any $\delta \in (0,1)$ and any list of at most $2^m \cdot \delta = 2^{m - \lg 1/\delta}$ subsets of $[N]$. It implies Theorem 1 since the competitiveness is $\alpha = \frac{m}{m - \lg 1/\delta}$.

Lemma 1. *Let C be a random $m \times N$ binary matrix. Then for any list L of $|L| \le 2^m$ subsets of $[N]$, and any $S \in L$:*

$$\Pr_C\left(\mathsf{Dec}_L^{\mathsf{ML}}(C \cdot \mathbb{1}_S) = S\right) \;\ge\; 1 - \big(|L| - 1\big)\, 2^{-m}.$$

Proof. For any pair of sets S, S' in the list L, we use $S \prec_L S'$ to denote that S appears before S' in L. We also let $S \triangle S' := (S \setminus S') \cup (S' \setminus S)$ denote the symmetric difference between S and S'. Finally, for $i \in [N]$ and $j \in [m]$, we let $c_i(j)$ denote the j-th entry of the code word c_i.

The decoder outputs a set $\hat{S} := \mathsf{Dec}_L^{\mathsf{ML}}(\mathsf{Enc}(S)) \neq S$ if and only if there is exists $S' \neq S$ such that $S' \prec_L S$ and $\sum_{i \in S'} c_i = \sum_{i \in S} c_i$. For any set $S' \prec_L S$ in L,

$$\begin{aligned}
\Pr\left(\sum_{i \in S'} c_i = \sum_{i \in S} c_i\right) \;&=\; \prod_{j=1}^{m} \Pr\left(\sum_{i \in S'} c_i(j) = \sum_{i \in S} c_i(j)\right) \\
&=\; \prod_{j=1}^{m} \Pr\left(\sum_{i \in S' \triangle S} c_i(j) = 0\right) \;=\; 2^{-m}.
\end{aligned}$$

By a union bound,

$$\begin{aligned}
\Pr\left(\mathsf{Dec}_L^{\mathsf{ML}}(\mathsf{Enc}(S)) \neq S\right) \;&=\; \Pr\left(\exists S' \prec_L S \,.\, \sum_{i \in S'} c_i = \sum_{i \in S} c_i\right) \\
&\le\; \sum_{S' \prec_L S} \Pr\left(\sum_{i \in S'} c_i = \sum_{i \in S} c_i\right) \\
&\le\; \big(|L| - 1\big)\, 2^{-m}.
\end{aligned}$$

□

In fact, one can prove a slightly stronger guarantee of success: that, for any fixed list L, with probability at least $1 - \delta$, the decoder decodes correctly *any* set $S \in L$. This leads to slightly worse competitiveness: $\alpha = 2 + o(1)$. In

particular, m-sized code can decode only lists of size 2^{m^*} where $m^* = \frac{1}{2}(m - \lg 1/\delta)$. The following corollary is immediate from the above.

Corollary 1. *Let C be a random $m \times N$ $0/1$ matrix. Then for any list L of subsets of $[N]$,*

$$\Pr_C \left(\forall S \in L \,.\, \mathsf{Dec}_L^{\mathsf{ML}}(C \cdot \mathbb{1}_S) = S \right) \;\geq\; 1 - |L| \cdot \big(|L| - 1\big)\, 2^{-m} \,.$$

III. Main Result: $O(\log \log N)$-competitive Coding Scheme

In this section, we prove Theorem 2, by designing a computationally efficient Huffman-asymmetric-coding scheme. The resulting algorithm is termed the *multi-level scheme* (for reason that will soon be apparent).

Let $\Delta([N])$ be the space of all distributions with support $[N]$. Our algorithm supports distributions μ from the following class

$$\mathcal{M} \;:=\; \big\{ \mu \in \Delta([N]) : 1/4N \leq \mu(i) < 1/2,\ \forall i \in [N] \big\} \,.$$

While this is a restriction from a general distribution $\mu \in \Delta([N])$, it is without loss of generality: we can transform any distribution into a distribution $\mu'' \in \mathcal{M}$ (up to a loss of at most factor 2 in the communication bound). First, if there are items i^* with probability more than 1/3, make them with probability $1/3$: set $\mu'(i^*) = 1/3$. Second, all the probabilities that are too small can be brought up to at least $1/4N$, while affecting the other probabilities only by a constant as follows: (1) construct $\mu'(i) = \max\{\mu(i), 1/2N\}$ (except for items i^*), (2) let $\zeta = \sum_i \mu'(i) \leq \sum_i (\mu(i) + 1/2N) = 1.5$, and (3) set $\mu''(i) = \frac{1}{\zeta}\mu'(i)$. It's not hard to verify now that $\mu'' \in \mathcal{M}$, as well as that $\mu''(i^*) \leq 1/2$ and for the other items $\lg 1/\mu''(i) \leq 2 \lg 1/\mu(i)$. We also assume that $m \geq \lg N + 4$.

Our scheme $\mathcal{A} = (\mathsf{Enc}, \mathsf{Dec})$ uses $T := \lg\lg(4N)$ levels, each parametrized by positive integers D_t, m_t to be determined later. We use uniformly random hash functions

$$h_t \colon [N] \to [D_t]$$

where the hash functions are determined using shared public randomness. The scheme also uses a family of T (deterministic) linear codes, $C^{(t)} = \left[\, c_1^{(t)}\ c_2^{(t)}\ \ldots\ c_{D_t}^{(t)} \,\right] \in \mathbb{F}_N^{m_t \times D_t}$ for $t \in [T]$, which are specified in the next subsection. Each matrix $C^{(t)}$ shall be designed to support efficient decoding of *every* $\left(\frac{m_t}{2 \lg D_t}\right)$-sparse vector. We now turn to the formal construction.

A. One level: sensing matrices $C^{(t)}$

For each level of our scheme, the basic building block is the compressed-sensing matrices designed in the work of [13]. These deterministic constructions produce $m \times N$ linear codes (matrices over some finite field) that can decode *any* k-sparse vector $x \in \mathbb{F}_2^N$ (i.e., any subset of size at most k), where $k := m/(2 \lg N)$, in time *polynomial* in m and N. Note that such a compression scheme is essentially optimal – the number of k-sparse subsets in $[N]$ is $\binom{N}{k} \approx 2^{k \lg(N/k)}$, hence any deterministic encoding scheme for this problem must use at least $k \lg(N/k) \approx m$ bits of communication.

We now state the formal theorem from [13]. The theorem relies on an elegant connection between channel coding and source coding (via "syndrome decoding"). The central object is the *parity check* matrix of a *Reed-Solomon*

code (see e.g., [23]). To this end, we denote by $[N, r, d]_q$ a Reed-Solomon code over the alphabet $\mathbb{F}_q$ ($q \geq \lg N$), whose codeword length is N, number of codewords is q^r, and the minimum Hamming distance between codewords is d (i.e., the code can correct up to $(d-1)/2$ errors). Our multi-level scheme uses the following theorem in a black-box fashion.

Theorem 3 (Efficient deterministic compressed sensing, [13])**.** *Let $\mathbf{P}_k^N \in \mathbb{F}_N^{m \times N}$ be the parity-check matrix of a $[N, N-2k, 2k+1]_{\mathbb{F}_N}$ Reed-Solomon code*[4]*, where $m = 2k\lceil \lg N \rceil$. There is a (deterministic) decoding algorithm that recovers any k-sparse vector in $\mathbb{F}_2^N$ (i.e., $x \in \binom{[N]}{k}$) from $\mathbf{P}_k^N \cdot x$ using $O(Nk \lg^2 N)$ operations over $\mathbb{F}_2$. In particular, $\mathbf{P}_k^N \cdot x$ uniquely determines x using $m = 2k\lceil \lg N \rceil$ linear measurements.*

The rough idea behind this result (which was used in the past) is to think of k-sparse vectors in $\mathbb{F}_2^N$ as a sparse *noise* vector introduced by a discrete memoryless channel, and then use the efficient *syndrome-decoding* algorithm for Reed-Solomon codes of Berlekamp and Massey (see [23]) which recovers the noise vector (i.e., our desired k-sparse subset) from the parity check matrix $\mathbf{P}_k^N$.

Of course, the main difference from the setup of Theorem 3 and our setup, is that in our case the original distribution on subsets (i.e., sparse vectors) may be very far from uniform. Nonetheless, our multi-level scheme uses the construction of [13] in each layer. More precisely, for level t of our scheme, our scheme shall set the matrix $C^{(t)}$ to be the parity-check matrix $\mathbf{P}_k^N$ with parameters $N := D_t$, $k := m_t/(2 \lg D_t)$ (i.e., it is a matrix of size $m_t \times D_t$). This will become clearer in the next section where we present the entire multi-level scheme.

B. Description and Analysis of the Multi-level Scheme

As mentioned in the previous section, the encoding and decoding of the input ($S \subseteq [N]$) is defined by an iterative procedure consisting of T levels, and crucially relies on the linearity of the encoding in each level. Let $\{D_t\}_{t \in [T]}$ and $\{m_t\}_{t \in [T]}$ be numbers to be determined later. The encoder is described in Algorithm 1, and the decoder is described in Algorithm 2.

We now turn to the analysis of the scheme, whose centerpiece is the following theorem.

Theorem 4. *Fix $\delta \in (0,1)$ and positive integer m^*. Set*

$$D_t \;:=\; \left\lceil \frac{T}{\delta} \cdot \left(2^{2 \cdot 2^t}/2 + \frac{(m^*)^2}{2^{2t}} \right) \right\rceil \;, \quad t \in [T]\,, \tag{2}$$

and

$$m_t \;:=\; \left\lceil 2 \lg D_t \cdot \min\left\{ \frac{m^*}{2^{t-1}}, \frac{4m^*}{\lg m^*} \right\} \right\rceil \;, \quad t \in [T]. \tag{3}$$

Then for any $\mu \in \mathcal{M}$ and S satisfying Eqn. (1) *with the fixed value of m^*, the Algorithm 2 outputs the set $\hat{S} = \mathsf{Dec}_\mu(\mathsf{Enc}(S))$ satisfying:*

$$\Pr[\hat{S} = S] \;\geq\; 1 - \delta\,.$$

[4]We assume here that N is a power of 2. Otherwise, replace it with $N' := 2^{\lceil \lg N \rceil}$.

Algorithm 1 Enc for multi-level scheme

input subset $S \subseteq [N]$ (represented as the indicator vector $\mathbb{1}_S \in \{0,1\}^N$).

output message $y \in \{0,1\}^m$.

For each $t \in [T]$, let $y_t := \sum_{i \in S} C^{(t)} \cdot \mathbb{1}_{\{h_t(i)\}}$, where $C^{(t)}$ is the $m_t \times D_t$ matrix $\mathbf{P}_{k_t}^{N_t}$ from Theorem 3, instantiated with $N_t := D_t$, $k_t := m_t/(2 \lg D_t)$. i.e., $y_t = \sum_{i \in S} c^{(t)}_{h_t(i)}$.

1: **return** concatenated string $y := (y^{(1)}, y^{(2)}, \ldots, y^{(T)})$

Algorithm 2 Dec_μ for multi-level scheme

input message $y = (y^{(1)}, y^{(2)}, \ldots, y^{(T)}) \in \{0,1\}^m$, and a prior distribution $\mu \in \mathcal{M}_m$.

output subset $\hat{S} \subseteq [N]$.

1: Let $B_t := \{i \in [N] : 2^{-2^t} \le \mu(i) < 2^{-2^{t-1}}\}$ for $t \in [T]$.
2: Initialize $\hat{S} := \emptyset$.
3: **for** $t = 1, 2, \ldots, T$ **do**
4: Let $\hat{z}^{(t)}$ be the output of the decoder for $C^{(t)}$ applied to $y^{(t)}$, guaranteed by Theorem 3.
5: **for** each $i \in B_t$ **do**
6: **if** $\hat{z}^{(t)}_{h_t(i)} = 1$ **then**
7: Let $\hat{S} := \hat{S} \cup \{i\}$.
8: **for** $\tau = t+1, t+2, \ldots, T$ **do**
9: Let $y^{(\tau)} := y^{(\tau)} - c^{(\tau)}_{h_\tau(i)}$.
10: **end for**
11: **end if**
12: **end for**
13: **end for**
14: **return** $\hat{S}$

We now briefly verify that Theorem 4 implies Theorem 2, when we set $m^* = m/\alpha$ where $\alpha = O(\lg \lg N + \lg 1/\delta)$. Since $\lg D_t \le \lg 2T/\delta + O(2^t) + O(\lg m^*)$, we have $m_t \le O(m^*(1 + 2^{-t+1} \lg 2T/\delta))$. The total message length over all the T levels is thus

$$\sum_{t=1}^{T} m_t = O(m^* \cdot T) + O(m^* \cdot \lg 2T/\delta) \le m^* \cdot \alpha = m.$$

Using Theorem 3, it is also clear that the running times of Algorithm 1 and Algorithm 2 are poly(N).

Proof of Theorem 4. Fix $\mu \in \mathcal{M}$ and S satisfying Eqn. (1). Because every $i \in S$ satisfies $\lg(1/\mu(i)) \le \lg(4N)$, we may partition S into $S_t := S \cap B_t$ for $t \in [T]$. Also let $S_{t:T} := S_t \cup S_{t+1} \cup \cdots \cup S_T$ for $t \in [T]$. Let E_t be the event in which the following hold:

1) $h_t(i) \neq h_t(j)$ for all distinct $i, j \in B_t$;
2) $h_t(i) \neq h_t(j)$ for all $i \in S_t$ and $j \in S_{t+1:T}$.

By definition, every $i \in B_t$ satisfies $\mu(i) \geq 2^{-2^t}$, and hence $|B_t| \leq 2^{2^t}$. Furthermore, every $i \in S_{t:T}$ satisfies $\mu(i) \leq 2^{-2^{t-1}}$, or equivalently, $1 \leq \frac{\lg(1/\mu(i))}{2^{t-1}}$. Therefore, it holds that

$$|S_{t:T}| \leq \sum_{i \in S_{t:T}} 1 \;\leq\; \sum_{i \in S_{t:T}} \frac{\lg(1/\mu(i))}{2^{t-1}} \;\leq\; \frac{\sum_{i \in S} \lg(1/\mu(i))}{2^{t-1}} \;\leq\; \frac{m^*}{2^{t-1}}\,,$$

where the final inequality follows since the set S satisfies Eqn. (1). For the size of the set S_t we note that $|S_t| \leq \min\{|S_{t:T}|, 4m^*/\lg m^*\}$. The last transition is due to the fact that at least half of the set is composed of items of probability mass at least $\lg \frac{2}{|S_t|}$, and thus, by Eqn. (1), $\frac{|S_t|}{2} \lg \frac{|S_t|}{2} \leq m^*$.

Now we note that

$$|S_t| \cdot |S_{t+1:T}| \;\leq\; \frac{1}{4} \cdot |S_{t:T}|^2 \;\leq\; \frac{(m^*)^2}{2^{2t}}\,.$$

Therefore, by a union bound, the probability that E_t holds is

$$\Pr(E_t) \;\geq\; 1 - \left(\binom{|B_t|}{2} + |S_t| \cdot |S_{t+1:T}| \right) \cdot \frac{1}{D_t} \;\geq\; 1 - \frac{\delta}{T}\,,$$

where the second inequality uses the choice of D_t in Eqn. (2). By another union bound over all $t \in [T]$, it follows that the event $E := E_1 \cap E_2 \cap \cdots \cap E_T$ holds with probability at least $1 - \delta$.

For the rest of the analysis, we condition on the occurrence of the event E. Let $\hat{S}_t$ be the set of items that Algorithm 2 adds to $\hat{S}$ in iteration t. It suffices to prove that if y is the encoding of items belonging only to buckets $B_t, B_{t+1}, \ldots, B_T$ (i.e., of the indicator vector $\mathbb{1}_{S_{t:T}}$), then upon reaching iteration t of the decoding algorithm, we have $\hat{S}_t = S_t$ (i.e., we argue that in level t we decode precisely the elements in S_t). Maintaining this invariant is indeed sufficient, because at the end of iteration t, Algorithm 2 subtracts the $C^{(\tau)}$-encoding of elements in $\hat{S}_t \cap B_t$ from $y^{(\tau)}$ for all $\tau > t$. Thus, if $\hat{S}_t = S_t$, then after iteration t, the *linearity* of the code implies that the message y (at least the parts relevant to rounds $> t$) no longer contains the items in S_t (and hence B_t).

Since we conditioned on the event E, the hash function h_t has no collisions between pairs of items in B_t, and moreover it has no collisions between items in S_t and items in $S \setminus S_t = S_{t+1:T}$ (where we use the assumption that $S = S_{t:T}$). Therefore, the items in S_t are in one-to-one correspondence with some subset of $\mathrm{supp}(z^{(t)})$, where

$$z^{(t)} \;:=\; \sum_{i \in S} e_{h_t(i)}\,.$$

The vector $z^{(t)}$ may have other non-zero entries not in the one-to-one correspondence with S_t, but they are not the image of any $i \in B_t$ under h_t. This implies that if $\hat{z}^{(t)} = z^{(t)}$, then $\hat{S}_t = S_t$.

We now argue that, indeed, we have $\hat{z}^{(t)} = z^{(t)}$. Observe that $z^{(t)}$ has at most $|S_{t:T}| \leq m^*/2^{t-1}$ non-entries in total (again, using the assumption that $S = S_{t:T}$), and $y^{(t)}$ is the encoding of $z^{(t)}$ under $C^{(t)}$, i.e., $y^{(t)} = C^{(t)} z^{(t)}$. Due to the choice of m_t from Eqn. (3) and Theorem 3, the decoding of $y^{(t)}$ returns $\hat{z}^{(t)} = z^{(t)}$ as required. □

IV. Lower Bound for Deterministic Schemes

We show that asymmetric coding schemes need to be randomized in order to gain advantage from using the side information. In particular we show that if the class of priors is sufficiently rich, then no *deterministic* asymmetric coding scheme can improve over the trivial baseline communication, even if we allow arbitrary (non-linear) schemes

and arbitrary decoding time. Note that this separates the asymmetric information case from the symmetric side information case—since the Huffman code is a deterministic (near)-optimal algorithm for the symmetric case.

We will prove the lower bound for the entropy-asymmetric-coding case (the weakest definition). We consider the family $\mathcal{M}_{N,k}$ of prior distributions that consists of all (product) distributions μ^k where μ is supported on some subset $M \subset [N]$ of cardinality $|M| = 2k$ (i.e., each μ defines a list $L = L(\mu)$ of all $\binom{2k}{k}$ subsets of $[M]$). More formally,

$$\mathcal{M}_{N,k} := \left\{ \mu^k \mid \operatorname{supp}(\mu) \subset M, \;\; M \subset [N], |M| = 2k \right\}.$$

Note that for any prior $\mu^k \in \mathcal{M}_{N,k}$, we have the information-theoretic minimum communication to be $m^* = H(\mu^k) = kH(\mu) \leq k \lg(2k)$. However, the following claim asserts that any deterministic scheme for $S \in \mathcal{M}_{N,k}$ must spend essentially the trivial communication of $\Omega(\lg \binom{N}{k}) = \Omega(k \lg N/k)$.

Claim 1 (Deterministic oblivious compression is impossible)**.** *Any entropy-asymmetric-coding scheme that handles priors $\sigma = \mu^k \in \mathcal{M}_{N,k}$, and achieves $\delta = 0$, must have $m = \Omega(k \lg(N/k))$ bits of communication even though the information-theoretic minimum is $m^* \leq k \lg 2k$. This remains true even without requiring linearity or computational efficiency.*

Proof. The idea is to use the fact that the encoder is oblivious to μ in order to argue that any deterministic encoding scheme can in fact be used to reconstruct *any* k-sparse vector in $\mathbb{F}_2^N$ (i.e., any subset $S \in \binom{[N]}{k}$). Clearly, the latter compression problem requires $\lg \binom{N}{k}$ bits of communication, hence the claim would follow. Indeed, we claim that a deterministic scheme $\mathcal{A} = (\mathsf{Enc}, \mathsf{Dec})$ that solves the entropy-asymmetric-coding problem, must satisfy

$$\forall \;\; S_1 \neq S_2 \subset \binom{[N]}{k} \;\; , \;\; \mathsf{Enc}(S_1) \neq \mathsf{Enc}(S_2).$$

Indeed, suppose this is false, then there is a pair of subsets $S_1 \neq S_2 \subset \binom{[N]}{k}$ which are mapped by $\mathcal{A}$ to the same message

$$\mathsf{Enc}(S_1) = \mathsf{Enc}(S_2) := \pi.$$

Now, consider the set $M := S_1 \cup S_2$ and let μ_M be the uniform distribution over M. Note that $|M| = |S_1 \cup S_2| \leq 2k$, and without loss of generality, assume that $|M| = 2k$ (otherwise, add arbitrary elements of $[N]$ to M). In this case, observe that $\mu_M^k \in \mathcal{M}_{N,k}$, and that $\Pr_{\mu_M^k}[S_1] = \Pr_{\mu_M^k}[S_2] = 1/|M|^k$. Therefore, with probability at least $\delta := 1/(2 \cdot |M|^k) = 1/(2 \cdot (2k)^k) > 0$, the decoding will fail, since

$$\Pr_{S \sim \mu_M^k} \left(\mathsf{Dec}_{\mu_M^k}(\mathsf{Enc}(S)) = S \right)$$
$$\leq 1 - 2\delta \cdot \min \left\{ \Pr \left(\mathsf{Dec}_{\mu_M^k}(\pi) = S_1 \right), \Pr \left(\mathsf{Dec}_{\mu_M^k}(\pi) = S_2 \right) \right\} \leq 1 - \delta < 1.$$

But this contradicts the premise that $\mathcal{A}$ is a deterministic communication scheme with respect to $\mathcal{M}_{N,k}$. This proves that the worst-case communication length of any deterministic scheme must be $\Omega(k \lg(N/k))$ bits even under the class of product distributions. □

Remark 1. If arbitrary (non-product) distributions are allowed, it is not hard to turn the above argument into an *average case* lower bound, for example, by considering the distribution σ that chooses S_1 or S_2 each with

probability $1/2$, where S_1, S_2 are the "colliding" sets from above (note that while $\sigma \notin \mathcal{M}_{N,k}$, $|L(\sigma)| = 2$). We also remark that this claim essentially states that prior-oblivious deterministic compression cannot perform any better than standard ("prior-free") compressed-sensing schemes for k-sparse vectors in $\mathbb{F}_2^N$, which indeed requires $\Theta(k \lg(N/k))$ bits/measurements.

V. Conclusions and Open Problems

We considered coding sets with asymmetric information, where each set is comprised of i.i.d. samples from a prior distribution μ over $[N]$, and μ is only known to the decoder. We showed that any such coding scheme must be randomized in order to gain advantage from the side information. Given an error probability δ, we designed a computationally efficient and linear coding scheme, which achieves an $O(\lg \lg N)$-competitive communication ratio compared to the optimal message length.

As we view this work as an initial step in the study of asymmetric compression, there are a few natural aspects of our assumptions that require further research:

- The most straightforward open question is whether the message length for product distributions over subsets of $[N]$ can be improved from $O_\delta(\lg \lg N)$ multiplicative overhead to $O(\lg(1/\delta))$ overhead, or even further to $O(\lg(1/\delta))$ *additive* overhead (matching the information bound of the baseline scheme from Theorem 1), while insisting on poly(N) decoding time. We note that even the scheme of [13] (for the uniform prior case) is only 2-competitive.
- As hinted before, we may also want decoding time which is sublinear in N, e.g., poly$(m, \log N)$. Note that this may be possible only if we allow the decoder to do preprocessing—otherwise, already its input μ has $\Omega(N)$ description size.
- Are the above goals simpler if we allow *non-linear* coding? Our scheme is linear, and we do not know if there exist more efficient non-linear coding schemes.
- Another important direction is to identify other natural instances of *non-product* distributions σ, where the problem is meaningful and poly-time, competitive coding schemes exist. As mentioned before, such a distribution σ must at minimum have a *succinct* description. A natural candidate family for modeling such succinct joint distributions on subsets of $[N]$ are *graphical models* [24]. It would be very interesting to develop compete with the (possibly much lower) entropy benchmark of joint distributions generated by low-order graphical models.
- Finally, one may want to construct schemes that have a somewhat better probability guarantee (somewhat akin to "for all" vs "for each" guarantee). While fully deterministic schemes are impossible, it may be possible to obtain the following guarantee: with probability $1 - \delta$, the decoder decodes correctly *any* set $S \in L$. It turns out that this is possible for the random code solution (see Corollary 1). It would be interesting if our main (computationally-efficient) result can be extended to this case as well.

APPENDIX A
CONNECTIONS BETWEEN DIFFERENT NOTIONS OF ASYMMETRIC-CODING SCHEMES

In this section, we show connections between different asymmetric coding schemes. First we show that a list-asymmetric-coding scheme implies a Huffman-asymmetric-coding scheme.

Claim 2. *If $\mathcal{A}$ is a list-asymmetric-coding scheme with parameters m_l^* and δ, then $\mathcal{A}$ is a Huffman-asymmetric-coding scheme with parameters $m_h^\star \le m_l^\star - \lg e$ and δ, and the same, fixed communication bound m.*

Proof. Consider any distribution μ over $[N]$. Let L be the list of subsets $S \subseteq [N]$ that satisfy Eqn. (1). We just need to show that the size of L is less than $e2^{m_h^\star} \le 2^{m_l^\star}$. A set S satisfies Eqn. (1) if and only if

$$\prod_{i \in S} \mu(i) \ge 2^{-m_h^\star}.$$

On the other hand

$$\begin{aligned}
\sum_{S \in L} \prod_{i \in S} \mu(i) &\le \sum_{S \subseteq [N]} \prod_{i \in S} \mu(i) \\
&= \sum_{(x_1,\ldots,x_N) \in \{0,1\}^N} \prod_{i=1}^{N} \mu(i)^{x_i} \\
&= \sum_{x_1 \in \{0,1\}} \mu(1)^{x_1} \sum_{x_2 \in \{0,1\}} \mu(2)^{x_2} \cdots \sum_{x_N \in \{0,1\}} \mu(N)^{x_N} \\
&= (1+\mu(1))(1+\mu(2)) \cdots (1+\mu(N)) \\
&\le e^{\mu(1)} e^{\mu(2)} \cdots e^{\mu(N)} \\
&= e.
\end{aligned}$$

Hence the size of list L is less than $e2^{m_h^\star} \le 2^{m_l^\star}$ and a list-asymmetric-coding scheme for list L, with parameters m_l^* and δ, yields an error probability δ. □

We now show that entropy-asymmetric-coding is the weakest of the three definitions, in that a list- or Huffman-asymmetric-coding scheme implies an entropy-asymmetric-coding scheme (with slightly weaker parameters). We first define, for any $\delta > 0$ and distribution $\sigma \in \Delta(2^{[N]})$, the *$\delta$-approximate cover size* of σ as

$$\mathcal{C}(\sigma, \delta) := \min_{m \in \mathbb{N}} \left\{ \exists L \subseteq \mathrm{supp}(\sigma), |L| \le 2^m \ , \ \sigma(L) \ge 1 - \delta \right\}.$$

The following claim asserts an upper bound on the cover number in terms of the Shannon entropy of σ.

Claim 3 (Cover-size vs. Entropy)**.** *For every distribution σ and $\delta > 0$, it holds that*

$$\mathcal{C}(\sigma, \delta) \ \le \ H(\sigma)/\delta.$$

We remark that the bound is essentially tight, as demonstrated by the distribution σ which has an "atom" of measure δ and otherwise uniform on the entire domain.

Proof. Let $\mathcal{G}_\delta := \{x : \lg(1/\sigma(x)) \le H(\sigma)/\delta\}$ be the set of elements with "large" mass under σ. Indeed, note that $\forall x \in \mathcal{G}_\delta$ we have $\sigma(x) \ge 2^{-H(\sigma)/\delta}$, thus it holds that $|\mathcal{G}_\delta| \le 2^{H(\sigma)/\delta}$. In order to conclude that $\mathcal{C}(\sigma,\delta) \le H(\sigma)/\delta$, it remains to show that $\sigma(\mathcal{G}_\delta) \ge 1-\delta$. Indeed, Markov's inequality implies that

$$\sigma(\mathcal{G}_\delta) = 1 - \sigma(\overline{\mathcal{G}_\delta}) = 1 - \Pr_{x\sim\sigma}\left(\lg\frac{1}{\sigma(x)} > \frac{H(\sigma)}{\delta}\right) = 1 - \Pr_{x\sim\sigma}\left(\lg\frac{1}{\sigma(x)} > \frac{\mathbb{E}\left[\lg\frac{1}{\sigma(x)}\right]}{\delta}\right) \ge 1-\delta.$$

□

The following is a corollary of Claim 3.

Claim 4. *If $\mathcal{A}$ is a list-asymmetric-coding scheme with parameters m_l^* and δ_l, then $\mathcal{A}$ can be converted into an entropy-asymmetric-coding scheme with parameters $m_e^* := \delta_l m_l^*$ and $\delta_e := 2\delta_l$ (and same, fixed communication bound m).*

Proof. For any prior σ on subsets of $[N]$, there is a list $L = L(\sigma)$ of size at most $2^{H(\sigma)/\delta_l}$ which is "responsible" to $1-\delta_l$ mass of the distribution.[5] So, when the encoding length is fixed to m, Claim 3 guarantees that decoding (w.p. $1-\delta_l$) all subsets with $\sigma(S) \ge 2^{-m_l^*}$ is equivalent to decoding (w.p. $1-\delta_l$) all distributions with Shannon entropy at most $\delta_l m_l^*$. □

Note that $\delta_l m_l^*$ bits are needed even in the standard compression setup when both parties know the distribution, hence this notion of decoding is competitive even with the Shannon entropy benchmark, which is the strongest possible.

Similarly, we can show that a Huffman-asymmetric-coding scheme implies an entropy-asymmetric-coding scheme (with some loss in the communication efficiency).

Claim 5. *If $\mathcal{A}$ is a Huffman-asymmetric-coding scheme with parameters m_h^* and δ_h, then for any $\epsilon \in (0,1)$, $\mathcal{A}$ is an entropy-asymmetric-coding scheme with parameters*

$$m_e^* := \left\lfloor \frac{1-\delta_h/(2N)}{1+\epsilon}\left(m_h^* - \left(\frac{1}{2\epsilon}+\frac{1}{3}\right)\lg(2N^2/\delta_h)\ln(2/\delta_h)\right)\right\rfloor, \quad \delta_e := 2\delta_h,$$

and same, fixed communication bound m.

Proof. Assume $\mathcal{A}$ is a Huffman-asymmetric-coding scheme with parameters m_h^* and δ_h. Take any $\mu \in \Delta([N])$ with $kH(\mu) \le m_e^*$. Define $\delta_0 := \delta_h/(2N^2)$. Let $\mathsf{Head} := \{i \in [N] : \mu(i) \ge \delta_0\}$ and $\mathsf{Tail} := [N] \setminus \mathsf{Head}$. Let E be the event where $S \sim \mu^k$ satisfies $S \subseteq \mathsf{Head}$. Since $(1-N\delta_0)^k \ge 1 - Nk\delta_0 \ge 1 - \delta_h/2$, it follows that

$$\Pr_{S\sim\mu^k}(E) \ge 1 - \delta_h/2.$$

Furthermore, conditional on E, we can bound the expected value of $\sum_{i\in S}\lg(1/\mu(i))$ as follows:

$$kH_E(\mu) := \mathbb{E}_{S\sim\mu^k}\left[\sum_{i\in S}\lg(1/\mu(i)) \;\middle|\; E\right] = \frac{k}{1-\mu(\mathsf{Tail})}\sum_{i\in\mathsf{Head}}\mu(i)\lg(1/\mu(i)) \le \frac{k}{1-\delta_h/(2N)}H(\mu).$$

[5] As mentioned before, this "truncation" of the tail of σ seems inherent to oblivious schemes, as they are *fixed-length* encodings.

By Bernstein's inequality, we have

$$\Pr_{S\sim\mu^k}\left(\sum_{i\in S}\lg\frac{1}{\mu(i)} \le kH_E(\mu) + \sqrt{2kH_E(\mu)\lg\left(\frac{2N^2}{\delta_h}\right)\ln\left(\frac{2}{\delta_h}\right)} + \frac{\lg\left(\frac{2N^2}{\delta_h}\right)\ln\left(\frac{2}{\delta_h}\right)}{3} \;\middle|\; E\right) \ge 1-\frac{\delta_h}{2}.$$

Therefore, with probability at least $1-\delta_h$ over the random draw $S\sim\mu^k$, we have

$$\begin{aligned}\sum_{i\in S}\lg(1/\mu(i)) &\le \frac{kH(\mu)}{1-\delta_h/(2N)} + \sqrt{\frac{2kH(\mu)\lg(2N^2/\delta_h)\ln(2/\delta_h)}{1-\delta_h/(2N)}} + \frac{\lg(2N^2/\delta_h)\ln(2/\delta_h)}{3}\\ &\le \frac{1+\epsilon}{1-\delta_h/(2N)}kH(\mu) + \left(\frac{1}{2\epsilon}+\frac{1}{3}\right)\lg(2N^2/\delta_h)\ln(2/\delta_h)\\ &\le m_h^*\end{aligned}$$

where the second inequality follows from the arithmetic-mean/geometric-mean inequality, and the last inequality uses the definition of m_e^*. Conditional on this event, $\mathcal{A}$ correctly decodes the set S with probability at least $1-\delta_h$. Thus, $\mathcal{A}$ is an entropy-asymmetric-coding scheme with parameters m_e^* and $\delta_e = 2\delta_h$. □

REFERENCES

[1] M. Adler and B. M. Maggs, "Protocols for asymmetric communication channels," in *Foundations of Computer Science, 1998. Proceedings. 39th Annual Symposium on*. IEEE, 1998, pp. 522–533.

[2] M. Gorlatova, P. Kinget, I. Kymissis, D. Rubenstein, X. Wang, and G. Zussman, "Challenge: ultra-low-power energy-harvesting active networked tags (enhants)," in *Proceedings of the 15th annual international conference on Mobile computing and networking*. ACM, 2009, pp. 253–260.

[3] T. Chen, J. Ghaderi, D. Rubenstein, and G. Zussman, "Maximizing broadcast throughput under ultra-low-power constraints," in *Proceedings of the 12th International on Conference on emerging Networking EXperiments and Technologies*. ACM, 2016, pp. 457–471.

[4] M. Buettner, B. Greenstein, and D. Wetherall, "Dewdrop: an energy-aware runtime for computational rfid," in *Proc. USENIX NSDI*, 2011, pp. 197–210.

[5] M. Adler, E. D. Demaine, N. J. Harvey, and M. Pătraşcu, "Lower bounds for asymmetric communication channels and distributed source coding," in *Proc. 17th ACM/SIAM Symposium on Discrete Algorithms (SODA)*, 2006, pp. 251–260.

[6] A. Gilbert and P. Indyk, "Sparse recovery using sparse matrices," *Proceedings of the IEEE*, vol. 6, no. 98, pp. 937–947, 2010.

[7] A. C. Gilbert, Y. Li, E. Porat, and M. J. Strauss, "Approximate sparse recovery: optimizing time and measurements," *SIAM Journal on Computing*, vol. 41, no. 2, pp. 436–453, 2012.

[8] E. Candes and T. Tao, "Near optimal signal recovery from random projections: Universal encoding strategies," *IEEE Transactions on Information Theory*, 2006.

[9] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52(4), pp. 1289 – 1306, 2006.

[10] S. C. Draper and S. Malekpour, "Compressed sensing over finite fields," in *Proceedings of the 2009 IEEE international conference on Symposium on Information Theory-Volume 1*. IEEE Press, 2009, pp. 669–673.

[11] J.-T. Seong and H.-N. Lee, "Necessary and sufficient conditions for recovery of sparse signals over finite fields," *Communications Letters, IEEE*, vol. 17, no. 10, pp. 1976–1979, 2013.

[12] W. Li, F. Bassi, and M. Kieffer, "Robust bayesian compressed sensing over finite fields: asymptotic performance analysis," *arXiv preprint arXiv:1401.4313*, 2014.

[13] A. K. Das and S. Vishwanath, "On finite alphabet compressive sensing," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, 2013, pp. 5890–5894. [Online]. Available: http://dx.doi.org/10.1109/ICASSP.2013.6638794

[14] D. A. Huffman, "A method for the construction of minimum-redundancy codes," *Proceedings of the IRE*, vol. 40, no. 9, pp. 1098–1101, 1952.

[15] E. S. Laber and L. G. Holanda, "Improved bounds for asymmetric communication protocols," *Information Processing Letters*, vol. 83, no. 4, pp. 205–209, 2002.

[16] M. Ghodsi and A. Saberi, "A new protocol for asymmetric communication channels: Reaching the lower bounds," *Scientia Iranica*, vol. 8, no. 4, pp. 297–302, 2001.

[17] J. Watkinson, M. Adler, and F. E. Fich, "New protocols for asymmetric communication channels," in *SIROCCO*, 2001.

[18] Z. Xiong, A. D. Liveris, and S. Cheng, "Distributed source coding for sensor networks," *Signal Processing Magazine, IEEE*, vol. 21, no. 5, pp. 80–94, 2004.

[19] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Transactions on Information Theory*, vol. 56, no. 4, pp. 1982–2001, 2010.

[20] G. Caire, S. Shamai, and S. Verdú, "Noiseless data compression with low-density parity-check codes," in *Advances in Network Information Theory, Proceedings of a DIMACS Workshop, Piscataway, New Jersey, USA, March 17-19, 2003*, 2003, pp. 263–284.

[21] S. H. Hassani and R. L. Urbanke, "Universal polar codes," in *2014 IEEE International Symposium on Information Theory, Honolulu, HI, USA, June 29 - July 4, 2014*, 2014, pp. 1451–1455. [Online]. Available: https://doi.org/10.1109/ISIT.2014.6875073

[22] J. Garcia-Frias and Y. Zhao, "Compression of binary memoryless sources using punctured turbo codes," *IEEE Communications Letters*, vol. 6, no. 9, pp. 394–396, 2002. [Online]. Available: https://doi.org/10.1109/LCOMM.2002.803484

[23] R. M. Roth, *Introduction to coding theory*. Cambridge University Press, 2006.

[24] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Foundations and Trends in Machine Learning*, vol. 1, no. 1-2, pp. 1–305, 2008. [Online]. Available: https://doi.org/10.1561/2200000001