Deterministic Compression with Uncertain Priors

[Extended Abstract] *

Elad Haramaty†
Department of Computer Science
Technion, Haifa.
eladh@cs.technion.ac.il.

Madhu Sudan
Microsoft Research
1 Memorial Drive, Cambridge, MA 02142.
madhu@mit.edu

ABSTRACT

Communication in "natural" settings, e.g., between humans, is distinctly different than that in classical designed settings, in that the former is characterized by the sender and receiver not being in perfect agreement with each other. Solutions to classical communication problems thus have to overcome an extra layer of uncertainty introduced by this lack of prior agreement. One of the classical goals of communication is compression of information, and in this context lack of agreement implies that sender and receiver may not agree on the "prior" from which information is being generated. Most classical mechanisms for compressing turn out to be non-robust when sender and receiver do not agree on the prior. Juba et al. (Proc. ITCS 2011) showed that there do exists compression schemes with *shared randomness* between sender and reciever that can compress information down roughly to its entropy.

In this work we explore the assumption of *shared* randomness between the sender and receiver and highlight why this assumption is problematic when dealing with natural communication. We initiate the study of deterministic compression schemes amid uncertain priors, and expose some the mathematical facets of this problem. We show some non-trivial determinstic compression schemes, and some lower bounds on natural classes of compression schemes. We show that a full understanding of deterministic communication turns into challenging (open) questions in graph theory and communication complexity.

Categories and Subject Descriptors

E.4 [**Coding And Information Theory**]: Data compaction and compression

*A full version of this paper is available as *ECCC* Technical Report TR 12-166, November 26, 2012.

†Work done in part when this author was visiting Microsoft Research New England.

Keywords

Communication complexity, Graph coloring, Source coding

1. INTRODUCTION

In this work we consider the task of compressing information deterministically, in settings where the sender and receiver are not in agreement on the distribution from which the information is being generated. We start by first describing the general motivation for the study of this problem before formally describing the problem and our results.

1.1 Natural Communication: Context and Uncertainty

Natural communication, say between two humans, differs in significant ways from "designed communication", say between a cell phone and its nearby tower. The latter are carefully engineered to optimize use of the channel of communication, while introducing careful redundancy to overcome any unreliability of the channel. The resulting problems and solution concepts, such as compression schemes and error-correcting encoding schemes are well understood by now.

Natural communication, as characterized by the vagaries of natural language, is much less understood mathematically. Natural communication, often characterized by dictionaries and grammars, do not follow the rules they prescribe. They tend to be ambiguous locally, and seemingly needlessly redundant in other cases; without offering the same reliability as error-correcting codes do. At the same time, natural communication has remarkable capacity to overcome lack of "perfect engineering" of the sender or receiver of information, and in particular do not seem to require perfect agreement between then two (on the design of the protocol). Understanding this resilience mathematically would be a fruitful pursuit and forms the motivation for this and prior works in this stream.

The main goal here is to understand what does "language" look like and what parameters does it try to optimize. Roughly language takes an intended message in the sender's mind and attempts to describe how to convert this message to a sequence of, say, bits. This part is similar to the encoding map of a error-correcting scheme, However to model language more precisely, one should really take into account its context-sensitivity. A more precise description of a language would be as an *encoding* map from a pair *(context,message)* to a *word* which is a sequence of bits. For the time being, one may view context and message as elements of some abstract space (each being possibly a countable set). The language also gives rules on how a receiver should re-

cover the message from the word: it does so by applying a *decoding* map that maps a pair *(context,word)* to a *message*. The goal hopefully is to make sure the receiver decodes to a message that is somehow *compatible* with the sender's intent. Defining compatibility is complex, and we will skirt this issue in this paper. (Such issues are considered, e.g., in [8, 4].) We will simply require receiver to decode to the *same* message as the sender wished to send (so in particular the message space for sender and receiver are identical). Even with this simplication, there remains a major hurdle to communication, namely that sender's context and receiver's context may not be identical! Remarkably, natural communication manages to function reasonably well even when these contexts are not identical, but reasonably close to each other, and it is this aspect that is the focus of this work.

In order to model contexts that are not identical, but are reasonably close, we need a more structured view of contexts (than merely as elements of an abstract set). Juba et al. [7] proposed a natural view: namely a context is simply a probability distribution on messages: The distribution that describes the message that the receiver is expecting to receive, or the distribution that the sender thinks the receiver is expecting to receive. In this framework, the encoding function becomes a map from a distribution P supported on some set U and message $m \in U$ to a word $w \in \{0,1\}^*$; and the decoding function becomes a map from a distribution Q also supported on U and word $w \in \{0,1\}^*$ to a message $\hat{m} \in U$ and the goal of the language is to ensure that $\hat{m} = m$ even if $P \neq Q$, provided they are reasonably close. At the same time the language would like to reduce the expected length of the communication, assuming say that the messages are generated from the distribution P. (More generally we could consider a setting where the messages are generated from some distribution R; our assumption that $R = P$ is mostly for simplicity.)

Classical communication dealt with the setting where $P = Q$. In the classical setting, it turns out that the entropy of the distribution P precisely describes the expected length of the transmission. Does entropy still give a good measure characterizing the expected length of the transmission in the natural setting, when $P \neq Q$? (We note that classical communication also considered the setting where $R \neq P = Q$ and this leads to notions such the KL-divergence, but the important aspect is that they did not consider settings where sender and receiver disagreed on the prior.)

Juba et al. consider the following notion of distance between distributions (a notion which is also used commonly in differential privacy): $\delta(P,Q) = max_{m \in U}\{|\log_2 P(m) - \log_2 Q(m)|\}$. (If $\delta(P,Q) \leq \Delta$ then for every $m \in U$ we have $2^{-\Delta}Q(m) \leq P(m) \leq 2^{\Delta}Q(m)$.) With this notion of distance in place, Juba et al. roughly show that entropy is a good measure of the compression length: Specifically they show that there exist encoding and decoding schemes that use shared randomness between sender and receiver and manage to compress information to a length of roughly $H(P)+2\delta(P,Q)$ where $H(\cdot)$ denotes the binary entropy function. (We note that works of Harsha et al. [6] and Braverman and Rao [1] also explore questions somewhat similar to the ones considered by Juba et al., though their motivations were quite different. Both works focus on the setting when sender and receiver have different priors and are trying to generate a random variable that is maximally correlated under their priors. In our case the sender gets a concrete message from its prior and wishes to communicate it. The focus in both works is on randomized solutions that get the communication complexity down to the minimum possible amount, whereas our thrust is to use less (or no) randomness at the expense of slightly larger communication complexity.)

The main goal of this paper is to explore the need for shared randomness in their work. The assumption of shared randomness takes the solution further away from the motivation of natural communication. Roughly their solution suggests that if the "dictionary", or any other codebook associated with language, is random, then information can be compressed. However in natural usage, the dictionary remains somewhat static whereas the contexts for communication vary vastly. Furthermore, it is even very plausible that the dictionary influences our communication and its context. We thus feel that a compression scheme based on shared randomness is not sufficient evidence to suggest that entropy is the natural complexity measure of capturing the complexity of communicating in natural settings. This leads us to study deterministic compression schemes in this paper, and as we stress below, this turns out to be surprisingly challenging to analyze. We illustrate the complexity by considering a rather simple problem in this space.

1.2 A toy problem

The following example illustrates the questions studied in this paper: Suppose Alice and Bob have a ranking of a set U of N elements, say, movies. Specifically Alice's rank function is $A : [N] \to U$ and Bob's rank function is $B : [N] \to U$ where $[N] = \{1, \ldots, N\}$ and A and B are bijections with $A(i)$ naming the ith ranked movie in Alice's ranking. Suppose further that Alice and Bob know that their rankings are "close", specifically for every $x \in U$, $|A^{-1}(x) - B^{-1}(x)| \leq 2$. How many bits does Alice have to send to Bob so that Bob knows her top-ranked movie, i.e., $A(1)$?

On the one hand Bob knows $A(1)$ is one of the three element set $S_1 = \{B(1), B(2), B(3)\}$ and so the information-content from his point of view is bounded by $\log_2 3$ bits. Indeed this leads to a randomized communication scheme, with Alice and Bob sharing common randomness with $O(1)$ bits of communication: Alice and Bob use their randomness to get a hash function hashing their universe to a constant number of bits. Alice sends the hash of $A(1)$, and Bob recovers the name of the movie provided the elements of S_1 hash to distinct values. However the deterministic communication complexity of the question is not as easily settled. Part of the reason is that Alice doesn't know S_1 and so has to "guess" it to communicate $A(1)$. Still she is not clueless: She knows it is contained in $T_2 = \{A(1), \ldots, A(5)\}$ and perhaps this can help her communicate $A(1)$ efficiently to Bob. The question of interest to us in this work is: Can Alice communicate $A(1)$ to Bob with a number of bits that is independent of N? (Unfortunately, we do not answer this question, though we do give a non-trivial upper bound. We will elaborate on this later.)

The question above is a prototypical example of "communication amid uncertainty", where the communicating players have fairly good information about each other (in the example above Alice and Bob know each others ranking of each movie to within ± 2), but are not sure of each other's information and do not have a common-ground to base communication on. As we elaborate on in this paper, solutions to this problem influence solutions to the general problems

of communication amid uncertainty, while this problem is itself a special case (when Alice and Bob's distributions are geometric).

We describe our problems and solutions shortly, but to give a gist of the findings: We show in this work that there is a solution to the above toy problem communicating roughly $2^{O(\log^* N)}$ bits, which is a very slowly growing, but nevertheless growing, function of U. We are unable to resolve if a growing function of N bits is necessary for this problem; however we do show that any solution that looks at only a constant number of Alice's top movies (or a constant number of Bob's top movies) must communicate $\log^{(\Omega(1))} N$ bits (where $\log^{(i)}$ denotes the ith iterated logarithm function.

1.3 Formal definitions and main results

We start by defining the notion of an "uncertain compression scheme".

We let $\{0,1\}^*$ denote the set of all finite length binary strings. For $x \in \{0,1\}^*$, let $|x|$ denote its length. Throughout U, the set of all messages, will be a finite set of size N. Let $\mathcal{P}(U)$ denote the space of all probability distributions over U.

Definition 1.1. ((Basic) Uncertain Compression Scheme) *For positive real Δ an* Uncertain Compression Scheme (UCS) *for distance Δ over the universe U is given by a pair of $E : \mathcal{P}(U) \times U \to \{0,1\}^*$ and $D : \mathcal{P}(U) \times \{0,1\}^* \to U$ that satisfy the following correctness condition: For every pair of distributions $P, Q \in \mathcal{P}(U)$ that are Δ-close (i.e $\delta(P,Q) \leq \Delta$) and for every $m \in U$, we have $D(Q, E(P,m)) = m$. The* performance *of a UCS (E,D) is given by the function $L : \mathcal{P}(U) \to \Re^+$, where $L(P) = \mathbf{E}_{m \leftarrow_P U}[|E(P,m)|]$, i.e., the expected length of the encoding under the distribution P. We refer to such a scheme as a (Δ, L)-UCS.*

In English, the definition above explicitly provides the distribution as input to the encoding and decoding schemes, and expect the schemes to work correctly even if the distributions used by the encoder and decoder are not the same, as long as they are Δ-close to each other. While in general we would like compression schemes which work for all possible distributions P, Q that are within Δ of each other, and with no error (as expected in the definition above), some of our schemes are weaker and work with some error, or only for some class of distributions. We define such general UCS's below.

Definition 1.2. ((General) Uncertain Compression Scheme) *For positive real Δ (for distance), $\epsilon \in [0,1]$ (for error), a class of distributions $\mathcal{F} \subseteq \mathcal{P}$, and performance function $L : \mathcal{F} \to \Re^+$ a* $(\Delta, \epsilon, \mathcal{F}, L)$-Uncertain Compression Scheme (UCS) *over the universe U is given by a pair of $E : \mathcal{F} \times U \to \{0,1\}^* \cup \{\perp\}$ and $D : \mathcal{F} \times \{0,1\}^* \cup \{\perp\} \to U \cup \{\perp\}$ that satisfy the following conditions:*

1. *For every pair of distributions $P, Q \in \mathcal{F}$ that are Δ-close and for every $m \in U$, it is the case that if $E(P,m) \neq \perp$ then $D(Q, E(P,m)) = m$. Furthermore $D(\perp) = \perp$.*
2. *$\Pr_{m \leftarrow_P U}[E(P,m) = \perp] \leq \epsilon$.*
3. *For every $P \in \mathcal{F}$, we have $\mathbf{E}_{m \leftarrow_P U}[|E(P,m)|] \leq L(P)$.*

Note that we do not distinguish the two definitions above by name, but rather just by the number of parameters. So if the number of parameters is just two, then it is assumed that there is no error, and the performance holds for all distributions.

We note that the definitions above only cover deterministic compression schemes. A compression scheme with shared randomness can be defined analogously, but we don't do so here. We also stress that the choice of P and Q is "worst-case" within the family $\mathcal{F}$ (as formalized by the universal quantifier in the correctness condition). There are no assumptions that $\mathcal{F}$ is small (has only finitely many elements), which tends to be the setting for universal compression. Similarly, we do not consider a sequence of messages that need to be transmitted: Rather, we are considering one-shot communication with no assumptions on the distributions P and Q, other than that they are from $\mathcal{F}$ and Δ-close.

We recall that Juba et al. present a $(\Delta, H(P) + 2\Delta + c)$-UCS (with shared randomness) for some constant $c \leq 3$. We give two *deterministic* schemes in this paper, both having complexity depending on N, but both using substantially less than $\log N$ bits.

Theorem 1.3. *For every $\Delta \geq 0$, there exists a $(\Delta, O(H(P)+\Delta+\log\log N))$-UCS, i.e., a deterministic universal compression scheme that works for all pairs P, Q that are within distance Δ of each other, and where the expected length of encoding is at most $O(H(P) + \Delta + \log\log N)$.*

The dependence on N of this scheme is non-trivial and thus may even be reasonable in "natural circumstances". However it is not clear if such a dependency on N is necessary. Motivated by the quest to understand the dependence on N more closely, we explore schemes whose performance is not necessarily linear in $H(P)$. Simultaneously we relax our schemes to allow them to "drop" messages with ϵ probability. We note that if we don't do the latter, then the former is not really a relaxation: Any error-free scheme with superlinear dependence on $H(P)$ can be converted to one with linear dependence on $H(P)$ by a simple reduction (see Lemma 3.11).

Our next theorem gives a scheme that is weaker than the one from Theorem 1.3 in its dependence on the entropy $H(P)$ and in that it errs with non-zero probability. But it does achieve significantly better dependence on N.

Theorem 1.4. *For every $\epsilon > 0$ and $\Delta \geq 0$ there exists a $(\Delta, \epsilon, \mathcal{P}(U), \exp(H(P)/\epsilon + \Delta \log^* N))$-UCS, i.e., the scheme has error probability at most ϵ, it works for all pairs of distributions P, Q within distance Δ and the expected length of the encoding is at most $\exp(H(P)/\epsilon + \Delta \log^* N)$.*

In the above the notation $\exp(x)$ denotes a function of the form c^x for some universal constant c, and $\log^* N$ denotes the minimum integer i such $\log^{(i)} N \leq 1$ and $\log^{(i)}$ is the logarithm function iterated i times.

An alternate way to get around the barrier of Lemma 3.11, which insists that schemes must have linear dependence on $H(P)$ or make some error, is to have schemes that do not work for all possible pairs of distributions P and Q. As it turns out the scheme from Theorem 1.4 does have this behavior for many natural distributions. In Theorem 3.8 we show that our scheme from Theorem 1.4 works without error and with same performance as long as P (or Q) are close to

a "flat distribution" (uniform over a subset), or a geometric distribution, or a binomial distribution. We stress that the scheme is not particularly carefully tailored to the class of distributions (though of course the encodings and decodings do depend on the distributions), but naturally adapts to being error-free for the above classes.

1.4 Techniques: Graph Coloring

While the most natural framework for studying our problem is as a question of communication complexity of a relational problem (as in [9]), this turns out not to be the most useful for studying the deterministic communication complexity. Indeed, as pointed out earlier, the modern stress in communication complexity is often on designing and understanding the limits of protocols that are interactive and use shared randomness, while in our case the thrust is in the opposite direction.

It turns out our questions are naturally also captured as graph-coloring questions. Furthermore such questions (or related ones) have been studied in the literature on distributed computing in the attempt to color graphs in a local distributed manner. In particular, the work of Linial [10] shows that a "local" algorithm for 3-coloring a cycle, due to Cole and Vishkin [2], implies that a large "high-degree graph" is 3-colorable. The ideas of Cole and Vishkin [2] and Linial [10] turn out to be quite useful in our context. Our work abstracts some of these techniques, and extends them to get combinatorial results, which we then convert to efficient compression schemes.

Uncertainty graphs and Chromatic number.

We start by defining a class of structured combinatorial graphs whose chromatic number turns out to be central to our problems. Let $[N] = \{1, \ldots, N\}$. Let S_N denote the set of all permutations on N elements, i.e., the set of all bijections from $[N]$ to itself. For $\pi, \sigma \in S_N$, let $\delta(\pi, \sigma) = \max_{i \in [N]} |\pi^{-1}(i) - \sigma^{-1}(i)|$.

DEFINITION 1.5 (UNCERTAINTY GRAPHS). *For integer N, ℓ the uncertainty graph $\mathcal{U}_{N,\ell}$ has as elements of S_N as its vertices, with $\pi \leftrightarrow \sigma$ if (1) $\pi(1) \neq \sigma(1)$ and $\delta(\pi, \sigma) \leq \ell$.*

It turns out that the chromatic number of the uncertainty graphs have a close connection to uncertain communication schemes. Roughly these graphs emerge from a very restricted version of the communication problem, where the distributions P and Q are geometric distributions (giving probability proportional to $\beta^{-\pi^{-1}(i)}$ and $\beta^{-\sigma^{-1}(i)}$ to the element $i \in [N]$). It follows that if $\delta(\pi, \sigma)$ is small, then P and Q are close to each other. Furthermore, for simplicity these graphs only consider the case that the message is the element with maximal probability under P. To understand how the chromatic number plays a role, fix a receiver with distribution Q and consider two possible senders P and P' that could communicate with this receiver. Consider coloring P and P' by $E(P, \text{argmax}_m\{P(m)\})$ and $E(P', \text{argmax}_m\{P'(m)\})$ respectively. This would lead to distinct colors on pairs P and P' that are too close to each other, provided their messages, i.e., $\text{argmax}_m\{P(m)\}$ and $\text{argmax}_m\{P'(m)\}$ are different. This exactly corresponds to adjacency in our graph: the underlying permutations π and σ are close, and the top ranked elements are different.

The results of Juba et al. imply that the "fractional chromatic number" of $\mathcal{U}_{N,\ell}$ is bounded by $O(\ell)$.[1] The (integral) chromatic number on the other hand does not immediately seem to be bounded as a function of ℓ alone. The implication of the low fractional chromatic number is that the chromatic number of $\mathcal{U}_{N,\ell}$ is at most $O(\ell N \log N)$, but this is worse that the naive upper bound of N, which can be obtained by setting the color of π to be $\pi^{-1}(1)$. (By definition of adjacency this is a valid coloring.) Our main technical contribution is in obtaining some non-trivial upper bounds on the chromatic number of this graph.

To derive our upper bounds, we look at "coarsened" versions of the graph $\mathcal{U}_{N,\ell}$. For positive integer k, we say that $\pi : [k] \to [N]$ is a *k-subpermutation* if π is injective. We let $S_{N,k}$ denoted the set of all k-subpermutations on $[N]$. For $k' \geq k$, we say subpermutation $\pi : [k] \to [N]$ *extends* the subpermutation $\sigma : [k'] \to [N]$ if $\sigma(i) = \pi(i)$ for all $i \in [k]$. For k-subpermutations π and σ, we let $\delta(\pi, \sigma) = \min_{\pi', \sigma' \in S_N \text{ extending } \pi, \sigma}\{\delta(\pi', \sigma')\}$.

DEFINITION 1.6 (RESTRICTED UNCERTAINTY GRAPHS). *For integers N, ℓ and k the* k-restricted uncertainty graph *$\mathcal{U}_{N,\ell,k}$ has elements of $S_{N,k}$ as its vertices, with $\pi \leftarrow \sigma$ if (1) $\pi(1) \neq \sigma(1)$ and $\delta(\pi, \sigma) \leq \ell$.*

Note that $\mathcal{U}_{N,\ell,N} = \mathcal{U}_{N,\ell}$. We derive our upper bounds on the chromatic number of $\mathcal{U}_{N,\ell}$ by giving non-trivial upper bounds on the chromatic number of $\mathcal{U}_{N,\ell,k}$.

LEMMA 1.7.
1. *For every $k \leq k'$, $\chi(\mathcal{U}_{N,\ell,k'}) \leq \chi(\mathcal{U}_{N,\ell,k})$.*
2. *For every N, ℓ, $\chi(\mathcal{U}_{N,\ell,2\ell}) \leq O(\ell^2 \log N)$.*
3. *For every N, ℓ and k that is an integral multiple of ℓ, we have $\chi(\mathcal{U}_{N,\ell,k}) \leq 2^{O(k \log \ell)} \log^{(k/\ell)} N$.*
4. *For every N, ℓ and k that is an integral multiple of ℓ, we have $\chi(\mathcal{U}_{N,\ell,k}) \geq \log^{(2k/\ell)}(N/\ell)$.*

As an immediate application we get the following theorem.

THEOREM 1.8. *For every N and ℓ, we have $\chi(\mathcal{U}_{N,\ell}) \leq O\left(\min\{\ell^2 \log N, 2^{O(\ell \log \ell \log^* N)}\}\right)$.*

Unfortunately, the lower bound from Part (4) of Lemma 1.7 goes to 0 as $k \to N$ and so we don't get a growing function of N as a lower bound. However, it does rule out most natural strategies for coloring $\mathcal{U}$, and shows limitations of the *intuition* that suggests $\mathcal{U}$ may be colorable with $f(\ell)$ colors independent of N. This is so since the intuition, as well most natural strategies, only use the top $O(\ell)$ ranking elements of a permutation π to determine its color; and such the lower bound shows that such strategies are inherently limited. In particular, it shows that there is no hope to extend the methods of Juba et al. (which was based on this intuition) in a simple way to get a deterministic UCS.

[1] The fractional chromatic number of a graph G is the smallest positive real w such that there exists a collection of independent sets $I_1, \ldots, I_t$ in G with weights $w_1, \ldots, w_t$ such that $\sum_{j=1}^{t} w_j = w$ and for every vertex $u \in V(G)$ it is the case that $\sum_{j: I_j \ni u} w_j \geq 1$.

1.5 Directions for further work

Given that most of the questions raised in this work haven't found tight answers, there is an obvious number of natural questions to resolve here — the most fundamental one being whether one can compress information down to its entropy (to within constant multiplicative factors) deterministically (or even just with private randomness) in the uncertain setting.

In addition to resolving open questions, a number of modelling challenges remain in trying to understand natural modes of communication. Language is an organically evolved concept with communicational, computational and societal pressures acting on it. The game that was played out with natural languages over the past millenia is now getting played out at a faster pace among computer networks: Protocols evolve, compete for survival, and develop a strange mix of tolerance for errors with intolerance for others. Most of the mechanics have not been studied mathematically and indeed little is known as to what the evolution process is trying to achieve, and what the steady state might look like, if at all one exists. Understanding aspects of language and its evolution definitely seem to be worthy causes.

One aspect in particular that we have not explored is the impact of "computational efficiency" of the encoding or decoding procedures. One of the reasons to set aside this concern for the time being is that ingredients like the dictionary suggest that natural language seems not to pay serious attention to the complexity of encoding/decoding relying instead on table look up for much of its performance; and tables don't appear to be particularly compact. Nevertheless, efficiency perhaps does play a significant role in the evolution of languages since some changes are more easy for humans to adapt to, as opposed to others. Understanding this aspect of efficiency is probably another challenge for the future.

Organization of this paper..

We start with the analysis of the chromatic number in Section 2. We then use the methods to build uncertain compression schemes in Section 3. Proofs omitted from this version may be found in the full version of the paper [5].

2. UNCERTAINTY GRAPHS

We start with some elementary material in Section 2.1 that already allows us to prove Parts (1) and (2) of Lemma 1.7. The lower bound mentioned in Part (4) of Lemma 1.7 follows also relatively easily from a result of Linial [10] and we show this in Section 2.2. Our main contribution, in Section 2.3, gives the upper bound from Part (3) of Lemma 1.7.

2.1 Preliminaries

We recall the concept of a homomorphism of graphs: For graph $G = (V, E)$ and $G' = (V', E')$, we say that $\phi : V \to V'$ is a homomorphism from G to G' if $(u, v) \in E \Rightarrow (\phi(u), \phi(v)) \in E'$. We say G is homomorphic to G' if there exists a homomorphism from G to G'.

PROPOSITION 2.1. *For every N, $\ell \geq 1$ and $k' \leq k \leq N$, the k-restricted uncertainty graph $\mathcal{U}_{N,\ell,k}$ is homomorphic to the k'-restricted uncertainty graph $\mathcal{U}_{N,\ell,k'}$.*

PROOF. We construct the homomorphism ϕ from $\mathcal{U}_{N,\ell,k}$ to $\mathcal{U}_{N,\ell,k'}$ as follows: For $\pi = \langle \pi(1), \ldots, \pi(k) \rangle \in S_{N,k}$ let $\phi(\pi) = \langle \pi(1), \ldots, \pi(k') \rangle \in S_{N,k'}$. From the definitions it follows that this is a homomorphism. □

PROPOSITION 2.2. . *For every G and G' such that G is homomorphic to G', we have $\chi(G) \leq \chi(G')$.*

PROOF. Follows from the composability of homomorphisms and the fact that G is k-colorable if and only if it is homomorphic to K_k, the complete graph on k vertices. □

Part (1) of Lemma 1.7 follows immediately from Propositions 2.1 and 2.2.

PROPOSITION 2.3. *For every N, ℓ, and $k \geq \ell + 1$ the fractional chromatic number of the restricted uncertainty graph $\mathcal{U}_{N,\ell,k}$ is at most 4ℓ.*

PROOF. For every function $f : [N] \to [2\ell]$ we associate the set $I_f = \{\pi \in S_{N,k} | f(\pi(1)) = 1 \text{ and } f(\pi(j)) \neq 1 \ \forall \ j \in \{2, \ldots, \ell + 1\}\}$.

We claim that I_f is an independent set of $\mathcal{U}_{N,\ell,k}$ for every f. To see this consider an edge (π, σ) and suppose $\pi \in I_f$. Then $\sigma(1) \in \{\pi(2), \ldots, \pi(\ell + 1)\}$ and so $f(\sigma(1)) \neq 1$ and so $\sigma \notin I_f$.

Next we note that for every π, the probability that $\pi \in I_f$ for f chosen uniformly at random is $1/(2\ell) \cdot (1 - 1/(2\ell))^{\ell} \geq 1/(4\ell)$.

Thus if we give each I_f a weight of $4\ell/(2\ell)^N$, then we have that the weight of independent sets containing any given vertex π is at least one, while the sum of all weights is 4ℓ, thus yielding the claimed bound on the fractional chromatic number. □

The following is a well-known connection between fractional chromatic number and chromatic number.

PROPOSITION 2.4. *For every graph G, $\chi(G) \leq \chi_f(G) \cdot \ln |V(G)|$.*

We are now ready to prove part (2) of Lemma 1.7.

LEMMA 2.5. $\chi(\mathcal{U}_{N,\ell}) \leq \chi(\mathcal{U}_{N,\ell,\ell+1})) \leq 4\ell(\ell + 1) \ln N$

PROOF. The first inequality follows from Propositions 2.1 and 2.2. The second one follows from Proposition 2.4 and 2.3 and the fact that $\mathcal{U}_{N,\ell,\ell+1}$ has at most $N^{\ell+1}$ vertices. □

2.2 Lower Bound on Chromatic Number

We now prove Part (4) of Lemma 1.7 giving a lower bound on $\chi(\mathcal{U}_{N,\ell,k})$. We use a lower bound on a somewhat related family of graphs due to Linial [10].

DEFINITION 2.6 (SHIFT GRAPHS). *For integers N and $k < N$, we say that $\pi \in S_{N,k}$ is a* left shift *of $\sigma \in S_{N,k}$ if $\pi(i) = \sigma(i + 1)$ for $i \in [k - 1]$ and $\pi(k) \neq \sigma(1)$. We say π is a* right shift *of σ if σ is a left shift of π, and we say π is a shift of σ if π is a left shift or a right shift of σ. For integers N and k, the shift graph $\mathcal{S}_{N,k}$ is given by $V(\mathcal{S}_{N,k}) = S_{N,k}$ with $(\pi, \sigma) \in E(\mathcal{S}_{N,k})$ if π is a shift of σ.*

THEOREM 2.7 (LINIAL [10, PROOF OF THEOREM 2.1]). *For every odd k, $\chi(\mathcal{S}_{N,k}) \geq \log^{(k-1)} N$.*

(We note that the notation in [10] is somewhat different: The graph $\mathcal{S}_{N,k}$ is denoted $B_{N,t}$ for $t = (k - 1)/2$ in [10].)

We show that the uncertainty graphs contain a subgraph isomorphic to the shift graph. This gives us our lower bound on the chromatic number of uncertainty graphs.

LEMMA 2.8. *For every N, ℓ and k that is an integral multiple of ℓ, we have $\chi(\mathcal{U}_{N,\ell,k}) \geq (\log^{(2k/\ell)}(N/\ell))$.*

PROOF. First without loss of generality we only consider the case of even ℓ. Then we reduce to the case $\ell = 2$, by considering only those permutations π which fix $\pi(i) = i$ if $\ell/2$ does not divide i. This still leaves us with $2N/\ell$ unfixed elements and subpermutations from $S_{2N/\ell,2k/\ell}$ that are within distance 2 of each other are within distance ℓ when mapped back to $S_{N,k}$.

So we assume $\ell = 2$ and show that $\mathcal{U}_{N,2,k}$ contains a subgraph isomorphic to the shift graph $\mathcal{S}_{N,k}$. Consider the map ϕ from $V(\mathcal{S}_{N,k})$ to $V(\mathcal{U}_{N,2,k})$ which sends $\pi = \langle \pi(1), \ldots, \pi(k) \rangle$ to $\phi(\pi) = \sigma = \langle \sigma(1), \ldots, \sigma(k) \rangle$ as follows: Let $t = \lfloor k/2 \rfloor$. Then $\sigma(2i) = \pi(t+i)$ and $\sigma(2i+1) = \pi(t-i)$ $\sigma(t+i) = \pi(2i)$ and $\sigma(t-i) = \pi(2i+1)$. It is easy to verify that the map is a bijection and if π and π' are shifts of each other, then $\phi(\pi)$ and $\phi(\pi')$ are within distance 2 of each other. It follows that $\mathcal{U}_{N,2,k}$ contains a copy of $\mathcal{S}_{N,k}$ and so $\chi(\mathcal{U}_{N,2,k}) \geq \chi(\mathcal{S}_{N,k}) \geq log^{(k-1)}N$. □

2.3 Upper Bound on Chromatic Number

In this section we give an upper bound on the chromatic number of the uncertainty graphs. We first describe our strategy. Fix N and ℓ. Now for every k, we know that there is a homomorphism from $\mathcal{U}_{N,\ell,k}$ to $\mathcal{U}_{N,\ell,k-1}$. However we note that if we jump from $\mathcal{U}_{N,\ell,k}$ to $\mathcal{U}_{N,\ell,k-\ell}$ then the homomorphism has an even nicer property. To describe this property, we introduce a new parameter associated with the homomorphism from $\mathcal{U}_{N,\ell,k}$ to $\mathcal{U}_{N,\ell,k-\ell}$. Let us denote this homomorphism ϕ_k. For $\pi \in S_{N,k}$ let $d_k(\pi) = |\{\phi_k(\sigma) \mid (\pi, \sigma) \in E(\mathcal{U}_{N,\ell,k})\}|$. Note that $d_k(\pi)$ is independent of π and so we just denote it d_k. We note first that d_k is small.

Recall that $\phi_k : S_{N,k} \to S_{N,k-\ell}$ and maps $\pi : [k] \to [N]$ to $\pi' : [k-\ell] \to [N]$ by setting $\pi'(i) = \pi(i)$.

CLAIM 2.9. *For every k, $d_k \leq (2\ell+1)^k$.*

PROOF. Let $(\sigma, \pi) \in E(\mathcal{U}_{N,\ell,k})$ then $\delta(\sigma, \pi) \leq \ell$. In particular for every $i \in [k-\ell]$, we have there exists $j(i) \in \{-\ell, \ldots, \ell\}$ such that $\sigma(i) = \pi(i + j(i))$. Thus the sequence $j(1), \ldots, j(k-\ell)$ completely specifies $\phi_k(\sigma)$. Since the number of such sequences is at most $(2\ell+1)^{k-\ell}$, we get our claim. □

The next lemma shows that a homomorphism with a small d-value yields especially good colorings.

LEMMA 2.10. *Let ϕ be a homomorphism from G to H and let $c = \chi(H)$ and $d = \max_{v \in V(G)} |\{\phi(w) \mid (v, w) \in E(G)\}|$. Then $\chi(G) \leq 2d(d+1)\log c = O(d^2 \log c)$.*

PROOF. For integers t and M, we start by building a small family of hash functions $\mathcal{H} = \{h_1, \ldots, h_M\} \subseteq \{h : [c] \to [t]\}$ with the property that for every subset $S \subseteq [c]$, with $|S| \leq d$, and for every $i \in [c] - S$, there exists $j \in [M]$ such that $h_j(i) \notin \{h_j(i') | i' \in S\}$.

Given such a hash family, we claim there is a coloring of G with $t \cdot M$ colors. To get such a coloring, let χ' be a coloring of H with colors $[c]$. Now, consider $v \in V(G)$ and let $S_v = \{\chi'(\phi(w)) \mid (w, v) \in E(H)\}$. By the definition of d, we have $|S_v| \leq d$. Also since χ' is a coloring of H and ϕ is a homomorphism, we have $\chi'(\phi(v)) \notin S_v$. Thus by the property of $\mathcal{H}$, we have that there exists a $j = j(v)$ such that $h_j(\chi'(\phi(v))) \notin \{h_j(i') | i' \in S_v\}$. We let the coloring χ of G be $\chi(v) = (j(v), h_{j(v)}(\chi'(\phi(v))))$. Syntactically it is clear that this is a $t \cdot M$ coloring of G. To see it is valid, consider $(v, w) \in E(G)$. If $j(v) \neq j(w)$ then we are done. Else, suppose $j(v) = j(w) = j$. Then by definition of S_v we have $\chi'(\phi(w)) \in S_v$ and so $h_j(\chi'(v)) \neq h_j(\chi'(w)) \in \{h_j(i) | i \in S_v\}$, and thus $\chi(v) \neq \chi(w)$ as desired.

To conclude we need to give an upper bound on t and M.

CLAIM 2.11. *There exists such a hash family with $t \leq 2d$ and $M \leq \log(c^{d+1})$.*

PROOF. The proof is an elementary probabilistic method argument. Let $t = 2d$. We pick members of $\mathcal{H}$ at uniformly at random from $\{h : [c] \to [t]\}$. Fix a set S with $|S| \leq d$ and $i \in [c] - S$. Say that h separates i from S if $h(i) \notin \{h(i') | i' \in S\}$. The probability that a random h separates i from S is at least $1/2$ and the probability that there does not exist $h \in \mathcal{H}$ separating i from S is at most 2^{-M}. The probability that there exists S and $i \in [c] - S$ such that there does not exist $h \in \mathcal{H}$ separating i from S is strictly less than $c^{d+1} \cdot 2^{-M}$. It follows that if $M = \log c^{d+1}$ then such a family $\mathcal{H}$ exists. □

The lemma follows.

We are now ready to prove Part (3) of Lemma 1.7, restated below.

LEMMA 2.12. *There exists a constant c such that for every N, ℓ, k, we have $\chi(\mathcal{U}_{N,\ell,k}) \leq 2^{ck \log \ell} \log^{(\lfloor (k-1)/\ell \rfloor - 1)} N$.*

PROOF. We prove the lemma by induction on k. For notational simplicity assume $k-1$ is a multiple of ℓ. For $k \leq \ell$ the lemma is immediate from the fact that $\chi(\mathcal{U}_{N,\ell,1}) \leq N$. Assume the lemma is true for $k-\ell$. Then, by Lemma 2.10 we have that for $\chi(\mathcal{U}_{N,\ell,k}) \leq 2d_k(d_k+1) \cdot \log(\chi(\mathcal{U}_{N,\ell,k-\ell})) \leq 4d_k^2 \log \chi(\mathcal{U}_{N,\ell,k-\ell})$. By Claim 2.9, $d_k \leq (2\ell+1)^k \leq (4\ell)^k$ and so for $\chi(\mathcal{U}_{N,\ell,k}) \leq 4(4\ell)^{2k} \log(2^{c(k-\ell)\log \ell} \log^{(k-\ell-1)/ell} N \leq 2^{ck \log \ell} \log^{(k-1)/\ell)} N$ for a suitably large c. □

3. UNCERTAIN COMMUNICATION

We now convert some of the methods from the previous section into schemes for uncertain compression. In Section 3.1 we derive a simple compression scheme based on the relationship between fractional chromatic number and chromatic number from Section 2.1. We then use the "nested series of homomorphisms" from Section 2.3 to derive a second compression scheme in Section 3.2. The compression scheme of Section 3.2 can make errors with positive probability and has a non-linear dependence on entropy. In Section 3.3 we show that for some natural distributions, this scheme is error-free. In Section 3.4 we show how an error-free scheme working for all distributions would automatically have linear dependence on the entropy, suggesting some of the weaknesses in Section 3.2 are necessary.

3.1 A simple, zero-error compression scheme

Our first construction uses the notion of an isolating hash family as defined implicitly in Section 2.3, which we make explicit now. For positive integers ℓ, N and $m \in [N]$ and $S \subseteq [N] - \{m\}$, we say that a function $h : [N] \to \{0,1\}^\ell$ isolates m from S if $h(m) \notin \{h(m') | m' \in S\}$. We say that a hash family $\mathcal{H}_\ell = \{h_{1,\ell}, \ldots, h_{M,\ell}\}$ is (N, ℓ)-*isolating* if for

every $S \subseteq [N]$ with $|S| \leq 2^{\ell-1}$, and for every $m \in [N] - S$, there exists $j = j(m,S)$ such that $h_{j,\ell}(m) \notin h_{j,\ell}(S) \stackrel{\triangle}{=} \{h_{j,\ell}(m') | m' \in S\}$.

We note first that small isolating families exist and then give a compression scheme based on small isolating families.

LEMMA 3.1. *For every ℓ and N, there exists an (N,ℓ)-isolating family of size at most $2^\ell \cdot \log N$.*

PROOF. The proof is straightforward application of the probabilistic method. We pick $\mathcal{H} = \{h_1, \ldots, h_M\}$ by picking h_i uniformly and independently from the set of all functions from $[N]$ to $\{0,1\}^\ell$. Fix $m \notin S \subseteq [N]$. The probability that a randomly chosen h isolates m from S is at least $1/2$. Thus the probability that some h_i in $\mathcal{H}$ does not isolate m from S is at most 2^{-M}. Taking the union bound over all m, S we find that the probability that $\mathcal{H}$ does not isolate some m from S is at most $N^{2^\ell}/2^M$. We conclude that $M \leq 2^\ell \cdot \log N$ suffices for the existence of such a $\mathcal{H}$. □

We are now ready to describe our encoding and decoding schemes.

Encoding: Given m, P let $S = \{m' \in [N] \setminus \{m\} \mid P(m') \geq P(m)/2^{2\Delta}\}$ and let $\ell = \log_2 1/P(m) + 2\Delta$. Let $\mathcal{H}$ be an (N,ℓ)-isolating family of size M and let $\mathcal{H} = \{h_{1,\ell}, \ldots, h_{M,\ell}\}$. Now let $j \in [M]$ be such that $h_{j,\ell}(m) \notin \{h_{j,\ell}(m') \mid m' \in S\}$. The encoding $E(P,m)$ is defined to be $(j, h_{j,\ell}(m))$.

Decoding: Given Q and $y = (j,z) \in \mathbb{Z}^+ \times \{0,1\}^*$, let $\ell = |z|$ and let $\hat{m} = \text{argmax}_{m \in [N]: h_{j,\ell}(m) = z}\{Q(m)\}$. The decoding of the pair (Q,y) is given by $D(Q,y) = \hat{m}$.

Our next proposition verifies the correctness of the compression scheme.

PROPOSITION 3.2. *For every pair of distributions P, Q such that $\delta(P,Q) \leq \Delta$, and for every message $m \in [N]$, it is the case that $D(Q, E(P,m)) = m$.*

PROOF. Fix P, Q and m such that $\delta(P,Q) \leq \Delta$. Let $E(m,P) = (j,z)$ with $\ell = |z|$ and let $D((j,z),Q) = \hat{m}$. We will show that $\hat{m} = m$. By definition of E, we have $h_{j,\ell}(m) = z$ and by definition of D we have $h_{j,\ell}(\hat{m}) = z$. Thus, by the condition that $\hat{m}$ maximizes probability under Q of messages satisfying $h_{j,\ell}(m') = z$, we have $Q(\hat{m}) \geq Q(m)$. Since the distance of P and Q is at most Δ, we have $P(m) \leq Q(m)2^\Delta$ and $P(\hat{m}) \geq Q(\hat{m})/2^\Delta$. Combining the inequalities we get $P(\hat{m}) \geq P(m)/2^{2\Delta}$. Now let $S = \{m' \in [N] - \{m\} \mid P(m') \geq P(m)/2^{2\Delta}\}$. We have $\hat{m} \in S \cup \{m\}$. But by definition of j, we have $h_{j,\ell}(m) \notin \{h_{j,\ell}(m') | m' \in S\}$ and since $h_{j,\ell}(m) = h_{j,\ell}(\hat{m})$, we must have $m = \hat{m}$. □

Finally we analyze the performance of our scheme.

LEMMA 3.3. *The expected length of the encoding E is $O(H(P) + \Delta + \log\log N)$.*

PROOF. Fix $m \in S$. Then we have $\ell \leq 1 + \log 1/P(m) + 2\Delta$ and $M \leq (2^\ell \log N)$. Thus, the length of $E(P,m)$ is at most $2\ell + \log\log N = O(\log 1/P(m) + \Delta + \log\log N)$. Taking expectation over m drawn from P, we have the expected length of the encoding is at most $O(H(P) + \Delta + \log\log N)$. □

Theorem 1.3 follows immediately from Proposition 3.2 and Lemma 3.3.

3.2 Compression with error in the low entropy setting

Our compression for the low entropy setting (with better dependence on N) relies on an extension of our coloring scheme for the uncertainty graphs. We describe this extension in the next section and then use that to present our compression scheme afterwards.

3.2.1 Compression for chains

We start with some terminology. We say that a finite sequence of sets $A_0, \ldots, A_k$ with $A_i \subseteq [N]$ is a *chain* in $[N]$ if $|A_0| = 1$ and $A_i \subseteq A_{i+1}$ for every i. We say that w is the *leader* of the chain if $A_0 = \{w\}$. We use $\text{Chain}(N)$ to denote the set of all chains in $[N]$.

In this section we will show how to compress the leader of a chain so that it is unambiguous relative to "nearby" chains. This is in the spirit of the coloring of uncertainty graphs. Indeed vertices of the uncertainty graph $\mathcal{U}_{N,\ell,k}$ correspond to chains with the vertex $\langle \pi(1), \ldots, \pi(k) \rangle$ corresponding to the chain $\mathcal{A}$ with $A_0 = \{\pi(1)\}$ and $A_i = \{\pi(1), \ldots, \pi(\ell \cdot i)\}$ for $i \geq 1$. The compressing scheme will thus be similar to the coloring scheme, however there are two distinguishing factors: We will want to compress some chains more than others - a notion that would correspond to asking some vertices to use small colors while allowing others to use larger ones. Furthermore our chains will now grow arbitrarily fast (and not just in steps of 1 or more generally ℓ). We now describe the precise problem.

For a chain $\mathcal{A} = \langle A_0, \ldots, A_k \rangle$ we say the length of the chain, denoted $\text{lgt}(\mathcal{A})$, is the parameter k. We use $\text{sz}(\mathcal{A})$ denote the size of the final set $|A_k|$. For a chain $\mathcal{A}$ of length at least i, we let $\mathcal{A}_i$ denote its prefix of length i, i.e., $\mathcal{A}_i = \langle A_0, \ldots, A_i \rangle$.

For chain $\mathcal{A} = \langle A_0, \ldots, A_k \rangle$ and chain $\mathcal{B} = \langle B_0, \ldots, B_{k-d} \rangle$, we say $\mathcal{B}$ is within distance d from $\mathcal{A}$ if for all $i \in \{0, ..., k-d\}$, $A_{i-d} \subseteq B_i \subseteq A_{i+d}$ (where we consider sets with negative index to be the empty set). We denote the set of all chains that are within d distance from $\mathcal{A}$ by $S^d(\mathcal{A})$. Our goal next is to compress the leader of chains so that the length of the compression is small as a function of $\text{sz}(\mathcal{A})$, while it remains unambiguous to chains that are nearby.

LEMMA 3.4. *There exists a coloring scheme $\text{Col} : \mathbb{Z}^+ \times \text{Chain}(N) \to \mathbb{Z}^+$ with the following properties:*

1. *If $\text{lgt}(\mathcal{A}) \geq 2k$, then for every $s \geq \text{sz}(\mathcal{A}_{2k})$, $\text{Col}(s, \mathcal{A}_{2k}) \leq 2^{6(s+1)} \log^{(k)} N$.*
2. *Let $\mathcal{A}$ and $\mathcal{A}'$ be chains of the same length, with $\text{lgt}(\mathcal{A}) \geq 2k$ and of size at most s. Then, if $S^1(\mathcal{A}) \cap S^1(\mathcal{A}') \neq \emptyset$ and $A_0 \neq A'_0$, then $\text{Col}(s, \mathcal{A}_{2k}) \neq \text{Col}(s, \mathcal{A}'_{2k})$.*

Proof omitted from this version.

3.2.2 The Compression Scheme

We are now ready to define our final compression scheme.

Encoding: Given m, P define $r = \lfloor -\log P(m) \rfloor$ and $f = 2\lfloor \log^* N \rfloor - 1$. Further define the chain $\mathcal{A}$ of length f as follows. $A_0 = \{m\}$ and $A_k = \{m' \in [N] \mid |\log 1/P(m') - r| \leq \Delta(k+1) + 1\}$ (so that A_k is the set of messages of probability roughly $P(m)$ with the

difference in logarithms being at most $(k+1)\Delta+1$). Let $s = \mathrm{sz}(\mathcal{A})$. The encoding $E_{\mathrm{low}}(P,m) = E(P,m)$ is

$$E(P,m) = \begin{cases} (s, r, \mathrm{Col}(s,\mathcal{A})) & \text{if } s \le 2^{\frac{H(P)}{\epsilon}+2\Delta\log^* N+1} \\ \perp & \text{otherwise.} \end{cases}$$

(We assume that s and r above are encoded in some prefix-free encoding, so that the receiver can separate the three parts.)

Decoding: The decoding function $D_{\mathrm{low}}(Q,y) = D(Q,y)$ works as follows: If $y = \perp$ then the decoder outputs $\perp$. Else let $y = (s,r,c)$ and let $f = 2\lfloor \log^* N \rfloor - 1$. Let $\mathcal{B} = \langle B_0, \ldots, B_{f-1}\rangle$ be as follows: $B_0 = \{w\}$ for some w such that $|\log 1/Q(w) - r| \le \Delta + 1$. For $k \ge 1$, $B_k = \{m' \mid |\log 1/Q(m') - r| \le (k+1)\Delta + 1\}$. Find a chain $\mathcal{A}'$ with the following properties: $\mathcal{B} \in S^1(\mathcal{A}')$, $\mathrm{lgt}(\mathcal{A}') = f$, $\mathrm{sz}(\mathcal{A}') \le s$ and $\mathrm{Col}(s,\mathcal{A}') = c$. Let $\hat{m}$ be the leader of $\mathcal{A}'$. The decoding $D(Q,y)$ is set to be $\hat{m}$.

We first analyze the correctness of the decoder.

LEMMA 3.5. *For every pair of distributions P, Q such that $\delta(P,Q) \le \Delta$ and for every message $m \in [N]$ such that $E_{\mathrm{low}}(P,m) \neq \perp$, it holds that $D_{\mathrm{low}}(Q, E_{\mathrm{low}}(P,m)) = m$.*

PROOF. Fix $P \in \mathcal{P}([N])$ and a message $m \in [N]$ such that $E_{\mathrm{low}}(P,m) \neq \perp$. The following claims will show that the decoding process is well defined (and then correctness will be essentially be immediate).

CLAIM 3.6. *There exists $w \in [N]$ such that $|\log 1/Q(w) - r| \le \Delta + 1$.*

PROOF. By our choice of r, we have $|\log 1/P(m) - r| \le 1$. Now using $\delta(P,Q) \le \Delta$, we have $|\log 1/P(m) - \log 1/Q(m)| \le \Delta$, and so $|\log 1/Q(m) - r| \le \Delta + 1$. So $w = m$ gives an element in $[N]$ with the desired property. □

Thus the chain $\mathcal{B}$ is now well-defined. It remains to show that there exists a chain $\mathcal{A}'$ satisfying the required properties. The next claim shows that $\mathcal{B} \in S^1(\mathcal{A})$, therefore $\mathcal{A}$ is a candidate for the role of $\mathcal{A}'$.

CLAIM 3.7. $\mathcal{B} \in S^1(\mathcal{A})$.

PROOF. The proof follows easily from our choice of $\mathcal{A}$, $\mathcal{B}$ and the fact that P and Q are Δ-close. Let $k \in \{0, ..., f-1\}$. We need to show that B_k is sandwiched between A_{k-1} and A_{k+1}.

First, We will show that $B_k \subseteq A_{k+1}$. When $k = 0$, we need to show that $w \in A_1$. Indeed,

$$\begin{aligned} & |\log 1/Q(w) - r| \le \Delta + 1 \\ \Rightarrow\ & |\log 1/P(w) - r| \le 2\Delta + 1 \\ \Rightarrow\ & w \in A_1 \ . \end{aligned}$$

Now consider $1 \le k \le f-1$. We have,

$$\begin{aligned} B_k &= \{m' \in [N] \mid |\log 1/Q(m') - r| \le (k+1)\Delta + 1\} \\ &\subseteq \{m' \in [N] \mid |\log 1/P(m') - r| \le (k+2)\Delta + 1\} \\ &= A_{k+1} \ . \end{aligned}$$

This shows that $B_k \subseteq A_{k+1}$. Next we show that $A_{k-1} \subseteq B_k$, for $2 \le k \le f-1$. We have

$$\begin{aligned} A_{k-1} &= \{m' \in [N] \mid |\log 1/P(m') - r| \le k\Delta + 1\} \\ &\subseteq \{m' \in [N] \mid |\log 1/Q(m') - r| \le (k+1)\Delta + 1\} \\ &= B_k \ . \end{aligned}$$

The case where $k = 1$ and $w \in B_1$ was proved in Claim 3.6. So we are done. □

To conclude, the decoder can find a chain $\mathcal{A}'$ such that $\mathrm{sz}(A') \le s$, $\mathrm{lgt}(\mathcal{A}') = \mathrm{lgt}(\mathcal{A})$, $\mathrm{Col}(s,\mathcal{A}') = \mathrm{Col}(s,\mathcal{A})$ and there exists a chain $\mathcal{B} \in S^1(\mathcal{A}') \cap S^1(\mathcal{A})$. From Lemma 3.4 the leader of A' is m as required.

We are now ready to prove Theorem 1.4.

PROOF. We now estimate the probability that the encoder will fail. Fix some probability P and a message m such that $E(P,m) = \perp$. We will first show that $P(m) \le 2^{-\frac{H(P)}{\epsilon}}$. Later, we will bound the probability that "m has such small probability" by ϵ.

Consider the chain $\mathcal{A} = \langle A_0, \ldots, A_f\rangle$ as defined by the encoder. In this case, the size of the largest set, $|A_f|$, is more then the threshold $T = 2^{\frac{H(P)}{\epsilon}+2\Delta\log^* N+1}$. So, there is some element $m' \in A_f$ such that $P(m') \le \frac{1}{T}$. By our choice of A_f, $P(m') \ge 2^{-\lfloor -\log P(m)\rfloor - (f+1)\Delta - 1} \ge P(m)2^{-2\Delta\log^* N - 1}$. Calculating,

$$\frac{1}{T} \ge P(m)2^{-2\Delta\log^* N-1} \Rightarrow P(m) \le \frac{2^{2\Delta\log^* N+1}}{T} = 2^{-\frac{H(P)}{\epsilon}}$$

Therefore, we can bound the failure probability by the probability that $P(m) \le 2^{-\frac{H(P)}{\epsilon}}$. Using the fact that $\mathbf{E}_{m\leftarrow P[N]}\left[\log\frac{1}{P(m)}\right] = H(P)$, we deduce the following by Markov's inequality,

$$\Pr_{m\leftarrow P[N]}\left[P(m) \le 2^{-\frac{H(P)}{\epsilon}}\right] = \Pr_{m\leftarrow P[N]}\left[\log\frac{1}{P(m)} \ge \frac{H(P)}{\epsilon}\right] \le \epsilon$$

We will finish the proof by bounding the performance of the scheme. To this end consider a distribution P and a message $m \in [N]$ such that $E(P,m) \neq \perp$ (i.e $\mathrm{sz}(\mathcal{A}) \le T$). The encoder sends $r = \lfloor -\log P(m)\rfloor$, $s = \mathrm{sz}(\mathcal{A})$ and $\mathrm{Col}(s,\mathcal{A})$. We first analyze the contribution of sending r to the performance. Because $\log|r| = O\left(\log(\frac{1}{P(m)})\right)$, the accepted length of sending r in a prefix-free encoding is at most $O\left(\mathbf{E}_{m\leftarrow P[N]}\log(\frac{1}{P(m)})\right) = O(H(P))$.

Now we analyze the length of $(s, \mathrm{Col}(s,\mathcal{A}))$. By Lemma 3.4:

$$C(s,\mathcal{A}) \le 2^{6(s+1)}\log^{(f)} N = 2^{O(s)}$$

Hence, the length of $(s, \mathrm{Col}(s,\mathcal{A}))$ is at most

$$O(\log s) + \log C(s,\mathcal{A}) = O(s) = 2^{\frac{H(P)}{\epsilon}+2\Delta\log^* n+O(1)} \ .$$

Thus, from the linearity of expectations, it follow that the total performance is at most $2^{\frac{H(P)}{\epsilon}+2\Delta\log^* n+O(1)}$. □

3.3 Error-free Compression for Natural Distributions

In this section we will show that for a large class of natural distributions, the above scheme is error free. We start by describing the natural distributions we can capture.

We say that a distribution $P \in \mathcal{P}([N])$ is *flat* it there exists a set $S \subseteq [N]$ such that P is uniform on S. The distribution is called *geometric* if there exists parameter $\alpha \in (0,1)$ and a permutation π on $[N]$ such that for all $k \in [N-1]$ it holds that $P(\pi(k+1)) = \alpha P(\pi(k))$. We call P *binomial* if there exists a parameter $p \in (0,1)$ and a permutation π on $[N]$

such that $\forall k \in [N]$, $P(\pi(k)) = \binom{N}{k} p^k (1-p)^{n-k}$. The sets of all flat, geometric and binomial distributions over $[N]$ are denoted by Flat_N, Geo_N and Bin_N respectively.

The following theorem shows that the scheme $(E_{\mathrm{low}}, D_{\mathrm{low}})$ performs well *without error* on all of the above natural distributions. Moreover, this theorem is stable in the sense that the guarantee on the performance holds even if a distribution is only close to one of the above-mentioned natural distributions.

THEOREM 3.8. *Let* $\mathcal{F} \triangleq \mathrm{Flat}_N \cup \mathrm{Geo}_N \cup \mathrm{Bin}_N$ *and* $L(P) \triangleq 2^{H(P)} \lceil \Delta \log^* N \rceil$. *Then the scheme* $(E_{\mathrm{low}}, D_{\mathrm{low}})$ *(with* ϵ *set to* 0*) is a* $(\Delta, 0, \mathcal{F}, O(L(P)))$*-UCS. Moreover, if* $P \in \mathcal{P}([N])$ *is* $\Delta \log^* N$*-close to a distribution* $\tilde{P} \in \mathcal{F}$ *then the performance of the scheme on* P *is* $\mathbf{E}_{m \leftarrow_P U}[|E(P,m)|] = O\left(L(\tilde{P})\right)$.

We prove the theorem above by identifying a broad condition on distributions, which we call the *capacity*, and showing that the performance of our scheme is good if the capacity is small. We define this notion next, show that it is small for the distributions under consideration in Lemma 3.9 next, and finally bound the performance as a function of the capacity in Lemma 3.10 afterwards, thus leading to a proof of Theorem 3.8.

Let $P \in \mathcal{P}([N])$ be a distribution and let $S \subseteq [N]$ be its support. We say that $U \subseteq S$ is a *unit set* of P if for any two elements $m_1, m_2 \in U$ the distance $|\log P(m_1) - \log P(m_2)| \leq 1$. We define the *capacity* of P, denoted by $\mathscr{C}(P)$, to be the minimal $c \in \Re$ such that the size of every unit set of P is bounded by 2^c.

Later, we will prove the following lemma, showing that for the previously discussed distributions, the capacity is roughly the entropy.

LEMMA 3.9. *Let* $P \in \mathrm{Flat}_N \cup \mathrm{Geo}_N \cup \mathrm{Bin}_N$. *Then* $\mathscr{C}(P) \leq H(P) + O(1)$.

Theorem 3.8 follows immediately from Lemma 3.9 combined with the following lemma.

LEMMA 3.10. *For every* P $(E_{\mathrm{low}}, D_{\mathrm{low}})$ *(with respect to* $\epsilon = 0$*) is a* $\left(\Delta, O\left(\log\left(H(P)\right) + 2^{\mathscr{C}(P)} \lceil \Delta \log^* N \rceil\right)\right)$ *scheme. Moreover, if* P *is* $\Delta \log^* N$ *close to a distribution* $\tilde{P}$*, then the performance of the scheme on* P *is* $O\left(\log\left(H(P)\right) + 2^{\mathscr{C}(\tilde{P})} \lceil \Delta \log^* N \rceil\right)$.

We omit the proofs of the lemmas above from this version.

3.4 Dependence of communication on entropy

In the previous sections we gave a scheme with performance that is exponential in the entropy. This scheme is error-free for some natural distributions and had positive error for general distributions. The next lemma shows that if we cannot find a scheme with performance that is linear in the entropy, then any scheme that we will find must have positive error for some distributions.

LEMMA 3.11. *For every non-decreasing function* $L : \Re^+ \to \Re^+$ *there exists a constant* $c = c_L$ *such that the following holds: If there exists* $(\Delta, L(H(P)))$*-UCS for some* $\Delta > 0$*, then there exists a* $(\Delta, c \cdot (1 + H(P)))$*-UCS.*

Proof omitted from this version.

4. REFERENCES

[1] M. Braverman and A. Rao. Information equals amortized communication. In R. Ostrovsky, editor, *FOCS*, pages 748–757. IEEE, 2011.

[2] R. Cole and U. Vishkin. Deterministic coin tossing with applications to optimal parallel list ranking. *Information and Control*, 70(1):32–53, 1986.

[3] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 1991.

[4] O. Goldreich, B. Juba, and M. Sudan. A theory of goal-oriented communication. *J. ACM*, 59(2):8, 2012.

[5] E. Haramaty and M. Sudan. Deterministic compression with uncertain priors. *Electronic Colloquium on Computational Complexity (ECCC)*, 19:166, 2012.

[6] P. Harsha, R. Jain, D. A. McAllester, and J. Radhakrishnan. The communication complexity of correlation. *IEEE Transactions on Information Theory*, 56(1):438–449, 2010. Preliminary version in IEEE CCC 2007.

[7] B. Juba, A. T. Kalai, S. Khanna, and M. Sudan. Compression without a common prior: an information-theoretic justification for ambiguity in language. In B. Chazelle, editor, *ICS*, pages 79–86. Tsinghua University Press, 2011.

[8] B. Juba and M. Sudan. Universal semantic communication I. In *Proceedings of the 2008 ACM International Symposium on Theory of Computing, Victoria, British Columbia, Canada, May 17-20, 2008*, pages 123–132. ACM, 2008.

[9] E. Kushilevitz and N. Nisan. *Communication Complexity*. Cambridge University Press, 1997.

[10] N. Linial. Locality in distributed graph algorithms. *SIAM J. Comput.*, 21(1):193–201, 1992.

[11] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.

[12] J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3):337–342, 1977.