Coding with Asymmetric Prior Knowledge

Alexandr Andoni\textsuperscript{1}, Javad Ghaderi\textsuperscript{2}, Daniel Hsu\textsuperscript{1}, Dan Rubenstein\textsuperscript{1}, and Omri Weinstein\textsuperscript{1}

\textsuperscript{1}Department of Computer Science, Columbia University
\textsuperscript{2}Department of Electrical Engineering, Columbia University

July 18, 2017

Abstract

We study the problem of compressed sensing with asymmetric prior information, as motivated by networking applications. We focus on the scenario in which a resource-limited encoder needs to report a small subset $S$ from a universe of $N$ objects to a more powerful decoder. The distinguishing feature of our model is asymmetry: the subset $S$ is an i.i.d. sample from a prior distribution $\mu$, and $\mu$ is only known to the decoder. This scenario implies that the encoder must use an oblivious compression scheme which can, nonetheless, achieve communication comparable to the entropy rate $|S| \cdot H(\mu)$, the standard benchmark when both encoder and decoder have access to the prior $\mu$ (as achieved by the Huffman code). In networks, this scenario models the omnipresent task of identity tracking: the encoder is a resource-limited tag in an Internet-of-Things universe, and is occasionally required to report a small subset of encountered tags to a more powerful gateway that resides closer to the datacenter.

We first show that in order to exploit the prior $\mu$ in a non-trivial way in such asymmetric information scenario, the compression scheme must be randomized. This stands in contrast to the symmetric case (when both the encoder and decoder know $\mu$), 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 essentially optimal communication rate of $O(k \cdot H(\mu))$ bits, nearly-matching the communication benchmark in the symmetric case. Alas, the resulting scheme has prohibitive decoding time: about $\left(\frac{N}{k}\right)^k \approx \frac{N}{k}$.

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 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 $\mu$ as a rather small convex combination of uniform (“flat”) distributions.

1 Introduction

This paper studies the problem of compressed sensing with asymmetric prior information. There is a universe $[N] := \{1, 2, \ldots, N\}$ of $N$ items (henceforth called tags), and the task is to encode a subset $S \subset [N]$ into an $m$-bit message so that a decoder can reconstruct the set $S$ efficiently. The new aspect of the problem we study is that the decoder has a prior distribution $\sigma$ over the sets $S$ that may be sent, which is not available at encoding time. The main goal is to design encoding/decoding schemes that (1) obtain communication rate as close as possible to the information-theoretical
minimum, namely the entropy bound with respect to the distribution $\sigma$, and (2) are computationally efficient.

The motivation for this problem stems from the need to design efficient communication protocols for emerging Internet-of-Things (IoT) environments in which a large collection of devices are constantly reporting their status wirelessly through a local gateway to a centralized processing point, likely located in the Cloud and/or in a Data Center. This centralized processing point aggregates the information from these reports and can make decisions about what actions need to be taken. Two central applications of such infrastructure are:

- **Object tracking.** Radio tags are attached to physical objects which can be located by identifying the gateway through which the tag communicated. Tracking can be further refined by utilizing multiple gateway points [WK13] or by allowing tags to perform short-range communications with one another to build a neighborhood map [CGRZ16].

- **Status reporting.** Beyond their identifier, devices are now able and sometimes expected to periodically report statistics of items it is assigned to monitor, e.g., their weight, or count (of multiple items), orientation (cardinal direction) [AIM10]. Health monitoring is an example of a domain where applications periodically report discrete statistics such as breathing rate and pulse [KLS+10].

Often the aggregate bandwidth required to send this information can be constraining, and any reduction in communication is critical. Our own interest in developing these compression techniques is so that we can apply them to reduce tag communication overheads, thereby reducing energy consumption of the tags [GKK+09, CGRZ16, BGW11]. To elaborate, consider the above setting of object (tag) tracking. To track these physical objects, the tags occasionally send messages to the more powerful devices (the gateways). Since the tags are not always in the proximity with the gateways, tags “accumulate” information into one message (say, via tag-to-tag communication) and send it to the gateway when available. Overall, the gateways receive a message (or more) which encodes a subset of tags that are in proximity with the gateway. We thus obtain the aforementioned coding problem, where the tags play the role of an encoder and the gateway of a decoder. Since the tags have extremely low power budgets, we are interested in protocols which use the minimal possible communication.

The crucial aspect of our setting is that the gateways may have side information (prior knowledge) about which tags are likely to be close to the gateway (e.g., the tags that were seen recently or were seen at a near-by gateway), or, say, to be together. This side-information may be modelled by a prior distribution $\sigma$ over the set of $S$ tag encoded in the message received by the gateway. Note that this side information is not available to the tags, at the encoding time.

We need to make further assumptions on the distribution $\sigma$, since assuming a general prior $\sigma$ not only raises the questions of efficient representation, but in fact doesn’t allow for efficient solutions. First, the description of a general prior $\sigma$ can be exponentially long in $N$, dooming the computational efficiency. Second and more importantly, it is not possible to improve the communication over the trivial $\sim N$ without further assumptions on $\sigma$ [AM98]. In this paper, we will focus on the most natural special class of priors: the product distributions $\sigma = \mu^k$, i.e., $S$ is comprised of $k$ items, each drawn independently from $\mu$.

A further desirable property of an encoding scheme is that the encoding is linear, i.e., the encoding is $C \cdot 1_S$ where $C$ is the coding matrix and $1_S$ is the indicator vector of the set $S$. This property is similar to the one imposed in compressed sensing (albeit, here, we count the number of bits of the encoding, as opposed to merely the dimension, as is common in many compressed sensing
schemes). Linearity facilitates quick and simple updates to the message in dynamic environments such as IoT as the message is simply updated as tags are added one by one to the set $S$.

Finally, we stress that, as is usual for such problems, we require that the decoding procedure is \textit{computationally efficient}, which we consider to be decoding time polynomial in $N$ (although it would also makes sense to ask for sublinear runtime, of the order of $\text{poly}(\lvert S\rvert, \lg N)$, as was accomplished in some compressed literature; see, e.g., [GI10, GLPS12]).

### 1.1 Relation to Problems in Prior Literature

Our problem naturally has connections to many problems studied previously. When there is \textit{no side information}, the problem is the classic problem of coding a set $S$. Without requiring linearity, a trivial solution is to append the indeces of items in $S$, yielding communication $k \lg N$ for sets $S$ of size $k$. 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 [CT06, Don06], and the vector $C \cdot 1_S$ 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., [DM09, SL13, LBK14, DV13b].

Another related model is when both the encoder and the decoder have access to some prior distribution $\sigma$ over the sets $S$. Then the optimal solution can be obtained via \textit{Huffman coding} [Huf52]. In that case, the expected length of the message is $\sum_S \sigma(S) \lg 1/\sigma(S) \leq H(\sigma) + |S|$, which is close to the information-theoretic optimum (at least up to the rounding issues). When considering a product distribution $\sigma = \mu^k$, the length of the compression of a set $S$ is $\sum_{i \in S} \lg 1/\mu(i)$, which, in expectation is again upper bounded by $k \cdot H(\mu) + k$. We remark that this variant of the problem is also related to the notion of \textit{model-based compressed sensing}, where both the encoder and decoder have prior knowledge on the possible structure (model) of the set $S$ (beyond, say, an upper bound on its size); see, e.g., [BCDH10].

Finally, when the side information is not known to the encoder (as it is in our case), the problem becomes a classic asymmetric transmission problem, related to the distributed source coding (Slepian-Wolf) problem [AM98, LH02, GS01, WAF01, ADHP06] (see also [XLC04]). In this problem, a client (tag), or more clients, generates elements $i$ from a probability distribution $\mu$ and needs to communicate their identities to the server (gateway). The goal is again to reach the information capacity of $\approx k \cdot H(\mu)$. While there are protocols that achieve such capacity, the protocols require \textit{two-way} communication (i.e., gateways need to communicate information to tags). In fact, classic impossibility results show that either the client or the server has to communicate $\Omega(\lg N)$ bits (the trivial solution) [AM98]. This setup differs from our problem in that: (1) we disallow gateway-to-tag communication (in the spirit of reducing the work done by tags), and (2) the tags need to send a set $S$ of items, instead of a single one. The latter condition in particular would allow to bypass the above lower bound: we can amortize the lower bound of $\Omega(\lg N)$ against $|S|$ items. In other words, in our setting, we want to encode the set $S$ using $m \geq \lg N$ bits, with the goal of achieving $m \ll O(|S| \cdot \lg N)$ where possible.

### 1.2 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 $N$ items (tags). For a given set $S \subseteq [N]$, the encoder $\text{Enc} : 2^{[N]} \to \{0, 1\}^m$ must construct a (possibly randomized) message $y := \text{Enc}(S)$ of at most $m$ bits, where $m$ is a pre-determined number that captures the

---

1We use $\lg$ to denote base-2 logarithm.
allowed message length. The decoder $\text{Dec}_\star: \{0,1\}^m \to 2^{|N|}$, for a fixed side-information $\star$, must produce a set $\hat{S} := \text{Dec}_\star(y)$ from the message $y$ such that $\hat{S} = S$ with, say, at least $1 - \delta$ probability, where $\delta$ is the error probability parameter (think $\delta = 0.1$). Note that, when the side information $\star$ is null, this task is generally impossible unless $m \geq \lg 2^{|N|} = N$. Note that the encoder’s message may not depend on the side information, i.e., the encoding function $\text{Enc}(S)$ must be oblivious (in the information theory literature this is referred to as universal compression [CSV03, HU14]; see also Section [1.3]).

One natural way to model the side information is via a prior distribution $\sigma$ on subsets of $[N]$. We further assume that the distribution $\sigma$ is a product distribution $\sigma = \mu^k$, where $\mu$ is a distribution on $[N]$ (although our theorems will be stated for a slightly more general setting). While probability distributions are common in information theory, we will shortly introduce a different way model the prior information as well.

As is usual, we will allow some slack in the optimality of a coding scheme. In particular, in the definitions below we use parameter $m^*$, which will have the following interpretation: the coding scheme uses $m$ bits to describe something whose “information-theoretic optimal” description is $m^*$ bits. In fact, this notion of “information-theoretic optimality” is what will primary differ between different definition. Typical functions $m^*$ will be $m^* = m - \alpha$ (i.e., a scheme which achieves optimal communication bound, up to $\alpha \geq 1$ extra bits) and $m^* = m/\alpha$ (optimal bound up to a multiplicative $\alpha \geq 1$ factor in the communication bound).

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. The first and most natural definition is the following.

**Definition 1.1.** With the parameters from above, a (randomized) scheme $A = (\text{Enc}, \text{Dec})$ is said to be an entropy-asymmetric-coding scheme if: for any integer $k$, and prior $\sigma = \mu^k$ such that $k \cdot H(\mu) \leq m^*$, we have that:

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

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

Note that $k \cdot H(\mu)$ is the theoretical minimum communication necessary to transmit $S \sim \mu^k$. The trivial scheme would achieve a bound of $k \lg N$, which can be much higher than $kH(\mu)$.

The above definition turns out to be a little restrictive, and therefore we consider a more general definition. One of its features is that it does not fix $S$ to be of a fixed size and is rather “adaptive” to the items in the set $S$.

**Definition 1.2.** With the parameters from above, a (randomized) scheme $A = (\text{Enc}, \text{Dec})$ is said to be a Huffman-asymmetric-coding scheme if: for any distribution $\mu$ over $[N]$, if the set $S$ satisfies the following:

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

then

$$\Pr_A[\text{Dec}_\mu(\text{Enc}(S)) = S] \geq 1 - \delta.$$  

The name of a Huffman-asymmetric-coding scheme stems from the fact that it essentially matches the performance of the aforementioned Huffman coding (where the encoder knows the prior $\mu$), for $m^* = m$ (modulo the rounding issues), and $\delta = 0$ (deterministically). We also note
that Eqn. (1) (with $m^* = m$) is the tightest condition we can require in order for a set $S$ to be decodable with a classic Huffman code. Hence, the above definition asks to match the efficiency of the Huffman code (symmetric information setting) in the asymmetric setting.

While the connection to the entropy-asymmetric-coding schemes may not be immediate, we prove that a Huffman-asymmetric-coding scheme implies an entropy-asymmetric-coding scheme (with some loss in communication efficiency); see Claim A.4 in Appendix A.

It is also possible to consider a even more general definition, which is the most natural from the algorithmic/TCS 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 $\sigma$, one could take these sets to be the “most likely” in $\sigma$ (with ties broken arbitrarily).

**Definition 1.3.** Fix the parameters as above, a (randomized) scheme $A = (\text{Enc}, \text{Dec})$ is said to be a list-asymmetric-coding scheme if: for any list $L$ of sets $S \subseteq [N]$, where $|L| \leq 2^m^*$, and any $S \in L$, we have that:

$$\Pr_A[\text{Dec}_L(\text{Enc}(S)) = S] \geq 1 - \delta.$$  

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 $\mu$, 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 $\leq e2^m^*$ (which results in just an additive $\lg e$ additive loss in communication); see Claim A.4 in Appendix A.

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

### 1.3 Our Results

First, we prove that any asymmetric-coding scheme must be randomized if it is to non-trivially exploit the prior $\mu$. In particular, if $\delta = 0$ (i.e., no randomization), then, there exists some priors where optimal communication in the symmetric case is $m^* = O(|S| \cdot \lg |S|)$, but the asymmetric-coding scheme must have $m \approx \Theta(|S| \cdot \lg N)$. See details in Section 3.

Second, as a warm-up we show a simple scheme that solves the most general definition, of list-asymmetric-coding scheme.

**Theorem 1.4** (Information-theoretic; see Section 2). Fix error probability $\delta > 0$. There is a list-asymmetric-coding scheme with $m = m^* + \lceil \lg 1/\delta \rceil$, while achieving error probability of $\delta$.

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 $\text{Enc}(S) = C \cdot 1_S$ (all computations are done in $\mathbb{F}_2$). The decoder $\text{Dec}(y)$ is the “maximum likelihood” decoder: for a given list $L$, go over the list in order and outputs the first set $\hat{S} \in L$ such that $C1_{\hat{S}} = y$. See Section 2 for further details and proofs.

While the above scheme achieves the information-theoretic bound (up to an additive constant), its fundamental downside is that it is not computationally-efficient (as is common for random linear coding schemes) and requires runtime of about $\Omega(2^m)$. This remains true even when the list $L$ is somehow more efficiently represented: e.g., all sets $S$ that satisfy the Huffman condition Eqn. (1). Obtaining faster decoding time is precisely the focal point of our work:
Main goal: Develop computationally efficient oblivious compression schemes, that have only \( \text{poly}(N) \) encoding/decoding time, at the expense of a (mild, multiplicative) overhead in communication cost compared to random codes.

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 to the message length.

**Theorem 1.5** (Main; see Section 4). Fix error probability \( \delta > 0 \) and \( m > \log N + 4 \). There is a linear Huffman-asymmetric-coding scheme with \( m = m^* \cdot O(\log(\log(1/\delta))) \), with \( \text{poly}(N) \) decoding time, while achieving error probability of \( \delta \).

1.4 Discussion

Noiseless compression in asymmetric scenarios was also previously studied in the information theory literature, in the context of universal compression (see e.g., [CSV03, HU14, DV13a] 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 1.5, see also the discussion in Section 4.1). 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 [CSV03] and Turbo codes [GZ02]. The main 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 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 BSC. This assumption is equivalent in our model to considering only \( i.i.d \) distributions \( \mu \) on the \( [N] \) coordinates (i.e., each item \( i \) is present \( i \in S \) iid with certain probability), whereas we wish to deal with arbitrary distributions \( \mu^k, \mu \in \Delta([N]) \) (where \( \Delta([N]) \) denotes the set of all distributions over \( [N] \)).

**Open questions.** 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 \( O(\delta \log \log N) \) multiplicative overhead to \( O(\log(1/\delta)) \) overhead, or even further to \( O(\log(1/\delta)) \) additive overhead, while insisting on \( \text{poly}(m) \) encoding and decoding time.

- It is natural and practical to seek an extension of our results to “simple” classes of non-product prior distributions with succinct descriptions. A natural candidate family for modeling such succinct joint distributions on subsets of \( [N] \) are graphical models [WJ08]. It would be very interesting to develop low communication oblivious compression schemes that compete with the (possibly much lower) entropy benchmark of joint distributions generated by low-order graphical models.

- The ideal goal is to obtain algorithms matching the information bound of the baseline scheme (Theorem 1.4), while achieving a recovery time linear (or polynomial) in \( |S| \) and \( \log N \). Alternatively, can one obtain such an efficient recovery time with a modest increase in the message length? Note that even achieving time \( O(N) \ll 2^m \) appears challenging, although more plausible, perhaps by extending the scheme from [DV13b].
• While the above problem has been formulated as a one-shot sketch, the networking application requires the sketch of the set $S$ to be computed in a streaming fashion. In particular, the sketch should be updated as tags are added one by one to the set $S$. The common approach is to use a linear sketch (as we do in our work). However, it is possible that a solution would proceed differently than a linear sketch. For example the naive solution which just writes out a list of $\lg N$-bit indices of items $S$ is not linear.

• Finally, one may want to construct schemes that have a somewhat better probability 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 in for the random code solution (see Corollary 2.2). It would be interesting if our main (computationally-efficient) result can be extended to this case as well.

2 A Basic Scheme: Random Linear Codes

We establish Theorem 1.4 by designing a list-asymmetric-coding scheme via 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 the random linear code 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 $\text{Enc}(S) = C \cdot \mathbb{1}_S$. The decoder for a list $L = (S_1, S_2, \ldots, S_{|L|})$ is the “maximum likelihood” decoder:

$$\text{Dec}_{L}^{\text{ML}}(y) := S_{\min \{t \in |L| : \text{Enc}(S_t) = y \}}.$$  

(The random matrix $C$ is determined using the public random bits). In other words, given the message $y$, the decoder returns the first set $S$ in the list $L$ such that $\text{Enc}(S) = y$. For brevity, we call this the random linear scheme.

Below we state standard guarantees for the $m \times N$ random scheme. Lemma 2.1 establishes that the random linear scheme is list-asymmetric-coding scheme for any $\delta \in (0, 1)$ and any list of at most $2^m$ subsets of $[N]$. It implies immediately Theorem 1.4.

**Lemma 2.1.** Let $C$ be a random $m \times N$ 0/1 matrix. Then for any list $L$ of $2^m$ subsets of $[N]$, and any $S \in L$,

$$\Pr \left( \text{Dec}_{L}^{\text{ML}}(C \cdot \mathbb{1}_S) = S \right) \geq 1 - \left(2^m - 1\right)2^{-m}.$$

**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 $\hat{S} := \text{Dec}_{L}^{\text{ML}}(\text{Enc}(S)) \neq S$ 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 $L$,

$$\Pr \left( \sum_{i \in S'} c_i = \sum_{i \in S} c_i \right) = \prod_{j=1}^{m} \Pr \left( \sum_{i \in S'} c_i(j) = \sum_{i \in S} c_i(j) \right)$$

$$= \prod_{j=1}^{m} \Pr \left( \sum_{i \in S' \triangle S} c_i(j) = 0 \right) = 2^{-m}.$$
By a union bound,

\[
\Pr \left( \text{Dec}^\text{ML}_L(\text{Enc}(S)) \neq S \right) = \Pr \left( \exists S' \prec_L S : \sum_{i \in S'} c_i = \sum_{i \in S} c_i \right) \\
\leq \sum_{S' \prec_L S} \Pr \left( \sum_{i \in S'} c_i = \sum_{i \in S} c_i \right) \\
\leq (|L| - 1) 2^{-m}.
\]

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 requires a slightly larger blow-up: we require that \( m^* = \frac{1}{2}(m - \lg 1/\delta) \). The following corollary is immediate from the above.

**Corollary 2.2.** Let \( C \) be a random \( m \times N \) 0/1 matrix. Then for any list \( L \) of \( 2^m \) subsets of \([N]\),

\[
\Pr \left( \forall S \in L : \text{Dec}^\text{ML}_L(C \cdot 1_S) = S \right) \geq 1 - 2^{m^*} \left( 2^{m^*} - 1 \right) 2^{-m}.
\]

### 3 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 algorithm for the symmetric case.

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

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

Note that for any prior \( \mu^k \in \mathcal{M}_{N,k} \), we have \( H(\mu^k) = kH(\mu) \leq k\lg(2k) \). 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 3.1 (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 \( m^* \leq k\lg 2k \). This remains true even without requiring linearity or computational efficiency.

**Proof.** The idea is to use the fact that the encoder is oblivious to \( \mu \) 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 \( A = (\text{Enc}, \text{Dec}) \) that solves the entropy-asymmetric-coding problem, must satisfy

\[
\forall \ S_1 \neq S_2 \subset \binom{[N]}{k}, \ \text{Enc}(S_1) \neq \text{Enc}(S_2).
\]
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
\[
\text{Enc}(S_1) = \text{Enc}(S_2) := \pi.
\]
Now, consider the set \( M := S_1 \cup S_2 \) and let \( \mu_M \) 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
\[
\Pr_{S \sim \mu_M^k} \left( \text{Dec}(\mu_M^k, \text{Enc}(S)) = S \right) \\
\leq 1 - 2\delta \cdot \min \left\{ \Pr \left( \text{Dec}(\mu_M^k, \pi) = S_1 \right), \Pr \left( \text{Dec}(\mu_M^k, \pi) = S_2 \right) \right\} \\
\leq 1 - \delta < 1.
\]
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.

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 \( \sigma \) 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.

\[\square\]

4 Main Result: Multi-level Scheme

In this section, we present our main result, Theorem 1.5, with is a coding algorithm we term the multi-level scheme. In particular, we describe a computationally efficient Huffman-asymmetric-coding scheme for distributions \( \mu \) in the class
\[
\mathcal{M} := \{ \mu \in \Delta([N]) : 1/4N \leq \mu(i) < 1/2 \forall i \in [N] \}.
\]
This restriction (from a general distribution) is without loss of generality because we can transform any distribution into a distribution in \( \mathcal{M} \) (up to a loss of at most factor 2 in the communication bound). First, if there’s an item with probability more than half, we can store its presence with just one bit. 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 \( \mu'(i) = \max\{\mu(i), 1/2N\} \), (2) let \( \zeta = \sum_i \mu'(i) \leq \sum_i (\mu(i) + 1/2N) = 1.5 \), and (3) setting \( \mu''(i) = \frac{1}{\zeta} \mu'(i) \). Now note that \( 1/2 \leq \frac{\mu''(i)}{\max(1/2N, \mu(i))} \leq 1 \), and hence \( \lg 1/\mu''(i) \leq \lg 1/\mu(i) + 1 \leq 2 \cdot \lg 1/\mu(i) \). We also assume that \( m \geq \lg N + 4 \).

Our scheme \( \mathcal{A} = (\text{Enc}, \text{Dec}) \) 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
\[
h_t : [N] \rightarrow [D_t]
\]
where the hash functions are determined using shared public randomness. The scheme also uses a family of \( T \) (deterministic) linear codes, \( C^{(t)} = [c_1^{(t)} \ c_2^{(t)} \ \ldots \ c_{D_t}^{(t)}] \in \mathbb{F}_N^{m_t \times D_t} \) for \( t \in [T] \), which are
specified in the next subsection. Each matrix $C^{(t)}$ shall be designed to support efficient decoding of every $\left(\frac{m}{2^{l_t} D_t}\right)$-sparse vector. We now turn to the formal construction, which is adapted from the work of [DV13a].

4.1 The Sensing Matrices $C^{(t)}$

The basic building block of our scheme are the compressed-sensing matrices designed in the work of [DV13a]. 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 [DV13a] that is used in our multi-level scheme. 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., [Rot06]). 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 4.1 (Efficient deterministic compressed sensing, [DV13a]).** Let $P_k^N \in \mathbb{F}_N^{m \times N}$ be the parity-check matrix of a $[N, N-2k, 2k+1]_{\mathbb{F}_N}$ Reed-Solomon code\(^2\), where $m = 2k\lfloor \lg N \rfloor$. 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 $P_k^N \cdot x$ using $O(Nk2^{\lg^2 N})$ operations over $\mathbb{F}_2$. In particular, $P_k^N \cdot x$ uniquely determines $x$ using $m = 2k\lfloor \lg N \rfloor$ 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 [Rot06]) which recovers the noise vector (i.e., our desired $k$-sparse subset) from the parity check matrix $P_k^N$.

Of course, the main difference from the setup of Theorem 4.1 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 [DV13a] 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 $P_k^N$ with dimensions $N := D_t$, $k := m_t/(2 \lg D_t)$. This will become clearer in the next section where we present the entire multi-level scheme.

4.2 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 in the analysis. 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 following theorem.

---

\(^2\)We assume here that $N$ is a power of 2. Otherwise, replace it with $N' := 2^\lceil \lg N \rceil$. 

Theorem 4.2. Pick any $\delta \in (0,1)$ and positive integer $m^*$. Set
\begin{equation}
D_t := \left\lfloor \frac{T}{\delta} \cdot \left( 2^{2t+1} - 1 + \frac{\log^2(4N)}{2t} \right) \right\rfloor, \quad t \in [T],
\end{equation}
and
\begin{equation}
m_t := \left\lfloor \frac{2\log D_t}{2^{t-1}} \cdot m^* \right\rfloor, \quad t \in [T].
\end{equation}
Then for any $\mu \in \mathcal{M}$ and $S$ satisfying Eqn. (1),
\[ \Pr(\hat{S} = S) \geq 1 - \delta. \]

To see that Theorem 4.2 implies Theorem 1.5, observe that the message length $m = \sum_{t=1}^{T} m_t$ satisfies
\[ \sum_{t=1}^{T} m_t = m^* \cdot O \left( \log(\log(N)/\delta) \right). \]
Using Theorem 1.1 it is also clear that the running times of Algorithm 1 and Algorithm 2 are \text{poly}(N).

Proof of Theorem 4.2. Fix $\mu \in \mathcal{M}$ and $S$ satisfying Eqn. (1). Because every $i \in S$ satisfies $\log(1/\mu(i)) \leq \log(4N)$, 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:

**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 $P^{N_t}_{k_t}$ from Theorem 4.1 instantiated with $N_t := D_t, k_t := m_t/(2 \log D_t)$. 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$_\mu$ 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 $B_t := \{ i \in [N] : 2^{-2^t} \leq \mu(i) < 2^{-2^t-1} \}$ for $t \in [T]$.
2: Initialize $\hat{S} := \emptyset$.
3: for $t = 1, 2, \ldots, T$ do
4: Let $\hat{y}^{(t)}$ be the output of the decoder for $C^{(t)}$ applied to $y^{(t)}$, guaranteed by Theorem 4.1.
5: for each $i \in B_t$ do
6: if $\hat{y}^{(t)}_{h_t(i)} = 1$ 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 $\hat{S}$
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 \( \mu(i) \geq 2^{-2^t} \), so \( |B_t| \leq 2^{2^t} \). Furthermore, since every \( i \in S_{t:T} \) satisfies \( \mu(i) \leq 2^{-2^{t-1}} \), it holds that

\[
|S_{t:T}| \leq \sum_{i \in S_{t:T}} \frac{\lg(1/\mu(i))}{2^{t-1}} \leq \frac{\lg(4N)}{2^{t-1}},
\]

and

\[
|S_t| \cdot |S_{t+1:T}| \leq \frac{1}{4} \cdot |S_{t:T}|^2 \leq \frac{\lg^2(4N)}{2^{2t}}.
\]

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

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

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 \( S_t \) 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 \( 1_{S_{t:T}} \)), 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 \( \text{supp}(\hat{z}(t)) \), where

\[
\hat{z}(t) := \sum_{i \in S} e_{h_t(i)}.
\]

The vector \( \hat{z}(t) \) 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 \( \hat{z}(t) \) has at most \( |S_{t:T}| \leq m_t^*/2^{t-1} \) 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., \( y(t) = C(t) \hat{z}(t) \). Due to the choice of \( m_t \) from Eqn. (3) and Theorem 4.1, the decoding of \( y(t) \) returns \( \hat{z}(t) = z(t) \) as required.

\[
\square
\]

References

[ADHP06] Micah Adler, Erik D. Demaine, Nicholas J.A. Harvey, and Mihai Pătraşcu. Lower bounds for asymmetric communication channels and distributed source coding. In Proc. 17th ACM/SIAM Symposium on Discrete Algorithms (SODA), pages 251–260, 2006.
[AIM10] Luigi Atzori, Antonio Iera, and Giacomo Morabito. The internet of things: A survey. *Computer networks*, 54(15):2787–2805, 2010.

[AM98] Micah Adler and Bruce M Maggs. Protocols for asymmetric communication channels. In *Foundations of Computer Science, 1998. Proceedings. 39th Annual Symposium on*, pages 522–533. IEEE, 1998.

[BCDH10] Richard G Baraniuk, Volkan Cevher, Marco F Duarte, and Chinmay Hegde. Model-based compressive sensing. *IEEE Transactions on Information Theory*, 56(4):1982–2001, 2010.

[BGW11] Michael Büttner, Ben Greenstein, and David Wetherall. Dewdrop: an energy-aware runtime for computational rfid. In *Proc. USENIX NSDI*, pages 197–210, 2011.

[CGRZ16] Tingjun Chen, Javad Ghaderi, Dan Rubenstein, and Gil Zussman. Maximizing broadcast throughput under ultra-low-power constraints. In *Proceedings of the 12th International on Conference on emerging Networking EXperiments and Technologies*, pages 457–471. ACM, 2016.

[CSV03] Giuseppe Caire, Shlomo Shamai, and Sergio 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*, pages 263–284, 2003.

[CT06] E. Candes and T. Tao. Near optimal signal recovery from random projections: Universal encoding strategies. *IEEE Transactions on Information Theory*, 2006.

[DM09] Stark C Draper and Sheida Malekpour. Compressed sensing over finite fields. In *Proceedings of the 2009 IEEE international conference on Symposium on Information Theory-Volume 1*, pages 669–673. IEEE Press, 2009.

[Don06] D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[DV13a] Abhik Kumar Das and Sriram 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*, pages 5890–5894, 2013.

[DV13b] Amal K Das and Sriram Vishwanath. On finite alphabet compressive sensing. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 5890–5894. IEEE, 2013.

[GI10] Anna Gilbert and Piotr Indyk. Sparse recovery using sparse matrices. *Proceedings of the IEEE*, 6(98):937–947, 2010.

[GKK+09] Maria Gorlatova, Peter Kinget, Ioannis Kymissis, Dan Rubenstein, Xiaodong Wang, and Gil Zussman. Challenge: ultra-low-power energy-harvesting active networked tags (enhants). In *Proceedings of the 15th annual international conference on Mobile computing and networking*, pages 253–260. ACM, 2009.

[GLPS12] Anna C Gilbert, Yi Li, Ely Porat, and Martin J Strauss. Approximate sparse recovery: optimizing time and measurements. *SIAM Journal on Computing*, 41(2):436–453, 2012.
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 A.1. If $A$ is a list-asymmetric-coding scheme with parameters $m_1^*$ and $\delta$, then $A$ is a Huffman-asymmetric-coding scheme with parameters $m_h^* \leq m_1^* - \log e$ and $\delta$, and the same, fixed communication bound $m$. 

14
Proof. Consider any distribution \( \mu \) over \( [N] \). Let \( L \) be the list of subsets \( S \subseteq [N] \) that satisfy Eqn. 11. We just need to show that the size of \( L \) is less than \( e^{2m^*} \). A set \( S \) satisfies Eqn. 11 if and only if

\[
\prod_{i \in S} \mu(i) \geq 2^{-m^*}.
\]

On the other hand

\[
\sum_{S \in L} \prod_{i \in S} \mu(i) \leq \sum_{S \subseteq [N]} \prod_{i \in S} \mu(i).
\]

\[
= \sum_{(x_1, \ldots, x_N) \in \{0,1\}^N} \prod_{i=1}^N \mu(i)^{x_i}
\]

\[
= \sum_{x_1 \in \{0,1\}} \mu(1)^{x_1} \sum_{x_2 \in \{0,1\}} \mu(2)^{x_2} \cdots \sum_{x_N \in \{0,1\}} \mu(N)^{x_N}
\]

\[
= (1 + \mu(1))(1 + \mu(2)) \cdots (1 + \mu(N))
\]

\[
\leq e^{\mu(1)}e^{\mu(2)} \cdots e^{\mu(N)}
\]

\[= e. \]

Hence the size of list \( L \) is less than \( e^{2m^*} \) and a list-asymmetric-coding scheme for list \( L \), with parameters \( m^* \) and \( \delta \), yields an error probability \( \delta \). \( \square \)

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 \( \sigma \) as

\[
C(\sigma, \delta) := \min_{m \in \mathbb{N}} \{ \exists L \subseteq \text{supp}(\sigma), |L| \leq 2^m, \sigma(L) \geq 1 - \delta \}.
\]

The following claim asserts an upper bound on the cover number in terms of the Shannon entropy of \( \sigma \).

**Claim A.2 (Cover-size vs. Entropy).** For every distribution \( \sigma \) and \( \delta > 0 \), it holds that

\[
C(\sigma, \delta) \leq H(\sigma)/\delta.
\]

We remark that the bound is essentially tight, as demonstrated by the distribution \( \sigma \) which has an “atom” of measure \( \delta \) and otherwise uniform on the entire domain.

**Proof.** Let \( \mathcal{G}_\delta := \{ x : \log(1/\sigma(x)) \leq H(\sigma)/\delta \} \) be the set of elements with “large” mass under \( \sigma \). Indeed, note that \( \forall x \in \mathcal{G}_\delta \) we have \( \sigma(x) \geq 2^{-H(\sigma)/\delta} \), thus it holds that \( |\mathcal{G}_\delta| \leq 2^{H(\sigma)/\delta} \). In order to conclude that \( C(\sigma, \delta) \leq H(\sigma)/\delta \), it remains to show that \( \sigma(\mathcal{G}_\delta) \geq 1 - \delta \). Indeed, Markov’s inequality implies that

\[
\sigma(\mathcal{G}_\delta) = 1 - \sigma(\overline{\mathcal{G}_\delta}) = 1 - \Pr_{x \sim \sigma} \left( \frac{1}{\sigma(x)} > \frac{H(\sigma)}{\delta} \right) = 1 - \Pr_{x \sim \sigma} \left( \log \frac{1}{\sigma(x)} > \frac{H(\sigma)}{\delta} \right) \geq 1 - \delta.
\]

\( \square \)

The following is a corollary of Claim A.2.
Claim A.3. If $A$ is a list-asymmetric-coding scheme with parameters $m^*_i$ and $\delta_i$, then $A$ can be converted into an entropy-asymmetric-coding scheme with parameters $m^*_e := \delta_i m^*_i$ and $\delta_e := 2\delta_i$ (and same, fixed communication bound $m$).

Proof. For any prior $\sigma$ on subsets of $[N]$, there is a list $L = L(\sigma)$ of size at most $2^{H(\sigma)/\delta_i}$ which is “responsible” to $1 - \delta_i$ mass of the distribution. So, when the encoding length is fixed to $m$, Claim A.2 guarantees that decoding (w.p. $1 - \delta_i$) all subsets with $\sigma(S) \geq 2^{-m^*_i}$ is equivalent to decoding (w.p. $1 - \delta_i$) all distributions with Shannon entropy at most $\delta_i m^*_i$. \hfill $\square$

Note that $\delta_i m^*_i$ 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 A.4. If $A$ is a Huffman-asymmetric-coding scheme with parameters $m^*_h$ and $\delta_h$, then for any $\epsilon \in (0, 1)$, $A$ is an entropy-asymmetric-coding scheme with parameters $m^*_e := \frac{1 - \delta_h/(2N)}{1 + \epsilon} \left( m^*_h - \left( \frac{1}{2\epsilon} + \frac{1}{3} \right) \log(2N^2/\delta_h) \ln(2/\delta_h) \right)$, $\delta_e := 2\delta_h$, and same, fixed communication bound $m$.

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

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

$$kH_E(\mu) := \mathbb{E}_{S \sim \mu^k} \left[ \sum_{i \in S} \log(1/\mu(i)) \mid E \right] = \frac{k}{1 - \mu(\text{Tail})} \sum_{i \in \text{Head}} \mu(i) \log(1/\mu(i)) \leq \frac{k}{1 - \delta_h/(2N)} H(\mu).$$

By Bernstein’s inequality, we have

$$\Pr_{S \sim \mu^k} \left( \sum_{i \in S} \log(1/\mu(i)) \leq kH_E(\mu) + \sqrt{2kH_E(\mu) \log \left( \frac{2N^2}{\delta_h} \right) \ln \left( \frac{2}{\delta_h} \right)} + \frac{\log(2N^2/\delta_h) \ln(2/\delta_h)}{3} \mid E \right) \geq 1 - \frac{\delta_h}{2}.$$

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

$$\sum_{i \in S} \log(1/\mu(i)) \leq \frac{kH(\mu)}{1 - \delta_h/(2N)} + \frac{2kH(\mu) \log(2N^2/\delta_h) \ln(2/\delta_h)}{1 - \delta_h/(2N)} + \frac{\log(2N^2/\delta_h) \ln(2/\delta_h)}{3} \leq \frac{1 + \epsilon}{1 - \delta_h/(2N)} kH(\mu) + \left( \frac{1}{2\epsilon} + \frac{1}{3} \right) \log(2N^2/\delta_h) \ln(2/\delta_h) \leq m^*_h$$

As mentioned before, this “truncation” of the tail of $\sigma$ seems inherent to oblivious schemes, as they are fixed-length encodings.
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, $A$ correctly decodes the set $S$ with probability at least $1 - \delta_h$. Thus, $A$ is an entropy-asymmetric-coding scheme with parameters $m_e^*$ and $\delta_e = 2\delta_h$. \qed