Broadcasting with Side Information

Ramakrishna Gummadi
CSL & ECE
UIUC, Urbana, IL, USA
gummadi2@illinois.edu

Amin Shokrollahi
ALGO & LMA
EPFL, Lausanne, Switzerland
amin.shokrollahi@epfl.ch

Ramavarapu Sreenivas
CSL & IESE
UIUC, Urbana, IL, USA
rsree@illinois.edu

***Abstract*—We consider the problem of multicasting data from a source to receivers that possess arbitrary subsets of the data apriori as *side information*. Fountain codes, which are an ideal solution to the standard multicasting problem without any side information, have also been proposed as a potential approach for the side information problem in multiple independent studies recently. Relevant to such a context, we formulate and study an optimization problem over degree distributions to minimize the overhead necessary for complete decoding, and prove that: (i) Degree distributions converging to the standard soliton distribution cannot exploit side information in terms of the overhead necessary for complete decoding. (ii) An asymptotic *shifted* soliton distribution achieves an overhead which is within a constant factor (< 2) of the optimal overhead (iii) There exist no degree distributions which achieve asymptotically optimal overhead for any non trivial constant fraction of the data as side information. While (iii) is discouraging, this limitation can be sidestepped by using systematic versions, where intermediate symbols are generated from the source symbols, to which the fountain code is then applied. One important implication of this is that the systematic versions are in a sense indispensable to achieve asymptotic rate optimality for the side information problem.**

I. Introduction

Multicasting via a common broadcast channel is important for a number of practical applications (especially relevant to the wireless medium), and its motivation hardly needs an elaboration. One aspect that makes it a non trivial problem is the presence of losses, also called erasures. Forward error correction using rateless codes ([13]) simultaneously achieves rate optimality with low encoding/decoding complexity. In some scenarios, one is faced with a complementary aspect, namely the apriori presence of arbitrary subsets of the source data at the receivers, also referred to as *side information*. In practice, this side information could have plausibly been gathered from alternate sources or perhaps from a previous incompletely decoded download session. The issue of designing codes which exploit the side information present at the receivers has been considered previously by a number of authors. Metzner ([9]) originally identified random linear coding as a useful approach to the context of designing an efficient broadcast retransmission protocol, which is essentially equivalent to the side information problem. More recently, Birk and Kol ([2]) have considered code design to minimize the number of transmissions, assuming that the encoder is provided with the complete side information pattern at the receivers. This problem has come to be known as *index coding* in its form where each receiver requests a unique packet. [1] showed that the optimal linear index code can be formulated as a rank minimization problem on finite field matrices. The unique requests constraint can also be generalized to arbitrary subsets of possibly non disjoint requests, for which multicast is a practically important non trivial subclass.

While understanding optimal *index codes* is undoubtedly a crucial aspect to the multicast side information problem (besides also being very generally related to network coding ([10])), this approach is not directly relevant to a practical setting. Firstly, conveying the side information at each receiver is a task that requires too much overhead in the form of ARQ. Further, processing the collected data to compute the optimal code is likely to be impractical because arbitrary rank minimization problems on finite field matrices are computationally intractable. Consequently, approaches that are oblivious to the precise pattern, and those that can also simultaneously deal with lossy transmissions are highly desirable.

An approach that avoids collecting feedback from the broadcast audience is to use random linear coding (RLC), where each coded packet is chosen as a uniformly random combination of the source packets. It can be shown that Random Linear Coding achieves optimal overhead for each receiver, because a random linear code remains a random linear code on any subset when the side information is subtracted out. However, its decoding involves Gaussian elimination, a step with undesirably high complexity. With no side information, fountain codes are a huge improvement to RLC in complexity. A key attribute to a fountain code is its *degree distribution*, which is a probability distribution on the integers. For each coded packet, its *degree* defines the number of uniform-randomly chosen packets which are XOR-ed to form the coded packet. A design issue is choosing the distribution appropriately, so that the decoder can perform an inexpensive iterative decoding (which involves picking a degree one packet, subtracting it from the rest and continuing till everything is decoded successfully). This can be achieved while keeping the complexity, defined by the expected value of the degree, of the order logarithmic in the size of the block being coded together. More generally, Raptor codes have an additional layer of coding on the source symbols to which the degree distribution is applied ([13]). This achieves an even lower complexity at the expense of a arbitrarily small loss of throughput. Fountain codes have been considered in the presence of side information previously by [5], [12], [6]. The iterative decoding process was studied by Darling and Norris in [4], in a context unrelated to coding. [8] first used this result in analyzing the decoding of LT codes. [4] gave a general result using fluid approximation to relate the number of received packets to the number of decoded packets for general limiting degree distributions. This result was also applied in [11] to investigate the intermediate performance of rateless codes.

Example I.1. Consider 10,000 *blocks* of data to be multicast to 1000 broadcast audience across an erasure free broadcast channel. Assume each user has some subset of at least 9000 blocks each with them as side information, which might be arbitrarily scattered over the source blocks. The results of this paper show the following implications on this example: (1) If we use the LT coding based on the standard distributions, it will take close to 10,000 block transmissions to complete iterative decoding (section II-A), and hence rendering the coding almost useless, as naive retransmission of everything itself takes 10,000 transmissions. (2) Degree distributions based on a simple truncation modification can ensure that the iterative decoding can be completed closer to less than 2000 retransmissions (section II-B). (3) There is no degree distribution that performs asymptotically optimal to capacity (i.e., loss of at least a constant factor is unavoidable asymptotically) (section II-C)

II. The Minimal Overhead for a Given fraction of the Data as Side Information

Let $\alpha > 0$ represent a parameter which corresponds to each receiver having an arbitrary subset of $(1-\alpha)n$ packets as side information. Given any code generated according to some degree distribution, each decoder first subtracts out its side information from each coded block received and then begins iterative decoding on the resultant encoded blocks. Let $P^{(n)}$ be a degree distribution with support on $[n]$ (i.e. corresponding to block length n), which was used at the encoder. Let $Q^{(n)}$ be the corresponding projection of $P^{(n)}$ obtained on any αn subset. We can explicitly write down the relation between them as follows (for $1 \leq i \leq \alpha n, 1 \leq j \leq n$):

$$Q^{(n)}(i) = \sum_{j=i}^{n} P^{(n)}(j) \frac{\binom{n\alpha}{i}\binom{n(1-\alpha)}{j-i}}{\binom{n}{j}} \tag{1}$$

Note that, for small i and j independent of n, the term $\frac{\binom{n\alpha}{i}\binom{n(1-\alpha)}{j-i}}{\binom{n}{j}}$ is approximated by $\binom{j}{i}\alpha^i(1-\alpha)^{j-i}$, but this not an approximation in general. Note that for LT code on block length n, the average degree is of the order $\log n$. Suppose as $n \to \infty$, the pointwise limits converge to valid probability distributions P and Q respectively on Z^+. The relation between the limiting distributions P and Q is given by the following relation, which follows by invoking the dominated convergence theorem.

Lemma II.1. *For limiting distributions P and Q formed as above, we have:*

$$Q(i) = \sum_{j=i}^{\infty} P(j) \binom{j}{i} \alpha^i (1-\alpha)^{j-i} \tag{2}$$

From the above lemma, it is easy to derive the following relation between the generating functions (defined as $P(z) = \sum_i P(i) z^i$):

Corollary II.2. *The generating function $Q(z)$ in terms of $P(z)$ is given as*

$$Q(z) = P(1 - \alpha + \alpha z)$$

Definition II.3. Let

$$r_n = \frac{\text{no. of coded packets received}}{\alpha n}$$

$$z_n = \frac{\text{no. of previously unkown packets recovered}}{\alpha n}$$

When the limits exist, let r denote the limit of r_n and $s_\alpha(r, P)$ denote the limit of z_n, where P is the generating function of the limit of the degree distributions used for coding on the entire source blocks.

Recall for a sequence of distributions that converges to a distribution with generating function, $P(z)$

when $\alpha = 1$, [4] shows that the recovered fraction (for a vanishingly small perturbed distribution) converges to[1]:

$$s(r, P) = \inf\{z \in [0, 1) : rP'(z) + \log(1-z) < 0\} \wedge 1$$

Using Corollary II.2, when the decoder subtracts out the side information from the received coded packets, it recovers an asymptotic fraction (represented on αn) corresponding to $s_\alpha(r, P) = s(r, Q)$. Thus:

Proposition 1. $s_\alpha(r, P) = \inf\{z \in [0, 1) : r\alpha P'(1-\alpha+\alpha z) + \log(1-z) < 0\} \wedge 1$

To define the overhead necessary for successful decoding of all source blocks, we thus need to solve the following optimization problem for each given $0 < \alpha \leq 1$:

$$r^*_\alpha = \min_P r \tag{3}$$

$$\text{Subject to} : s_\alpha(r, P) = 1 \tag{4}$$

A trivial lower bound for r^*_α for any $0 < \alpha \leq 1$ is 1. We can also obtain an easy upper bound by ignoring the side information and using the soliton distribution. Since we know that asymptotically n packets suffice to recover all the n packets without even considering the side information, this achieves an $r_\alpha = \frac{n}{n\alpha} = \frac{1}{\alpha}$. This gives us:

$$1 \leq r^*_\alpha \leq \frac{1}{\alpha} \tag{5}$$

Clearly, this implies $r^*_1 = 1$, which is attained by the capacity achieving soliton distribution used for LT codes [7]. Some natural questions arise for the problem under consideration: Is $r^*_\alpha = 1$ for $\alpha < 1$? (i.e., can we have capacity achieving degree distributions for general α?) How bad is the Soliton distribution as a solution to the optimization problem in 3? (We know that its no worse than the upper bound even after throwing away the side information, but could it actually achieve a better ratio because of side information?) If the Soliton is bad, how do we design degree distributions that do well for $\alpha < 1$?

A. Performance of the Soliton distribution

Proposition 2. *For a sequence of distributions with a limiting Soliton distribution, for recovering all unknown packets ($z_n \to 1$), we need $r > 1/\alpha$. This means that side information gives no advantage for complete recovery.*

Proof: Consider a sequence of distributions that converge to the Soliton distribution, whose generating function, $P(z) = \sum_{i\geq 2} \frac{z^i}{i(i-1)}$. The recovered fraction, z_n (of a vanishingly small perturbed distribution) converges to: $s(r, Q) \triangleq \inf\{z \in [0, 1) : rQ'(z) + \log(1-z) < 0\} \wedge 1$. Consider:

$$\begin{aligned} & rQ'(z) + \log(1-z) \\ &= r\frac{d}{dz}P(1-\alpha+\alpha z) + \log(1-z) \\ &= r\alpha\frac{d}{dz}\left(\sum_{i\geq 2}\frac{(1-\alpha+\alpha z)^i}{i(i-1)}\right) + \log(1-z) \\ &= r\alpha\sum_{i\geq 1}\frac{(1-\alpha+\alpha z)^i}{i} + \log(1-z) \\ &= r\alpha|\log\alpha| + (r\alpha-1)|\log(1-z)| \end{aligned}$$

From the above equation, it is clear that $rQ'(z) + \log(1-z) > 0 \ \ \forall z \in [0, 1)$ iff $r\alpha > 1$. ■

B. k-lifted Soliton distribution

Consider the $k-$lifted Soliton distribution defined by the following generating function (where k will be chosen later):

$$P(z) = \sum_{i\geq k+1}\frac{k}{i(i-1)}z^i$$

The idea of shifting the distributions was also considered by the authors in [6] (although the parameters are different).

Proposition 3. *For $k = \lfloor 1/1.82\alpha \rfloor$, the k-lifted Soliton distribution requires at most $r = 1/(\alpha\lfloor 1/1.82\alpha \rfloor)$ for complete recovery.*

Proof: First we lower bound $Q'(z)$ for $z \in [0, 1)$.

$$\begin{aligned} Q'(z) &= \sum_{i\geq k}\alpha k\frac{(1-\alpha+\alpha z)^i}{i} \\ &= \alpha k\left(-\log(\alpha-\alpha z) - \sum_{i=1}^{k-1}\frac{(1-\alpha+\alpha z)^i}{i}\right) \\ &\geq \alpha k\left(|\log\alpha| + |\log(1-z)| - H_{k-1}\right) \\ &\Big(\text{ where } H_k \text{ is the } k^{th} \text{ Harmonic number}\Big) \\ &\geq \alpha k\Big(|\log\alpha| + |\log(1-z)| \\ &\qquad - \left(\log k + \gamma + \frac{1}{2n+\frac{1}{3}}\right)\Big) \\ &\text{(Using a bound from [3],} \\ &\text{where } \gamma \approx 0.578 \text{ is the Euler constant)} \\ &\geq \alpha k\left(|\log\alpha| - \log(1.82k) + |\log(1-z)|\right) \end{aligned}$$

[1] Although the conditions stated don't apply verbatim here either, the justification for it comes from Lemma 1 of [11]

Given r, we thus have for $z \in [0,1)$:

$$\begin{aligned} & rQ'(z) + \log(1-z) \\ & \geq r\alpha k\left(|\log\alpha| - \log(1.82k)\right) + \\ & (r\alpha k - 1)|\log(1-z)| \end{aligned}$$

The choice of $k = \lfloor 1/1.82\alpha \rfloor$ ensures that the term above is non negative. For such a k, that the second term is also non-negative for $z \in [0,1)$ for the given choice of r, implying $s(r,Q) = 1$.

C. Lower Bounds on the Optimal Overhead

In this section, we prove lower bounds for the optimal decoding overhead required for any degree distribution for any given fraction α defining the side information. This shows that r^*_α could be strictly greater than 1 for general α, which was also conjectured in [12]. This is accomplished by considering an intermediate performance problem inspired by [11]. We now move on to the details of the argument. The optimization in Eq (3) can be rewritten as the following:

$$r^*_\alpha = \min_P r$$

Subject to ($\forall\ 0 \leq t < 1$):

$$\sum_{i\geq 1} r\alpha i p_i (1-\alpha+\alpha t)^{i-1} + \log(1-t) \geq 0 \quad (6)$$

Choose some large integer m and consider the following related problem (which can also be interpreted as an intermediate performance problem):

$$r_{\alpha,m} = \min_P r$$

Subject to:

$$\sum_{i\geq 1} r\alpha i p_i (1-\alpha+\alpha t)^{i-1} + \log(1-t) \geq 0 \quad (7)$$

$$\forall \quad 0 \leq t < 1 - 1/\alpha(m+1)$$

We just replaced the constraints in Eq (6) with a proper subset in Eq (7) and thus,

$$r_{\alpha,m} \leq r^*_\alpha \quad \forall \quad m$$

Further it can be verified that:

$$\begin{aligned} i(1-\alpha+\alpha t)^{i-1} < m(1-\alpha+\alpha t)^{m-1} \\ \forall \quad i > m,\ t \leq 1 - 1/\alpha(m+1) \end{aligned}$$

The above condition can be shown to imply that we can restrict attention to $p_i = 0\ \forall\ i > m$ in the above optimization defining $r_{\alpha,m}$. Hence we obtain:

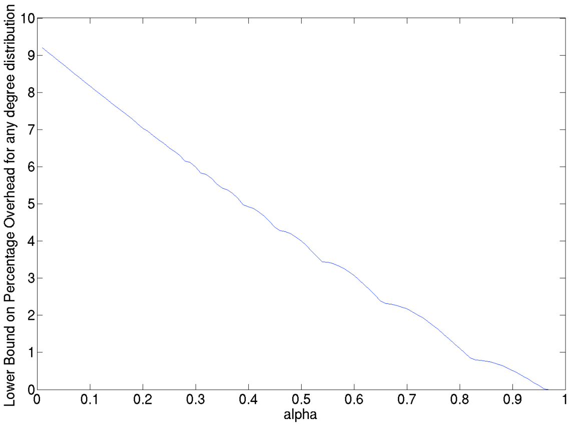

Fig. 1. A plot of our lowerbound on the overhead, represented as a percentage of αn.

$$r_{\alpha,m} = \min_P r$$

Subject to:

$$\sum_{i=1}^{m} r\alpha p_i i (1-\alpha+\alpha t)^{i-1} + \log(1-t) \geq 0$$

$$\forall \quad 0 \leq t < 1 - 1/\alpha(m+1)$$

By setting $a_i = rp_i$, we get the following LP:

$$r_{\alpha,m} = \min_A \sum_{i=1}^{m} a_i$$

subject to:

$$\sum_{i=1}^{m} \alpha i (1-\alpha+\alpha t)^{i-1} a_i \geq -\log(1-t)$$

$$\forall \quad 0 \leq t < 1 - 1/\alpha(m+1)$$

The dual of the above LP can be written as :

$$\xi_{\alpha,m} = \max_\mu \int_{t\in[0,1-1/\alpha(m+1))} -log(1-t)d\mu(t)$$

Subject to: $(\forall\ 1 \leq i \leq m)$

$$\int_{t\in[0,1-1/\alpha(m+1))} \alpha i (1-\alpha+\alpha X)^{i-1} d\mu(t) \leq 1$$

where, μ is a measure on $[0, 1-1/\alpha(m+1)]$. Further, any feasible solution above is a lower bound to

$$\xi_{\alpha,m} \leq r_{\alpha,m} \leq r^*_\alpha$$

Thus, we can optimize over a subclass to obtain the required lower bounds. One possibility is to take μ as a discrete distribution with finite support.

In this case, the dual defined above becomes a finite dimensional linear program, for which we can use linear programming to compute the bound. We plot the best bounds we have numerically computed using this approach in the Figure 1 rescaled to be expressed as percentage overhead corresponding to each α in the range $0, 1$. As we can see, for most α (except when it is very close to 1), we were able to find a lower bound strictly greater than the 1.

III. A Solution using Systematic LT Codes

Systematic fountain codes were introduced in [13]. Let $x_1, \ldots, x_k$ be the original source symbols. These are first transformed into *intermediate symbols*, $y_1, \ldots, y_k$ to which a degree distribution is then applied to generate the actual code symbols. The intermediate symbols are designed such that when the fountain code is applied to them, it results in the original source symbols at some subset of k positions in the first $k(1 + \epsilon)$ generated code symbols, where $\epsilon > 0$ is small. The code to be applied with side information is a slight modification of the systematic code in the following sense: After creating the intermediate symbols, the encoder starts generating output symbols omitting the first $k(1+\epsilon)$ symbols which contain the systematic part. The decoders put together the side information they have along with the received code symbols, and these symbols together form an LT code over the *intermediate symbols*, by design. Thus, each receiver is able to optimally recover the intermediate symbols, and from them subsequently, the source symbols. Approaches based on systematic raptor codes are also discussed in [12], [5].

Remark 1. *Note that this resultant systematic code may not be generated by employing any degree distribution directly on the source symbols $x_1, \ldots, x_k$. This is the reason why the lower bound computed previously does not apply to this type of coding.*

Note that for systematic LT codes, the complexity per symbol can be made to scale logarithmically in block length. Alternately, one could also use a general systematic raptor code. In such a case, the decoding complexity at the receivers is constant time, but the encoder needs to perform a Gaussian elimination step for calculating the intermediate symbols, which involves a linear complexity per each intermediate symbol generated. In this sense, one has the option of suffering a per symbol linear complexity at the encoder (instead of the logarithmic complexity) in exchange for a constant per symbol complexity at each of the decoders (as opposed to logarithmic) by using more general systematic Raptor codes for the side information problem.

References

[1] Z. Bar-Yossef, Y. Birk, T. Jayram, and T. Kol. Index coding with side information. *IEEE FOCS*, 2006.

[2] Y. Birk and T. Kol. Coding-on-demand by an informed source (iscod) for efficient broadcast of different supplemental data to caching clients. *IEEE Transactions on Information Theory*, 52(6):28252830 2006. Earlier version appeared in INFOCOM 98.

[3] Ch.-P Chen and F. Qi. The best bounds of the n-th harmonic number. *Global Journal of Mathematics and Mathematical Sciences*.

[4] R. Darling and J. Norris. Structure of large random hypergraphs. *Annals of Applied Probability*, Vol 15, pp 125-152 2005.

[5] M. Fresia and L. Vandendorpe. Distributed source coding using raptor codes. *Proc. of IEEE GLOBECOM*, 2007.

[6] A. Hagedorn, S. Agarwal, D. Starobinski, and A. Trachtenberg. Rateless coding with feedback. *Proc. of IEEE Infocom*, 2009.

[7] M. Luby. Lt codes. *IEEE FOCS*, 2002.

[8] E. Maneva and A. Shokrollahi. New model for rigorous analysis of lt codes. *IEEE International Symposium on Information Theory (ISIT)*, pages 2677-2679 2006.

[9] J. Metzner. An improved broadcast retransmission protocol. *IEEE Transactions on Communications*, Vol 32 No. 6, June 1984.

[10] S. El Rouayheb, A. Sprintson, and C. N. Georghiades. On the index coding problem and its relation to network coding and matroid theory. *submitted to IEEE Trans. on Information theory*, 2009.

[11] S. Sanghavi. Intermediate performance of rateless codes. *IEEE Information Theory Workshop*, September 2007.

[12] D. Sejdinovic, R.J. Piechocki, and A. Doufexi. Fountain coding with decoder side information. *Proc. of IEEE ICC*, pp. 4477 - 4482 2008.

[13] A. Shokrollahi. Raptor codes. *IEEE Trans on Information Theory*, vol. 52, pp. 2551-2567 2006.