The price of uncertainty in communication

Mark Braverman[1] and Brendan Juba[2]

***Abstract*— We consider the problem of one-way communication when the recipient does not know exactly the distribution that the messages are drawn from, but has a *"prior"* distribution that is known to be close to the source distribution, a problem first considered by Juba et al. [5]. This problem generalizes the classical source coding problem in information theory, in which the receiver knows the source distribution exactly, that was first considered in Shannon's work [6]. This *"uncertain priors"* coding problem was intended to illuminate aspects of natural language communication, and has applications to adaptive compression schemes. We consider the question of how much longer the messages need to be in order to cope with the uncertainty that the sender and receiver have about the receiver's prior and the source distribution, respectively, as compared to the source coding problem.**

We obtain lower bounds for one-way communication using uncertain priors that are tight up to low-order terms. Specifically, we consider two variants of the uncertain priors problem. First, we consider the original setting of Juba et al. [5] in which the receiver is required to correctly recover the message with probability 1. We find that in this setting, the overhead of $2\log\alpha + O(1)$ bits achieved by that scheme when the prior is α-*close* to the source is optimal up to an additive $O(\log\log\alpha)$ bits. We also consider a setting introduced in the work of Haramaty and Sudan [3], in which the receiver is permitted to fail to recover the message with some positive probability ϵ. In this setting, we find that the optimal overhead is essentially $\log\alpha + \log 1/\epsilon$ bits by presenting both a variant of the coding scheme of Juba et al. with an overhead of $\log\alpha+\log 1/\epsilon+1$ bits, and a lower bound that matches up to an additive $O(\log\log\alpha)$ bits. Our lower bounds are obtained by observing that a worst-case, one-way communication complexity problem can be embedded in the sources and priors that any any uncertain priors coding scheme must address.

I. Introduction

In a seminal work, Shannon [6] considered the problem of how to encode a message so that it can be transmitted and decoded reliably across a channel that introduces errors. Shannon's contribution in that work was two-fold: first, he identified how large any encoding of messages would need to be in the absence of noise – the *"source coding"* problem – and then identified the optimal length for encodings that can be decoded in spite of errors introduced by the channel. The difference between these two lengths – the number of extra, *redundant* bits from the standpoint of source coding – may be viewed as the *"price" of noise-resilience*.

Such work in information theory essentially settled the basic quantitative questions of noise-tolerant transmission in telecommunications. In natural communication, however, errors frequently arise not due to any kind of interference, but instead due to a lack of shared context. In the interest of understanding why natural language is structured so that such errors can occur and how they might be addressed, Juba et al. [5] introduced a model of *communication with uncertain priors*. This is a variant of the source coding problem in which the sender and receiver do not agree on the source distribution that the messages are drawn from. Thus, errors arise because the sender and receiver do not agree on which messages should be considered more likely, and should therefore receive shorter codewords so as to minimize the expected encoding length. We note that this problem also has applications to adaptive data compression, in which parties use their empirical observations of message frequencies to encode messages. This would be useful since the distribution over messages generally changes over time for a variety of reasons; this clearly occurs in natural language content, for example, as new words are introduced and old ones fall out of use. Since different parties on a network will in general observe different empirical distributions of messages, they must tolerate some (limited) inconsistency about the relative frequency of the different messages.

In the model of Juba et al., "uncertainty" about the priors is captured by the following kind of distance between the priors used by the sender and receiver. Namely, if the sender has a source distribution P and the receiver expects a distribution Q, then we say that P and Q are α-*close* when $\frac{1}{\alpha}Q(x) \leq P(x) \leq \alpha Q(x)$ for every message x. Juba et al. then presented a scheme (building on the coding technique of Braverman and Rao [1]) in which every source P can be encoded by $H(P)+2\log\alpha+O(1)$ bits so that every decoder using an α-close prior Q will recover the message correctly. Thus, the Juba et al. scheme uses $2\log\alpha+O(1)$ bits beyond the $H(P)$ bits required of the basic source coding problem.

Juba et al. had noted that $H(P)+\log\alpha$ bits is necessary for such "uncertain priors" coding, and asked whether the redundancy could be reduced to this $\log\alpha$ lower bound. In this work, we address this question by showing that it cannot. Indeed, in the original errorless coding setting, we show that the original scheme is optimal up to terms of size $O(\log\log\alpha)$, and hence the *"price of uncertainty"* is, up to lower-order terms, an additional $2\log\alpha$ bits. We also consider the variant of the problem introduced by Haramaty and Sudan [3] in which the decoding is allowed to fail with some positive probability ϵ. We also identify the price of uncertainty in this setting: we note that the scheme of Juba et al./Braverman-Rao can be modified to give an uncertain priors coding of length $H(P)+\log\alpha+\log 1/\epsilon+1$, and show a lower bound of $H(P)+\log\alpha+\log 1/\epsilon-O(\log\log\alpha)$ when $\epsilon \geq 1/\alpha$. The price of uncertainty α when error $\epsilon > 1/\alpha$ is allowed is thus essentially reduced to $\log\alpha + \log 1/\epsilon$ bits.

[1]Princeton University mbraverm@cs.princeton.edu
[2]Washington University in St. Louis bjuba@wustl.edu

We obtain our results by reducing a one-way communication complexity problem to uncertain prior coding: Consider the problem where Alice receives as input a message from a domain of size roughly α^2, Bob receives as input a set S of size roughly α containing Alice's message, and Alice's task is to send the message to Bob. For our reduction, we note that there is a low-entropy distribution P with most of its mass on Alice's input and an α-close family of distributions corresponding to the sets S that essentially capture this problem. Thus, a lower bound for the communication complexity problem translates directly to the desired lower bound on the redundancy since the entropy of P is negligible compared to the overhead. Our lower bounds for the two variants, errorless and positive error, of the uncertain priors coding problem are then obtained from lower bounds for this same problem in the analogous variant of the one-way communication complexity model. Specificaly, we obtain a $2\log\alpha - O(\log\log\alpha)$ lower bound in the errorless model and a $\log\alpha + \log 1/\epsilon - O(\log\log\alpha)$ lower bound in the model with ϵ error, yielding our main results.

A. Aside: why not the relative entropy/KL-divergence?

A common misconception upon first learning about the model of Juba et al. [5] is that (a) the problem had already been solved and (b) the correct overhead is given by the *relative entropy* (or "*KL-divergence*"), $RE(P\|Q) = \sum_x P(x)\log\frac{P(x)}{Q(x)} \leq \log\alpha$. Of course, our lower bounds imply that this is incorrect, but it is useful to understand the reason. Indeed, our problem is somewhat unusual in that the relative entropy is essentially the correct answer to a few similar problems, including (i) the problem where the *sender* does not know the source distribution P, only an approximation Q and (ii) the problem where the communication is two-way (and the receiver can tell the sender when to stop)—this variant essentially follows from the work of Braverman and Rao [1]. The difference is that in our setting unlike (i), the sender does not know Q and unlike (ii), has no way to learn anything about it, apart from the fact that it is α-close to P. The sender's message must simultaneously address decoding by all possible α-close priors using only this knowledge, hence the connection to a worst-case communication complexity set-up.

We also note that the problem we consider is not addressed by the *universal compression* schemes of Lempel and Ziv [8]. Lempel and Ziv's compression schemes provide *asymptotically* good compression of many messages from, e.g., a Markovian source. By contrast, the problem we consider here concerns the compression of a *single message.*

II. The Model and Prior State of the Art

We now recall the model we consider in more detail and review the existing work on this model. Our work concerns the uncertain priors coding problem, originally introduced by Juba et al [5]. In the following we will let M denote a set of *messages* and $\Delta(M)$ denote the probability distributions on M.

Definition 1 (α-close): We will say that a pair of distributions $P, Q \in \Delta(M)$ are *α-close* for $\alpha \geq 1$ if for every $m \in M$,

$$\frac{1}{\alpha}Q(m) \leq P(m) \leq \alpha Q(m).$$

In uncertain priors coding, one party (*"Alice"*) wishes to send a message drawn from a source distribution $P \in \Delta(M)$ using a one-way (noiseless) binary channel to another party (*"Bob"*) who does not know P exactly. Bob does know a distribution $Q \in \Delta(M)$, however, that is guaranteed to be α-close to P, where Alice in particular knows α. We assume that Alice and Bob share access to an infinitely long common random string R. Our objective is to design an encoding scheme that, regardless of the pair of *"prior"* distributions P and Q, enables Bob to successfully decode the message using as short a transmission as possible:

Definition 2 (Errorless uncertain priors coding): An *errorless uncertain priors coding scheme* is given by a pair of functions $E : M \times \mathbb{N} \times \{0,1\}^{\mathbb{N}} \times \Delta(M) \to \{0,1\}^*$ and $D : \{0,1\}^* \times \mathbb{N} \times \{0,1\}^{\mathbb{N}} \times \Delta(M) \to M$ such that for every $m \in M$, $\alpha \in \mathbb{N}$, $R \in \{0,1\}^{\mathbb{N}}$, $P \in \Delta(M)$, and α-close $Q \in \Delta(M)$, if $c = E(m, \alpha, R, P)$, then $D(c, \alpha, R, Q) = m$. When $R \in \{0,1\}^{\mathbb{N}}$ is chosen uniformly at random and m is chosen from P, we will refer to $\mathbb{E}_{m,R}[|E(m, \alpha, R, P)|]$ as the *encoding length* of the scheme (E, D) for P and α.

The key quantity we focus on in this work is a measure of the overhead introduced by uncertain priors coding, as compared to standard source coding where $P = Q$. Huffman codes [4], for example give a means to encode a message in this setting using on average no more than one more bit than the *entropy* of P, $H(P) = \sum_{m\in M} P(m)\log\frac{1}{P(m)}$. Then:

Definition 3 (Redundancy): The *redundancy* of an uncertain priors coding scheme (E, D) for a given $\alpha \in \mathbb{N}$ is given by the maximum over $P \in \Delta(M)$ of the encoding length of (E, D) for α and P minus the entropy of P.

As an uncertain priors coding scheme gives a means to perform standard source coding (again, by taking $P = Q$, $\alpha = 1$) it follows that the redundancy is always nonnegative since (as is well known) the entropy of P is always a lower bound on the encoding length for any solution to the source coding problem.

Juba et al. [5], using a coding technique introduced by Braverman and Rao [1], showed an errorless uncertain priors coding scheme that achieved redundancy bounded as a function of α:

Theorem 4 (Juba et al. [5]): There is an errorless uncertain priors coding scheme that achieves redundancy $2\log\alpha + 2$.

They also noted that a lower bound on the redundancy of $\log\alpha$ is easy to obtain. This follows essentially by taking a distribution over possible source distributions P that are α-close to a common distribution Q, and noting that the resulting distribution over messages has entropy $H(P) + \log\alpha - o(1)$. Our first main theorem, in Section III-A, will improve this lower bound to $2\log\alpha - O(\log\log\alpha)$, so that it nearly matches the upper bound given by Theorem 4 (up to lower-order terms).

We also consider a variant of the original uncertain priors coding problem, introduced by Haramaty and Sudan [3], in which we allow an error in communication with some bounded but positive probability:

Definition 5 (Positive-error uncertain priors coding): For any $\epsilon > 0$, an *ϵ-error uncertain priors coding scheme* is given by a pair of functions $E : M \times \mathbb{N} \times \{0,1\}^{\mathbb{N}} \times \Delta(M) \to \{0,1\}^*$ and $D : \{0,1\}^* \times \mathbb{N} \times \{0,1\}^{\mathbb{N}} \times \Delta(M) \to M$ such that for every $\alpha \in \mathbb{N}$, $P \in \Delta(M)$, and α-close $Q \in \Delta(M)$, when $m \in M$ is chosen according to P and $R \in \{0,1\}^{\mathbb{N}}$ is chosen uniformly at random,

$$\Pr_{m,R}[D(E(m,\alpha,R,P),\alpha,R,Q) = m] \geq 1 - \epsilon.$$

Again, we will refer to $\mathbb{E}_{m,R}[|E(m,\alpha,R,P)|]$ as the *encoding length* of the scheme (E,D) for P and α.

The redundancy for positive-error uncertain priors coding is then defined in exactly the same way as for errorless coding.

We briefly note that the definition of Haramaty and Sudan [3] differs in two basic ways. First, their definition does not include a common random string since they were primarily interested in deterministic coding schemes. Second, they required that the decoder output a special symbol $\perp$ when it makes an error. Our one positive result (Theorem 7) can be easily modified to satisfy this condition, but we prove our lower bound, Theorem 8, for the slightly more lenient model stated here.

We also briefly note that Canonne et al. [2] have considered another variant of the basic model in which Alice and Bob do not share the common random string R perfectly, only correlated random strings. In both this imperfect randomness model and the deterministic model of Haramaty and Sudan, the known encoding schemes feature substantially greater redundancy than $2\log\alpha$—the redundancy for these schemes is linear in the entropy and, in the case of the deterministic schemes of Haramaty and Sudan, furthermore depends on the size of M. It is an interesting open question for future work whether or not lower bounds of this form can be proved for these other settings.

III. The Price of Uncertainty

We now establish lower bounds on the redundancy for uncertain priors coding schemes, in both the errorless and positive error variants. We will see that both of these lower bounds are tight up to lower-order terms. Hence, at least to the first order, we identify the "price" incurred by uncertainty about a recipient's prior distribution, beyond what is inherently necessary for successful communication in the absence of such uncertainty.

In both cases, our lower bounds are proved by exploiting the worst-case nature of the guarantee over priors P and Q to embed a (worst-case) one-way communication complexity problem into the uncertain priors coding problem. We then essentially analyze both errorless and positive-error variants of the communication complexity problem to obtain our lower bounds for the respective uncertain priors coding problems.

A. Lower bound for errorless communication

We first consider the original errorless variant of uncertain priors coding, considered by Juba et al. [5]. Earlier, in Theorem 4, we recalled that they gave a errorless uncertain priors coding scheme with redundancy $2\log\alpha + 2$. We now show that this is essentially tight:

Theorem 6: Any errorless compression scheme for uncertain priors suffers redundancy at least $2\log\alpha - 3\log\log\alpha - O(1)$ for every sufficiently large α, for some pair of α-close priors.

Proof: Let k be the largest integer such that $k\sqrt{\log k} \leq \alpha$ (so $\log\alpha = \log k + \frac{1}{2}\log\log k + O(1)$) and consider a message set of size $k^2 + 1$. Now, suppose that Alice has a prior P that gives a distinguished message m probability $1 - 1/\log k$ and gives the other k^2 messages probability $1/k^2 \log k$. Bob, on the other hand, has a prior Q which gives a set S of $k+1$ messages that includes m probability $1/k\sqrt{\log k}$, and gives the other $k^2 - k$ messages uniform probability $\frac{1}{k^2-k}\left(1 - \frac{k+1}{k\sqrt{\log k}}\right)$.

Lemma 1: For sufficiently large α and k, P and Q are α-close.

Lemma 2: P has entropy $O(1)$.

It therefore suffices to give a lower bound on the expected codeword length to obtain a lower bound on the redundancy up to an additive $O(1)$ term, since the entropy of the messages themselves only contribute $O(1)$ bits.

Lemma 3: Any errorless coding scheme must have expected codeword length at least $2\log k - 2\log\log k$ for some α-close pair P and Q as described above.

Now, since our choice of α ($\log\alpha = \log k + \frac{1}{2}\log\log k + O(1)$ and $\alpha > k$) ensures that $2\log k - 2\log\log k$ is at least $2\log\alpha - 3\log\log\alpha - O(1)$, we find that the redundancy is indeed likewise at least $2\log\alpha - 3\log\log\alpha - O(1)$ since, as shown in Lemma 2, $H(P)$ is also $O(1)$. ∎

We note that apart from the encoding length, this lower bound is essentially tight in another respect: If there are fewer than α^2 messages, then these can be indexed by using less than $2\log\alpha$ bits, and so clearly a better scheme is possible by ignoring the priors entirely and simply indexing the messages in this case. Thus, no hard example for uncertain priors coding can use a substantially smaller set of messages.

Although our hard example uses a source distribution with very low entropy, we also briefly note that it is possible to extend it to examples of hard source distributions with high entropy. Namely, suppose that there is a second, independent, high-entropy distribution T, and that Alice and Bob wish to solve two independent source coding problems: one with their α-close priors P and Q, and the other a standard source coding problem for T. The joint distributions PT and QT are then α-close and have entropy essentially $H(T)$. The analysis is now similar to the Slepian-Wolf Theorem [7]: The "side problem" of source coding for T cannot reduce the coding length for uncertain priors coding of P and Q since Alice and Bob can simulate drawing an independent message m'' from T using the shared randomness; this shared message m'' can then be used by Bob to decode a

hash of the message Alice would send in the joint coding problem.

B. The price of uncertainty when errors are allowed

We now turn to the setting where Alice and Bob are allowed to fail at the communication task with positive probability ϵ. Haramaty and Sudan [3] introduced this setting, but only considered deterministic schemes for this problem. We first note that the original techniques from Braverman and Rao [1] (used in the original work of Juba et al. [5]) give a potentially much better upper bound when errors are allowed:

Theorem 7: For every α and ϵ, there is an uncertain priors compression scheme with expected code length $H(P) + \log\alpha + \log\frac{1}{\epsilon} + 1$ that is correct with probability $1-\epsilon$ when P and Q are α-close. (Actually, we only require $Q(x) \geq P(x)/\alpha$ for all x.)

We now give a lower bound that essentially matches this upper bound, up to terms of size $O(\log\log\alpha)$. Thus, in this ϵ-error setting, we also identify the "price" of uncertainty α up to lower-order terms. The proof is a natural extension of the proof of Theorem 6, showing that when the codes are so short, a miscommunication is not only possible but must be likely.

Theorem 8: Any compression scheme for uncertain priors that is correct with probability $1-\epsilon$ suffers redundancy at least $\log\alpha + \log\frac{1}{\epsilon} - \frac{9}{2}\log\log\alpha - O(1)$ for every sufficiently large $\alpha > \frac{1}{\epsilon}$ and some pair of α-close priors.

IV. Suggestions for future work

We now conclude with a handful of natural questions that are unresolved by our work. First of all, in both the errorless and positive error settings, the known upper and lower bounds still feature gaps of size $\Theta(\log\log\alpha)$. It would be nice to tighten these results further, reducing the gap to $\Theta(1)$ if possible in particular (as is known for standard source coding).

Second, we do not have any better lower bounds for the deterministic or imperfect-randomness settings, respectively from Haramaty and Sudan [3] and Canonne et al. [2]. The known upper bounds in these settings are *much* weaker, featuring redundancy that grows at least linearly in the entropy of the source distribution, and in the case of the deterministic codes of Haramaty and Sudan, some kind of dependence on the size of the distribution's support. Is this inherently necessary? Haramaty and Sudan give some reasons to think so, noting a connection between graph coloring and uncertain priors coding schemes: namely, they identify geometric distributions with the nodes of a graph and include edges between all α-close geometric distributions. They identify the "colors" with the messages sent by the high-probability element in each distribution. They then point out that a randomized uncertain priors scheme is essentially a "fractional coloring" of this graph, whereas a deterministic scheme is a standard coloring. Thus, there is reason to suspect that this problem may be substantially harder, and so a stronger lower bound (e.g., depending on the size of M) may be possible for the deterministic coding problem.

As for the imperfectly shared randomness setting, Canonne et al. [2] give some lower bounds for a communication complexity task (sparse gap inner product) showing that imperfectly shared randomness may require substantially more communication than perfectly shared randomness. Although this does not seem to immediately yield a strong lower bound for the uncertain priors problem (as it is for a very specific communication problem), their technique may provide a starting point for a matching lower bound on the redundancy in that setting as well.

Finally, none of these works address the question of the *computational* complexity of uncertain priors coding, a question originally raised in the work of Juba et al. [5]. In particular, codes such as arithimetic coding (first considered by Elias in an unpublished work) can encode a message m by making a number of queries to a CDF for the source distribution P that is *linear in the code length*, and has similar computational complexity. Although the code length is not as tight as Huffman coding [4] for one-shot coding, it is nearly optimal. The question is whether or not a similarly computationally efficient, near-optimal compression scheme for uncertain priors coding exists.

Acknowledgements

This work was partially performed while B. Juba was affiliated with Harvard University and supported by ONR grant number N000141210358; B. Juba is also supported by an AFOSR Young Investigator Award. M. Braverman is partially supported by NSF Award CCF-1525342, NSF CAREER award CCF-1149888, a Packard Fellowship in Science and Engineering, and the Simons Collaboration on Algorithms and Geometry.

References

[1] M. Braverman and A. Rao. Information equals amortized communication. In *Proc. 52nd FOCS*, pages 748–757, 2011.

[2] C. L. Canonne, V. Guruswami, R. Meka, and M. Sudan. Communication with imperfectly shared randomness. In *Proc. 6th ITCS*, pages 257–262, 2015.

[3] E. Haramaty and M. Sudan. Deterministic compression with uncertain priors. In *Proc. 5th ITCS*, pages 377–386, 2014.

[4] D. Huffman. A method for the construction of minimum-redundancy codes. In *Proc. I.R.E.*, pages 1098–1102, 1952.

[5] B. Juba, A. Kalai, S. Khanna, and M. Sudan. Compression without a common prior: an information-theoretic justification for ambiguity in language. In *Proc. 2nd Innovations in Computer Science*, pages 79–86, 2011.

[6] C. Shannon. A mathematical theory of communication. *Bell Sys. Tech. J.*, 27:379–423, 623–656, 1948.

[7] D. Slepian and J. K. Wolf. Noiseless coding of correlated information sources. *IEEE Trans. Inf. Theory*, 19(4):471–480, 1973.

[8] J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Trans. Inf. Theory*, 23(3):337–343, 1977.