Coding for Network Coding

Andrea Montanari* and Rüdiger Urbanke†

November 19, 2007

Abstract

We consider communication over a noisy network under randomized linear network coding. Possible error mechanism include node- or link- failures, Byzantine behavior of nodes, or an over-estimate of the network min-cut. Building on the work of Kötter and Kschischang, we introduce a probabilistic model for errors. We compute the capacity of this channel and we define an error-correction scheme based on random sparse graphs and a low-complexity decoding algorithm. By optimizing over the code degree profile, we show that this construction achieves the channel capacity in complexity which is jointly quadratic in the number of coded information bits and sublogarithmic in the error probability.

1 Introduction

Consider a wire-line communication network modeled as a directed acyclic (multi-)graph with edges of unit capacity. A sources wants to communicate information to a set of receivers. If we allow *processing* of information at nodes in the network then the achievable throughput is in general higher than what can be achieved by schemes that only allow *routing* [1, 9]. Schemes that employ processing are referred to as *network coding* schemes.

The standard assumption in the network coding literature is that no errors are introduced within the network or, equivalently, that sufficiently powerful error-correcting codes are employed on the links at the physical layer. However a number of error sources (e.g., malicious or malfunctioning nodes) cannot be neglected. We consider a probabilistic model for transmission errors that builds upon the work of Kschischang and Kötter [7, 14]. We compute the information theoretic limit on point-to-point communication for this model (the channel capacity) and define a coding scheme based on a sparse-graph construction and a low-complexity iterative decoding algorithm. We show that the parameters of the construction can be optimized analytically and, remarkably, the optimized scheme achieves the channel capacity. This is the second channel model for which iterative schemes can be shown to achieve capacity (the first one being the binary erasure channel; this was shown in the seminal work of Luby, Mitzenmacher, Shokrollahi, Spielman, and Steman [10]).

2 Network Coding: Background and Related Work

Assume that an information source generates h symbols per unit time. The integer h is referred to as the source rate. Information is encoded at the sender in packets of length N with entries from a finite field $\mathbb{F}_q$. The network is assumed to be synchronous and without delay. As a consequence, packets are aligned at the destination at regular time intervals.

*Departments of Electrical Engineering and Statistics, Stanford University, Stanford CA-9305, USA

†School of Computer and Communication Sciences, EPFL, 1015 Lausanne, CH

Keywords: Sparse graph codes, probabilistic channel models, Shannon channel capacity, network coding

The most common scenario studied in this context is a multi-cast one in which the source aims at communicating the same information to a set of receivers (distinct nodes in the same network.) The fundamental theorem of network coding states that this is possible using network coding (i.e., processing at the nodes) if the values of the min-cuts from the source to any of the receivers is at least h [1]. Moreover, *linear* network coding suffices [9]. This means that processing at the nodes can be limited to forwarding packets which are linear (over $\mathbb{F}_q$) combinations of incoming packets. Finally, it is not necessary to choose the local encoding functions at the nodes carefully. *Random* linear combinations are sufficient with probability close to one, provided the cardinality of the field is large enough [6, 8]. For a general introduction into network coding we refer the reader to [4, 16, 17].

The preferred method to implement random linear network codes is to include "headers" in the packets of length N [2]. The role of the headers is to "record" the coefficients used in the local encoding functions so that the receiver can be oblivious to the network topology and to the specific local encoding functions used. In more detail, assume that we send ℓ packets. The header of each packet is then an element of $(\mathbb{F}_q)^\ell$, where the header of the i-th source packet, $i \in [\ell]$, is the all-zero tuple, except for an identity element at position i. Recall that nodes forward packets which are linear combinations of the incoming packets. Therefore, if the header of a packet somewhere in the network reads $(\beta_1, \ldots, \beta_\ell)$, $\beta_i \in \mathbb{F}_q$, then we know that this packet is the linear combination of the ℓ original source packets, where the i-th original source packet has "weight" β_i. The significant advantage of such a scheme is that the receivers can be oblivious to the topology and the local encoding functions. Of course we pay some price; if we use headers then only $m = N - \ell$ of the N symbols of each packet are available for information transmission. Our subsequent discussion assumes this "oblivious" model.

So far we assumed that errors neither occur during transmission nor during processing. If the channel or the processing are noisy, one can use coding to combat the noise. Note that if we stack the ℓ source packets of length N on top of each other then we get an $\ell \times N$ matrix over $\mathbb{F}_q$ whose, lets say, left $\ell \times \ell$ submatrix (the collection of headers) is the identity matrix.

Formally, a code $\mathfrak{C}$ is a collection of $\ell \times N$ matrices with elements in $\mathbb{F}_q$, such that each $M \in \mathfrak{C}$ takes the form $M = [\underline{1} \,|\, \underline{x}]$. Here, $\underline{1}$ is the $\ell \times \ell$ identity matrix and $\underline{x}$ is an $\ell \times m$ matrix ($m = N - \ell$). We say that M is in *normal* form. The code $\mathfrak{C}$ is thus equivalently described by a collection of $\ell \times m$ matrices $\{\underline{x}\}$. The *rate* of the code is defined as the ratio of the number of information q-bits that can be conveyed by the choice of codeword ($\log_q |\mathfrak{C}|$) to the number of transmitted symbols ($N\ell$):

$$R(\mathfrak{C}) = \frac{\log_q |\mathfrak{C}|}{N\ell}\,. \tag{1}$$

Before the source packets are transmitted we multiply M from the left by an $\ell \times \ell$ random invertible matrix with components in $\mathbb{F}_q$. This "mixes" the rows of M and ensures that regardless of the network topology and the location where the errors are introduced, the effect of the errors on the normalized form is uniform. We then transmit each resulting row as one packet.

Upon transmission of M, a "corrupted" version Q of the codeword is received. Without loss of generality, we assume that Q is brought back into normal form $Q = [\underline{1} \,|\, \underline{y}]$ by Gaussian elimination.[1] Following Kötter and Kschischang [7], we model the net effect of the transmission- and the processing-"noise" as a low-rank perturbation of $\underline{x}$. More precisely, we assume that

$$\underline{y} = \underline{x} + \underline{z}, \tag{2}$$

where $\underline{z}$ is an $\ell \times m$ matrix over $\mathbb{F}_q$ of $\mathrm{rank}(\underline{z}) = \ell\omega$, $\omega \in [0, 1]$. We call $\ell\omega$ the *weight* of the error, and ω the *normalized* weight.

Define the *distance* of two codewords $\underline{x}$ and $\underline{x}'$ as $d(\underline{x}, \underline{x}') = \mathrm{rank}(\underline{x} - \underline{x}')$ and the *minimum* distance $d(\mathfrak{C})$ of the code $\mathfrak{C}$ as the minimum of the distances between all distinct pairs of codewords. The *normalized* minimum distance is $\delta(\mathfrak{C}) = d(\mathfrak{C})/\ell$. It is shown in [7] that $d(\cdot, \cdot)$ is a true distance metric;

[1] In principle it might be that the received matrix cannot be brought in this form because its first ℓ columns have rank smaller than ℓ. However, within the probabilistic model which we discuss in the following, the rank deficiency is small with high probability and can be eliminated by a small perturbation.

in particular it fulfills the triangle inequality. Therefore, given a code $\mathfrak{C}$ of minimum distance $d(\mathfrak{C})$ a simple *bounded distance* decoder can correct all errors of weight $s = (d(\mathfrak{C}) - 1)/2$ or less. A bounded distance decoder is an algorithm that, given a received word $\underline{y}$, decodes $\underline{y}$ to the unique word within distance s if such a word exists and declares an error otherwise. Bounded distance decoders are popular since a suitable algebraic structure on the code often ensures that bounded distance decoding can be accomplished with low complexity.

The *bounded-distance* error-correcting capability of a code is defined as $\omega(\mathfrak{C}) = d(\mathfrak{C})/(2\ell) = \delta(\mathfrak{C})/2$. Kötter and Kschischang showed that the optimal trade-off between $R(\mathfrak{C})$ and $\omega(\mathfrak{C})$ is given by an appropriate generalization of the "Singleton bound." In the limit $N \to \infty$, with $\ell = \lambda N$, the maximal achievable rate for the parameters $\omega, \lambda \in [0, 1/2]$, call it $C_{\text{Singleton}}(\lambda, \omega)$, is given by

$$C_{\text{Singleton}}(\lambda, \omega) \quad = (1 - \lambda)(1 - 2\omega). \tag{3}$$

Note that $C_{\text{Singleton}}(\lambda, \omega)$ is the maximum achievable rate for a guaranteed error correction in an adversarial channel model. It is also the maximal achievable rate in a probabilistic setting if we are limited to bounded distance decoding. Remarkably, Kötter and Kschischang found a generalization of Reed-Solomon codes that achieves this bound.

3 Main Results

We are interested in a probabilistic (as opposed to adversarial) channel model. More precisely, we assume that in (2) the perturbation $\underline{z}$ is chosen uniformly at random from all matrices in $(\mathbb{F}_q)^{\ell \times m}$ of rank $\ell\omega$. We assume that the parameters λ and ω are fixed and consider the behavior of the channel as we increase N. We refer to our channel model as the *symmetric network coding channel* with parameters λ and ω, denoted by $\text{SNC}(\lambda, \omega)$.

Proposition 3.1 (Channel Capacity)**.** *The capacity of* $\text{SNC}(\lambda, \omega)$ *is*

$$C(\lambda, \omega) = 1 - \lambda - \omega + \lambda\omega^2. \tag{4}$$

Discussion: In the definition of capacity we implicitly assume that the error probability ω is not a function of N. Depending on the underlying physical error mechanism this may or may not be the case. Note that for small ω, $C(\lambda, \omega) \approx 1 - \lambda - \omega$, whereas $C_{\text{Singleton}}(\lambda, \omega) \approx 1 - \lambda - 2(1 - \lambda)\omega$. Fig. 1 compares $C(\lambda, \omega)$ with $C_{\text{Singleton}}(\lambda, \omega)$ and shows the points that are achievable according to Theorem 3.2.

Theorem 3.2 (Capacity-Achieving Iterative Code Construction)**.** *For any* $\lambda, \omega \in (0, 1)$ *such that* $(1 - \lambda)/\lambda$ *is an integer multiple of* ω*, any* $R < C(\lambda, \omega)$*, and any* $\pi > 0$ *there exists an error correcting code and a decoding algorithm that achieves symbol error probability smaller than* π*, with* $O(N^4 \log\log(1/\pi))$ *decoding complexity.*

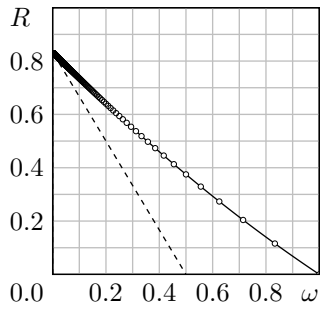

Figure 1: Comparison of $C(\lambda, \omega)$ (solid line) with $C_{\text{Singleton}}(\lambda, \omega)$ (dotted line) for $\lambda = 1/6$. The points on the curve $C(\lambda, \omega)$ that are achievable by the low-complexity iterative scheme are shown as dots.

Discussion: The complexity of the scheme is given as $O(N^4)$. But note that the number of transmitted information symbols is $N^2\lambda R$. Therefore, if we measure the complexity per transmitted information symbol then it is only quadratic.

Note also that the complexity scales much better with the target error probability than for usual sparse graph codes (where it is at least linear in $\log(1/\pi)$).

3.1 Code Construction

Fig. 2 shows our coding scheme. Each row corresponds to a packet of length N. The $\ell \times \ell$ identity matrix

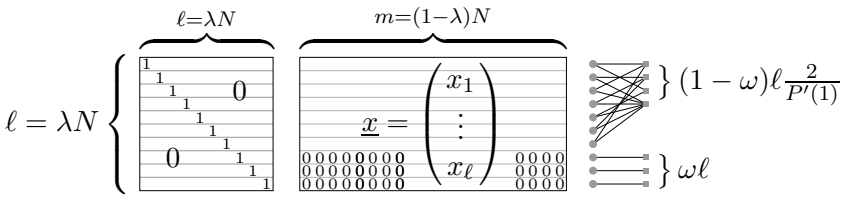

Figure 2: Coding scheme. The last $\omega\ell$ rows of $\underline{x}$ are zero. The first $(1-\omega)\ell$ obey a set of linear constraints represented by the bipartite graph shown on the right-hand side.

$\underline{1}$ is shown on the left-hand side, whereas the $\ell \times m$ matrix on the right-hand side represents $\underline{x}$. Each $\underline{x}$ corresponds to a codeword of $\mathfrak{C}$. Not all $\underline{x}$ are allowed. Here are the constraints that $\underline{x}$ must fulfill to be a codeword. The bottom $\omega\ell$ rows are identical to zero. The top $(1-\omega)\ell$ rows are constrained by a linear system of equations. These are indicated by the bipartite graph on the right-hand side, according to the standard graphical representation used for low-density parity-check codes [5, 13]. More precisely, we have

$$\hat{\mathbb{H}} \begin{pmatrix} x_1^T \\ \vdots \\ x_{(1-\omega)\ell}^T \end{pmatrix} = 0. \tag{5}$$

The matrix $\hat{\mathbb{H}}$ has the following structure. Start with a "sparse" $((1-\omega)\ell r) \times ((1-\omega)\ell)$ $\{0,1\}$-valued matrix $\mathbb{H}$. The matrix $\mathbb{H}$ has exactly 2 non-zero entries along each column. Further, the fraction of rows that contain exactly i non-zero entries is equal to P_i, where $P(x) = \sum_i P_i x^i$ is a given *degree distribution* (in particular, it fulfills $P_i \geq 0$ and $P(1) = 1$.) In the following, we shall say that P has *bounded support* if $P_i = 0$ for i larger than some $n_{\max} < \infty$ or, equivalently, if $P(x)$ is a polynomial.

The matrix $\mathbb{H}$ is represented by the graph. *Circles* (on the right-hand side in Fig. 2) correspond to the columns of $\mathbb{H}$ and *squares* (on the left-hand side) correspond to the rows of $\mathbb{H}$. There is an edge between a circle and an edge iff there is a non-zero entry at the corresponding row and column of $\mathbb{H}$. Following the iterative coding literature, we refer to the circles as the *variable* nodes, to the squares as the *check* nodes, and we call this graph a *Tanner* graph. To get the matrix $\hat{\mathbb{H}}$ we "lift" $\mathbb{H}$ by replacing each of its non-zero elements by an $m \times m$ invertible matrix with elements in $\mathbb{F}_q$. We can visualize this by attaching these invertible matrices as labels to the corresponding edges.

We claim that for any choice of the matrix $\hat{\mathbb{H}}$ compatible with the degree distribution $P(x)$ the rate of the code is *at least*

$$R(\omega, \lambda, P) = (1-\lambda)(1-\omega)\Big(1 - \frac{2}{P'(1)}\Big). \tag{6}$$

To see this, note that the matrix $\underline{x}$ is of dimension $\ell \times m$ and has entries in $\mathbb{F}_q$. Since the last $\omega\ell$ rows have to be zero this reduces the degrees of freedom by $m\omega\ell$. Further, there are $m(1-\omega)\ell\frac{2}{P'(1)}$ linear constraints, taking away *at most* that many further degrees of freedom (and possibly less because of linear dependencies). We get the claim by dividing the remaining degrees of freedom by $N\ell$, in accordance with (1).

So far we have explained how to construct a code. We define an *ensemble* of codes by (i) picking a matrix $\mathbb{H}$ uniformly from all matrices that have degree profile $P(x)$ according to the configuration model and, (ii) picking the labels (the $m \times m$ invertible matrices) for all edges uniformly and independently for each edge. We denote the resulting ensemble by $\mathcal{C}(N, \lambda, \omega, P(x))$.

Discussion: In Fig. 2 all linear constraints are on the rows of $\underline{x}$. An entirely equivalent formulation is to apply the linear constraints to the columns of $\underline{x}$ instead; i.e., set the last ωm columns of $\underline{x}$ to zero and apply a set of linear constraints on the first $(1-\omega)m$ columns of $\underline{x}$. All subsequent statements apply also to this case and yield identical results if we let N tend to infinity. For the sake of simplicity, we limit our discussion to the scheme of Fig. 2. In a practical implementation, however, there can be reasons to prefer one scheme over the other. For instance, the iterative decoder discussed in the next section might be more effective on a larger Tanner graphs. This suggests to use the construction in Fig. 2 if $\ell > m$ and the 'transposed' one otherwise.

3.2 Encoding and Decoding Algorithm

Assume that the parameters of the model (N, λ, ω, and $P(x)$) are fixed and that we have chosen one particular code from the ensemble $\mathcal{C}(N, \lambda, \omega, P(x))$. At the source we are given $RN\ell$ symbols over $\mathbb{F}_q$ (the information we want to transmit). We need to map each of these $q^{RN\ell}$ possible information vectors to a distinct codeword $\underline{x}$. This is the *encoding* task. In principle this can be done by solving a linear system of equations, starting with (5). A brute force approach, however, has complexity $O(N^6)$. Fortunately, one can exploit the sparseness of the matrix $\mathbb{H}$ to reduce the encoding complexity to $O(N^3)$. The basic idea is to bring $\mathbb{H}$ into upper-triangular form by using only row and column permutations but no algebraic operations. As proved in [12], this can be done with high probability if $P''(1)/P'(1) > 1$. We will see in Section 4, cf. Lemma 4.5, that this condition is always fulfilled. Further details on the efficient implementation of the encoder will be discussed in a forthcoming publication. We are currently mainly concerned with the decoding problem.

The receiver sees the perturbed matrix $\underline{y}$. An equivalent description of our channel model is the following. Each row of $\underline{y}$ is the result of adding to the corresponding row of $\underline{x}$ a uniformly random element of a subspace W of $(\mathbb{F}_q)^m$. The subspace W is itself uniformly random under the condition $\dim(W) = \omega\ell$.[2]

Recall that by assumption the last $\ell\omega$ rows of $\underline{x}$ are zero. In fact, in order to achieve reliable transmission we need to modify the scheme described so far and set the last $\ell\omega'$ rows of $\underline{x}$ to 0, where $\omega' > \omega$ is arbitrarily close to ω. This modification reduces the rate by a quantity that can be made arbitrarily small. Since the perturbation has dimension $\ell\omega$, the last $\ell\omega'$ rows of $\underline{y}$ will span W with high probability as $N \to \infty$. A basis of W is then obtained by reducing these rows via gaussian elimination.

We therefore assume hereafter that W is known and, to avoid cumbersome notation, we set $\omega' = \omega$. The decoding task consists in finding the perturbations for the first $(1-\omega)\ell$ rows of $\underline{y}$. If we subtract these perturbations form $\underline{y}$, we have found $\underline{x}$. Throughout the description, given two sets of vectors U_1 and U_2, we let $U_1 + U_2 \equiv \{u_1 + u_2 : u_1 \in U_1, u_2 \in U_2\}$ and, for a given vector x, $x + U \equiv \{x\} + U$. Finally, given a matrix $\mathfrak{h} \in (\mathbb{F}_q)^{m\times m}$, $U\mathfrak{h} \equiv \{u\mathfrak{h} : u \in U\}$ (vectors are always thought as row vectors).

We proceed in an iterative fashion. The basic principle is easily understood. We know that $x_i \in y_i + W$. In words, we know that x_i lies in a given affine subspace. Consider a check node a and, without loss of generality, let its neighbors be $1, \dots, d$. Let $\mathfrak{h}_{ia}$, $i = 1, \dots, d$, denote the corresponding edge labels. As we discussed earlier, each such edge label is an $m \times m$ invertible matrix with entries in $\mathbb{F}_q$. By the definition of the code, $\sum_{i=1}^{d} x_i\mathfrak{h}_{i,a} = 0$. In particular, this means that $x_1 \in (\sum_{i=2}^{d} x_i\mathfrak{h}_{i,a})\mathfrak{h}_{1,a}^{-1}$. Since we know that $x_i \in y_i + W$, this implies that

$$x_1 \in [(y_2 + W)\mathfrak{h}_{2,a} + \cdots + (y_d + W)\mathfrak{h}_{d,a}]\mathfrak{h}_{1,a}^{-1}\,.$$

[2] As discussed in the introduction, the underlying physical process is the following: we add the headers to the rows of $\underline{x}$; we scramble the rows of M multiplying it by a random invertible matrix in $(\mathbb{F}_q)^{m\times m}$; we send the resulting packets; the channel perturbs these packets; the receiver collects the perturbed packets, stacks them up to a matrix Q, brings the matrix back into normal form, and "strips off" the headers.

Since we also know that $x_1 \in y_1 + W$, this implies that

$$x_1 \in (y_1 + W) \cap \{[(y_2 + W)\mathfrak{h}_{2,a} + \cdots + (y_d + W)\mathfrak{h}_{d,a}]\mathfrak{h}_{1,a}^{-1}\}\,. \tag{7}$$

The actual decoder is most conveniently described (and analyzed) as a 'message passing' algorithm, with messages being sent along the edges of the Tanner graph. Messages are affine subspace of $(\mathbb{F}_q)^m$. They are sent in rounds. First we send messages from the variable nodes to the check nodes. We process the incoming messages at the check nodes and then send messages on all edges from the check nodes to the variable nodes. This concludes one *iteration of message passing*.

In more detail, the message sent from variable node i to check node a in the t-th iteration is an affine subspace $W_{i\to a}^{(t)}$ of $(\mathbb{F}_q)^m$. If variable node i is connected to check node a, let $\bar{a}$ denote the second check node that is connected to i (recall that each variable node has exactly two neighbors). Variable nodes do not perform any non-trivial processing of the messages, and check-to-variable node messages coincide with variable-to-check ones $W_{i\to a}^{(t)} = W_{\bar{a}\to i}^{(t)}$.

For $t = 0$ we have $W_{i\to a}^{(0)} = y_i + W$ for all variable nodes i and all check nodes a. Further, let ∂a denote all neighbors of a check node a. According to the above discussion, we apply for $t \geq 0$ the recursion

$$W_{i\to a}^{(t+1)} = (y_i + W) \cap \Big\{\Big[\sum_{j\in\partial\bar{a}\setminus i} W_{j\to\bar{a}}^{(t)}\mathfrak{h}_{j,\bar{a}}\Big]\mathfrak{h}_{i,\bar{a}}^{-1}\Big\}\,. \tag{8}$$

If, after some iterations, $\dim(W_{i\to a}^{(t)} \cap W_{i\to\bar{a}}^{(t)}) = 0$, then have determined the i-th row of $\underline{x}$, namely $W_{i\to a}^{(t)} \cap W_{i\to\bar{a}}^{(t)} = \{x_i\}$.

Our (main) Theorem 4.5 affirms that. for given parameters λ and ω, the degree distribution $P(x)$ can be chosen in such a way that the rate of the overall code approaches the capacity arbitrarily closely and that the decoder succeeds with high probability when the packet size tends to infinity.

4 Proofs

In the next section we state a few auxiliary lemmas on the behavior of the message-passing decoder and prove Theorem 3.2. The lemmas are then proved in Section 4.2. Finally, the capacity of the network coding channel is computed in Section 4.3.

4.1 Auxiliary Results and Proof of the Main Theorem

To start we can simplify our proof in two manners. First, by symmetry of the channel and the message-passing rules, we can assume that the all-zero matrix $\underline{x}$ was transmitted and we need only analyze the behavior of the decoder for this case. Notice that, under this assumption, the messages $W_{i\to a}^{(t)}$ are *linear* subspaces (as they must contain the transmitted vectors $x_i = 0$.) Second, as we discussed in Section 3.2, the first step of the decoding procedure consists of learning the perturbing subspace W. Because of the special structure of the matrix $\underline{x}$ (the last $\omega\ell$ rows are zero) this is accomplished by a simple inspection. We therefore assume in all that follows that W is known and that the all-zero matrix was transmitted.

Throughout this section we let P be a distribution over the integers and let G be a random multi-graph over $\ell(1-r)$ nodes with degree distribution P. The graph G is drawn according to the configuration model and the code is constructed from G as described in the previous section. Since variable nodes have degree 2, we can think of G either as a multi-graph over the check nodes, or as a bipartite graph over check *and* variable nodes.

It is also useful to define the 'edge perspective' degree distribution

$$\rho_n = \frac{nP_n}{\sum_{n'\geq 0} n'P_{n'}}\,. \tag{9}$$

For a uniformly random edge in G, let $W^{(t)}$ be the associated message (that, we recall, is an affine subspace in $(\mathbb{F}_q)^m$). The key step in the analysis is to notice that the dimension of $W^{(t)}$ satisfies a simple recursion.

First consider $n-1$ independent and uniformly random linear subspaces $V_1, \dots, V_{n-1} \subseteq (\mathbb{F}_q)^m$ of dimensions $d_1, \dots, d_{n-1}$, respectively. Let V be a fixed subspace of $\dim(V) = D$, and define

$$K^{(n)}_{m,D}(d \,|\, d_1, \dots, d_{n-1}) \equiv \mathbb{P}\{\dim(V \cap (V_1 + \cdots + V_{n-1})) = d\},. \tag{10}$$

The probability kernel $K^{(n)}_{m,D}$ admits an explicit albeit cumbersome expression in terms of Gauss polynomials. Fortunately, we do not need its exact description in the following.

We define a sequence of integer-valued random variables $\{D^{(t)}\}_{t\geq 0}$ recursively as follows. For $t = 0$ we let $D^{(0)} = \ell\omega$ identically. For $t \geq 0$, choose n with distribution ρ_n, and draw $D^{(t)}_1, \dots, D^{(t)}_{n-1}$ iid copies of $D^{(t)}$. Then, the probability of $D^{(t+1)} = d$, conditioned on the values $D^{(t)}_1 = d_1, \dots, D^{(t)}_{n-1} = d_{n-1}$ coincides with Eq. (10) where $D = \ell\omega$. In formulae,

$$\mathbb{P}\{D^{(t+1)} = d\} = \sum_{n\geq 1} \rho_n \sum_{d_1 \dots d_{n-1}} K^{(n)}_{m,D}(d \,|\, d_1, \dots, d_{n-1})\, \mathbb{P}\{D^{(t)} = d_1\} \cdots \mathbb{P}\{D^{(t)} = d_{n-1}\}\,. \tag{11}$$

The sequence $\{D^{(t)}\}$ accurately tracks the dimension of $W^{(t)}$ as stated below.

Lemma 4.1 (Density Evolution on a Graph versus Density Evolution on a Tree)**.** *For any degree distribution P with bounded support (i.e. such that $P_n = 0$ for all n large enough) and any $t \in \mathbb{N}$ there exists a sequence $\epsilon(\ell, t)$ with $\epsilon(\ell, t) \downarrow 0$ as $\ell \to \infty$, such that, for any m, and ℓ,*

$$||\mathbb{P}\{\dim(W^{(t)}) \in \cdot\,\} - \mathbb{P}\{D^{(t)} \in \cdot\,\}||_{\mathrm{TV}} \leq \epsilon(\ell, t)\,, \tag{12}$$

where we recall that $||\mathbb{P}_X - \mathbb{P}_Y||_{\mathrm{TV}} = \sup_A |\mathbb{P}(X \in A) - \mathbb{P}(Y \in A)|$.

Controlling the sequence of random variables $\{D^{(t)}\}_{t\geq 0}$ is quite difficult. Luckily, its behavior simplifies considerably if we let $m \to \infty$ and consider the scaled dimensions $D^{(t)}/(\ell\omega)$.

More precisely, we define the sequence of random variables $\{\xi^{(t)}\}_{t\geq 0}$ with values in $[0, 1]$ recursively as follows. We let $\xi^{(0)} = 1$ identically. For any $t \geq 0$, let n be drawn with distribution ρ_n, and $\xi^{(t)}_1, \dots, \xi^{(t)}_{n-1}$ be iid copies of $\xi^{(t)}$. Further, for $a, b, x \in \mathbb{R}$ with $a \leq b$, define $[x]^b_a = \min(\max(x, a), b)$. Then, the distribution of $\xi^{(t+1)}$ is given by

$$\xi^{(t+1)} \stackrel{\mathrm{d}}{=} \left[\sum_{i=1}^{n-1} \xi^{(t)}_i + 1 - \left(\frac{1-\lambda}{\lambda\omega}\right)\right]^1_0. \tag{13}$$

We will prove that the rescaled dimensions $D^{(t)}/(\ell\omega)$ are accurately tracked by $\xi^{(t)}$.

Lemma 4.2 (Density Evolution versus Rescaled Density Evolution)**.** *For any $n_{\max}$, ω, and λ there exists $\varepsilon > 0$ such that, for any degree distribution P with support in $[0, n_{\max}]$:*

$$\lim_{m\to\infty} \mathbb{P}\{D^{(t+1)} > 0\} \leq n_{\max}\, \mathbb{P}\{\xi^{(t)} \geq \varepsilon\}\,. \tag{14}$$

The previous lemma shows that it suffices to consider the behavior of $\xi^{(t)}$ for which we have the explicit simple recursion (13). Even so, finding a degree distribution ρ which results in codes of large rates and so that $\xi^{(t)}$ converges to 0 for large values of δ, seems challenging. The key to our analysis is the observation that the recursion (13) simplifies significantly if $(1-\lambda)/\lambda$ is an integer multiple of ω. In this case the distribution of $\xi^{(t)}$ trivializes: $\xi^{(t)}$ only takes on the values 0 or 1 regardless of the degree distribution ρ. Density evolution therefore collapses to a scalar recursion, making it possible to find the optimum degree distribution ρ.

Lemma 4.3 (Capacity Achieving Degree Distributions for Rescaled Density Evolution)**.** *Let $\lambda, \omega \in (0, 1)$ be such that $(1-\lambda)/\lambda$ is an integer multiple of ω and let $r < C(\lambda, \omega)/((1-\lambda)(1-\omega))$. Then there exists ρ with bounded support and $1 - 2\int_0^1 \rho(x)\mathrm{d}x \geq r$, and two constants $A > 0$, $\gamma > 1$ such that, for any t, $\varepsilon > 0$,*

$$\mathbb{P}\{\xi^{(t)} \geq \varepsilon\} \leq \exp\{-A\gamma^t\}\,. \tag{15}$$

Proof of the Main Theorem 3.2. Let λ, ω, R be as in the statement of the theorem and $r \in (R/((1-\lambda)(1-\omega)), C(\lambda,\omega)/((1-\lambda)(1-\omega))$. We claim that there exists a degree distribution P with support in $[0, n_{\max}]$, with $1 - 2/P'(1) \geq r$ (equivalently, from the edge perspective, $1 - 2\int_0^1 \rho(x)\,\mathrm{d}x \geq r$) such that the iterative decoder achieves error probability smaller than π in $O(\log\log(1/\pi))$ iterations. Let us check that this indeed proves the theorem. As mentioned above, the perturbation subspace W (i.e., the linear subspace of $(\mathbb{F}_q)^m$ spanned by the rows of $\underline{z}$) can be inferred with high probability by the last $\omega'\ell$ of the output $\underline{y}$. This requires Gaussian elimination of an $m \times (\ell\omega')$ matrix with elements in $\mathbb{F}_q$, which can be accomplished at a cost of $O(N^3)$ operations.

The rest of the codeword $\underline{x}$ is decoded by message passing. Each iteration requires updating $O(N)$ messages (because P has bounded support). Each update, cf. Eq. (8), requires finding a basis for a space spanned by, at most $(\ell\omega)n_{\max}$ vectors in $(\mathbb{F}_q)^m$. This can be done, again via Gaussian elimination, in $O(N^3)$ operations. We thus get $O(N^4)$ operations per iteration. Since running $O(\log\log(1/\pi))$ iterations achieves error probability smaller than π, this implies the thesis.

Let us now prove this claim. First, we fix the degree distribution in such a way that Lemma 4.3 holds for some $A > 0$, $\gamma > 1$. We let $t_*(\pi) = O(\log\log(1/\pi))$ be such that $\mathbb{P}\{\xi^{(t)} \geq \varepsilon\} \leq \exp\{-A\gamma^t\} \leq \pi/(3n_{\max})$ for any $t \geq t_*(\pi)$. Then, for any fixed $t \geq t_*(\pi)$ the decoded error probability is upper bounded by π if N is large enough.

Indeed, ε can be chosen in such a way that Lemma 4.2 holds and therefore, for m large enough, $\mathbb{P}\{D^{(t+1)} > 0\} \leq n_{\max}\pi/(3n_{\max}) + \pi/3 \leq 2\pi/3$. The i-th row of codeword $\underline{x}$ is decoded correctly if any of the two messages $W_{i\to a}^{(t+1)}$ or $W_{i\to \bar{a}}^{(t+1)}$ has dimension 0. Therefore, the symbol error probability is upper bounded by $\mathbb{P}\{\dim(W^{(t+1)}) > 0\}$. By Lemma 4.1, for ℓ large enough, this is at most $\mathbb{P}\{D^{(t+1)} > 0\} + \pi/3 \leq \pi$, which proves the theorem. □

4.2 Proofs of Lemmas

Proof of Lemma 4.1. The proof is based on the 'density evolution' technique [13], and on some remarks that allow to simplify the resulting distributional recursion. A similar result appeared already in the context of erasure decoding for non-binary codes [11]: in order to be self-contained we nevertheless sketch the proof here.

Let $\vec{e}$ be a uniformly random directed edge in G and let $W^{(t)}$ the associated message after t iterations of the message-passing algorithm. Denote by $\mathsf{B}(\vec{e}, t)$ the 'directed neighborhood' of $\vec{e}$ with radius t, i.e., the induced sub-graph containing all non-reversing walks in G of length at most t that terminate in $\vec{e}$. We regard this as a labeled graph with variable node labels given by the received vectors and edge labels by the $m \to m$ matrices that define the code. It is well known that such a neighborhood converges to a (labeled) Galton-Watson tree $\mathsf{T}(t)$.

More precisely, $\mathsf{T}(t)$ is a t-generations tree rooted in a directed edge $\vec{e}_{\mathsf{T}}$ and with offspring distribution ρ_n. We have

$$||\mathbb{P}\{\mathsf{B}(\vec{e}, t) \in \cdot\} - \mathbb{P}\{\mathsf{T}(t) \in \cdot\}||_{\mathrm{TV}} \leq \epsilon(\ell, t)\,, \tag{16}$$

for some $\epsilon(\ell, t)$ as in the statement of Lemma 4.1.

Note that the message $W^{(t)}$ is a function only of the neighborhood $\mathsf{B}(\vec{e}, t)$. Suppose that we apply the message-passing algorithm to $\mathsf{T}(t)$ and let $W_{\mathsf{T}}^{(t)}$ be the message passed through the root edge after t iterations. It follows from the definition of total variation distance that

$$||\mathbb{P}\{\dim(W^{(t)}) \in \cdot\} - \mathbb{P}\{\dim(W_{\mathsf{T}}^{(t)}) \in \cdot\}||_{\mathrm{TV}} \leq \epsilon(\ell, t)\,. \tag{17}$$

The proof is completed by showing that $\dim(W_{\mathsf{T}}^{(t)})$ is distributed as the random variable $D^{(t)}$ defined recursively by Eq. (11). First, note that $W_{\mathsf{T}}^{(t)}$ is a uniformly random subspace, conditional on its dimension $\dim(W_{\mathsf{T}}^{(t)})$. This follows from the message-passing update rule (8) together with the remark that, given any fixed subspace W_* and a uniformly random full-rank $m \times m$ matrix $\mathbb{L}$, $\mathbb{L}W_*$ is a uniformly random subspace with the same dimension as W_*.

We prove that $\dim(W_{\mathsf{T}}^{(t)})$ is distributed as $D^{(t)}$ by recursion. The statement is true for $t = 0$ by definition of our channel model. Consider the tree $\mathsf{T}(t+1)$ and condition on the offspring number at

the root $n-1$. Denote by $W^{(t)}_{\mathsf{T},1},\dots,W^{(t)}_{\mathsf{T},n-1}$ the corresponding messages towards the root and condition on $\dim(W^{(t)}_{\mathsf{T},1}) = d_1,\dots,\dim(W^{(t)}_{\mathsf{T},n-1}) = d_{n-1}$. Then the distribution of $\dim(W^{(t+1)}_{\mathsf{T}})$ is given by the kernel (10) with $D = \ell\omega$ by uniformity of the subspace. The claim follows from the fact that $W^{(t)}_{\mathsf{T},1},\dots,W^{(t)}_{\mathsf{T},n-1}$ are iid because of the tree structure. □

In the proof of Lemma 4.2 we require an estimate of the probability that true density evolution deviates significantly from the the rescaled density evolution.

Proposition 4.4 (Deviations from Asymptotic Density Evolution)**.** *Let V_1 be a subspace of dimension d_1 in $\mathbb{F}_q^m$, and V_2 a uniformly random subspace of dimension d_2. Define $d_1 \odot d_2 \equiv \max(0, d_1+d_2-m)$, and $d_1 \boxplus d_2 \equiv \min(m, d_1+d_2)$. Then*

$$\mathbb{P}\{d_1 \odot d_2 \le \dim(V_1\cap V_2) < d_1\odot d_2 + k\} \ge 1 - q^{-k-\max(0,m-d_1-d_2)}\,, \tag{18}$$

$$\mathbb{P}\{d_1 \boxplus d_2 - k \le \dim(V_1+V_2) < d_1\boxplus d_2\} \ge 1 - q^{-k-\max(0,m-d_1-d_2)}\,. \tag{19}$$

Further, let V be a subspace of dimension d and let $V_1,\dots,V_{n-1}$ be uniformly random subspaces of dimensions (respectively) $d_!,\dots,d_{n-1}$ and $d \equiv [d_1+\cdots+d_{n-1}+d-m]_0^d$. Then

$$\mathbb{P}\{|\dim((V_1+\cdots+V_{n-1})\cap V) - d| \ge k\} \le n\, q^{-k/n}\,. \tag{20}$$

Proof. Notice that Eq. (19) follows from Eq. (18) together with the identity $\dim(V_1+V_2) = d_1+d_2-\dim(V_1\cap V_2)$. Further $\dim(V_1\cap V_2) \ge d_1\odot d_2$ for any two subspaces V_1, V_2 of the given dimensions.

We are left with the task of bounding the probability of $\dim(V_1\cap V_2) \ge d_1\odot d_2 + k$. Notice that this event is identical to $|V_1\cap V_2| \ge q^{d_1\odot d_2+k}$ (we denote by $|S|$ the cardinality of the set S). By the Markov inequality we have

$$\mathbb{P}\{\dim(V_1\cap V_2) \ge d_1\odot d_2 + k\} \le q^{-k-d_1\dot{d}_2}\mathbb{E}|V_1\cap V_2| = q^{-k-d_1\odot d_2}\, q^{d_1+d_2-m}, \tag{21}$$

where the equality on the right-hand side follows by multiplying the number of vectors in V_1 (that is q^{d_1}) with the probability that one of them belongs to V_2 (by uniformity this is q^{-m+d_2}).

Eq. (20) follows by applying the previous bound recursively. bounds. Explicitly, we define $W_1 = V_1$, $W_2 = W_1+V_2,\dots,W_{n-1} = W_{n-2}+V_{n-1}$, and $W_n = W_{n-1}\cap V$. The corresponding (typical) dimensions are $c_1 = d_1$, $c_2 = c_1\boxplus d_2,\dots,c_{n-1} = c_{n-2}\boxplus d_{n-1}$, $c_n = c_{n-1}\odot d_n = d$. By the union bound, with probability at least $1-n\,q^{-k/n}$ we have $|\dim(W_n) - (d_n\odot\dim(W_{n-1}))| \le k/n$ and $|\dim(W_i) - (d_i\boxplus\dim(W_{i-1})| \le k/n$ for $i\in\{2,\dots,n-1\}$. The thesis follows by the triangle inequality. □

Proof of Lemma 4.2. We will first prove that there exists a coupling between $D^{(t)}$ and $\xi^{(t)}$ such that $|D^{(t)} - (\ell\omega)\xi^{(t)}| \le \ell\varepsilon$ with high probability as ℓ, $m\to\infty$ (with λ, ω fixed). Subsequently, we shall prove that this claim implies the thesis.

The coupling is constructed recursively. For $t=0$ we have $D^{(0)} = (\ell\omega)\xi^{(t)} = \ell\omega$ deterministically. This defines the coupling of $D^{(t)}$ and $\xi^{(t)}$ for $t=0$. Assume we have shown how to construct a coupling of $D^{(t)}$ and $\xi^{(t)}$ for some $t\in\mathbb{N}$. To define the coupling for $t+1$ we draw an integer n with distribution ρ_n. We then generate $n-1$ coupled pairs $(D_i^{(t-1)},\xi_i^{(t-1)})$. From those we generate a coupled pair $(D_i^{(t)},\xi_i^{(t)})$ via the recursions (11) and (13), respectively.

In order to prove the claim it is sufficient to show the following. If $V_1,\dots,V_{n-1}$ are uniformly random subspaces of dimensions $(\ell\omega)\xi_1,\dots,(\ell\omega)\xi_{n-1}$ in $\mathbb{F}_q^m$, and if V has dimension $(\ell\omega)$, then, with high probability, $|\dim((V_1+\cdots+V_{n-1})\cap V) - (\ell\omega)\xi| \le \ell\varepsilon$ for any $\varepsilon>0$. This in turns follows from Proposition 4.4 (Eq. (20)) together with the observation that the degree n is bounded.

Let us now consider the thesis of the lemma, Eq. (14). We can assume without loss of generality that $n_{\max}\ge 1$ and $m > \ell\omega$, whence $1-\lambda > \lambda\omega$ follows. Let $n_{\max}\ge 2$ be the largest integer in the support of ρ_n and take $\varepsilon>0$ small enough so that $2(n_{\max}-1)\varepsilon \le (1-\lambda)/(\lambda\omega) - 1 - \gamma$ for some $\gamma>0$. Draw $n_{\max}$ iid copies of $D^{(t)}$, denoted $D_1^{(t)},\dots,D_{n_{\max}}^{(t)}$. Since under the coupling $|D^{(t)} - (\ell\omega)\xi^{(t)}| \le \ell\varepsilon$ with high probability,

$$\mathbb{P}\left\{\max\{D_1^{(t)},\dots,D_{n_{\max}}^{(t)}\} \ge 2\,\varepsilon\,(\ell\omega)\right\} \le n_{\max}\mathbb{P}\{\xi^{(t)}\ge\varepsilon\} + o_m(1)\,. \tag{22}$$

Now draw n with distribution ρ_n and $D^{(t+1)}$ conditional on $D_1^{(t)}, \dots, D_{n-1}^{(t)}$ according to the kernel (10). Namely, $D^{(t+1)}$ is the dimension of $V \cap (V_1 + \cdots + V_{n-1})$ when $\dim(V) = \ell\omega$ and $V_1, \dots, V_{n-1}$ are uniformly random subspaces of $\mathbb{F}_q^m$ with dimensions $D_1^{(t)}, \dots, D_{n-1}^{(t)}$.

Let $W \equiv V_1 + \cdots + V_{n-1}$. Then W is uniformly random conditioned on its dimension $\dim(W) \le D_1^{(t)} + \cdots + D_{n-1}^{(t)} \le \ell(1-\lambda-\lambda\omega)/\lambda - \ell\omega\gamma$ with probability lower bounded as in Eq. (22). Assume this to be the case. By Proposition 4.4, Eq. (18), and recalling that $m = \ell(1-\lambda)/\lambda$, the probability that $D^{(t+1)} = \dim(V \cap W) > 0$ is at most $q^{-\ell\gamma\omega}$. This proves the thesis. □

In order to prove our last auxiliary result, Lemma 4.3, we need some algebraic properties of the edge-perspective capacity-achieving degree distribution and of the corresponding generating function:

$$\rho_k^*(x) = \sum_{i=k+1}^{\infty} \frac{k-1}{(i-1)(i-2)}\, x^{i-1} \equiv \sum_{i=0}^{\infty} \rho_{k,i}^*\, x^{i-1}\,. \tag{23}$$

Lemma 4.5 (Basic Properties of Capacity-Achieving Degree Distribution)**.** *Let $k \in \mathbb{N}$ and define $f_{k,i}(\alpha) = \sum_{j=k}^{i-1} \binom{i-1}{j} \alpha^j (1-\alpha)^{i-1-j}$. Then $\rho_k^*(1) = 1$, $\mathrm{d}\rho_k^*(x)/\mathrm{d}x\,|_{\,x=1} \ge k$, $\int_0^1 \rho_k^*(x)\mathrm{d}x = 1/(2k)$, and $\sum_i \rho_{k,i}^* f_{k,i}(\alpha) = \alpha$.*

Proof. By a reordering of the terms in the sum,

$$\rho_k^*(1) = \lim_{j\to\infty} \sum_{i=k+1}^{k+j} \frac{k-1}{(i-1)(i-2)} = \lim_{j\to\infty} \sum_{i=k+1}^{k+j} \left(\frac{k-1}{i-2} - \frac{k-1}{i-1}\right) = \lim_{j\to\infty} \left(1 - \frac{k-1}{k+j-1}\right) = 1\,.$$

In a similar manner, we have

$$\int_0^1 \rho_k^*(x)\mathrm{d}x = \lim_{j\to\infty} \sum_{i\ge k+1}^{k+j} \frac{k-1}{i(i-1)(i-2)} = \lim_{j\to\infty} \sum_{i\ge k+1}^{k+j} \left(\frac{k-1}{2(i-1)(i-2)} - \frac{k-1}{2i(i-1)}\right) = \frac{1}{2k}\,.$$

The claim $\int_0^1 \rho_k^*(x)\mathrm{d}x = 1/(2k)$ follows since $\rho_k^*(x) = 1$, $\rho_{k,i}^* \ge 0$, and since $\rho(x)$ only contains powers of x of at least k. In order to prove the last assertion we recall the identity [15]

$$\sum_{n=i}^{\infty} \binom{n}{i} x^n = \frac{x^i}{(1-x)^{i+1}}\,. \tag{24}$$

We then obtain (here $\bar\alpha \equiv (1-\alpha)$):

$$\begin{aligned}\sum_i \rho_{k,i}^* f_{k,i}(\alpha) &= \sum_{i\ge k+1} \frac{k-1}{(i-1)(i-2)} \sum_{j=k}^{i-1} \binom{i-1}{j} \alpha^j \bar\alpha^{i-1-j} = (k-1)\sum_{j=k}^{\infty} \left(\frac{\alpha}{\bar\alpha}\right)^j \sum_{i=j+1}^{\infty} \frac{\binom{i-1}{j}\bar\alpha^{i-1}}{(i-1)(i-2)} \\ &= (k-1)\sum_{j\ge k}^{\infty} \left(\frac{\alpha}{\bar\alpha}\right)^j \frac{\alpha^{1-j}\bar\alpha^j}{j(j-1)} = (k-1)\sum_{j=k}^{\infty} \frac{\alpha}{j(j-1)} = \alpha\,,\end{aligned}$$

where we applied the identity obtained by integrating Eq. (24) twice with respect to x. □

Proof of Lemma 4.3. Let $k = (1-\lambda)/(\lambda\omega)$, $k \in \mathbb{N}$. Then $C(\lambda,\omega)/((1-\omega)(1-\lambda)) = 1 - 1/k$.

It is clear from the recursive definition (13) together with the initial condition $\xi^{(0)} = 1$ that, for any $t \ge 0$, $\xi^{(t)}$ only takes values 0 and 1. Let $\alpha_t \equiv \mathbb{P}\{\xi^{(t)} = 1\}$. Then $\alpha_0 = 1$, and Eq. (13) implies that

$$\alpha_{t+1} = \sum_{n=k+1}^{\infty} \rho_n\, f_{k,n}(\alpha_t) \equiv \mathsf{F}_{k,\rho}(\alpha_t), \tag{25}$$

where $f_{k,n}(\alpha)$ is defined as in the statement of Lemma 4.5 (note that $f_{k,k}(\alpha) \equiv 0$). We claim that for any $r < 1 - 1/k$ there exists an edge-perspective degree distribution ρ of bounded support such that: (i)

$1 - 2\int_0^1 \rho(x)\,\mathrm{d}x \geq r$; (ii) $\mathsf{F}_{k,\rho}(\alpha) < \alpha$ for any $\alpha \in (0,1]$; (iii) $\mathsf{F}_{k,\rho}(\alpha) = O(\alpha^k)$ as $\alpha \downarrow 0$. Then the lemma follows by standard calculus, with $\gamma \in (1,k)$ and A sufficiently small.

In order to exhibit such a degree distribution, fix $b \in \mathbb{N}$, $b \geq k$, and define $\rho(x) = \sum_{i=k} \rho_i x^{i-1}$, where $\rho_i = 0$ except for $\rho_i = \rho^*_{k,i}$, $i = k+1, \dots, b$, and $\rho_k = 1 - \sum_{i=k+1}^b \rho^*_{k,i}$. Then

$$\int_0^1 \rho(x)\,\mathrm{d}x = \sum_{i=k}^{b} \rho_i/i = \sum_{i=k}^{b} \rho^*_{k,i}/i + \sum_{i=b+1}^{\infty} \rho^*_{k,i}/k\,. \tag{26}$$

By Lemma 4.5 the right-hand side converges to $1/(2k)$ as $b \to \infty$. Therefore we can chose b large enough so that claim (i) above is fulfilled.

Consider now claim (ii). We write

$$\mathsf{F}_{k,\rho}(\alpha) = \sum_{i=k+1}^{b} \rho_i f_{k,i}(\alpha) = \sum_{i=k+1}^{\infty} \rho^*_{k,i} f_{k,i}(\alpha) - \sum_{i=b+1}^{\infty} \rho^*_{k,i} f_{k,i}(\alpha) = \alpha - \sum_{i=b+1}^{\infty} \rho^*_{k,i} f_{k,i}(\alpha),$$

where the last identity follows from Lemma 4.5. The claim is implied by the remark that $f_{k,i}(\alpha) > 0$ for $i \geq k+1$ and $\alpha \in (0,1]$.

Finally, claim (iii) is a consequence of the fact that $f_{k,i}(\alpha) = \binom{i-1}{k}\alpha^k + O(\alpha^{k+1})$ together with $i \leq b$. □

4.3 Capacity

Proof of Proposition 3.1. By standard information-theoretic arguments [3], the channel information capacity is given by

$$C(\omega,\lambda) = \lim_{N\to\infty,\, \ell = N\lambda} \frac{1}{N\ell} \sup_{\mathbb{P}_{\underline{X}}} I(\underline{X};\underline{Y})\,. \tag{27}$$

Here $I(\underline{X};\underline{Y}) = \sum_{\underline{x},\underline{y}} \mathbb{P}_{\underline{X},\underline{Y}}(\underline{x},\underline{y}) \log\{\mathbb{P}_{\underline{X},\underline{Y}}(\underline{x},\underline{y})/\mathbb{P}_{\underline{X}}(\underline{x})\mathbb{P}_{\underline{Y}}(\underline{y})\}$ is the *mutual information* between $\underline{X}$ and $\underline{Y}$ and the supremum is taken over all possible input distributions.

Writing the mutual information in terms of entropy and conditional entropy, and using our channel model (2), we have $I(\underline{X};\underline{Y}) = H(\underline{Y}) - H(\underline{Y}|\underline{X}) = H(\underline{Y}) - H(\underline{Z})$. Since $H(\underline{Z})$ does not depend on the input distribution, the mutual information is maximized when the latter is uniform. This implies that the output is uniform as well, and we get $H(\underline{Y}) = \log(q^{m\ell})$.

Finally, $H(\underline{Z})$ is the logarithm of the number $A(s,\ell,m)$ of $\ell \times m$ matrices of rank $\mathrm{rank}(\underline{Z}) = \ell\omega \equiv s$. We have $A(s,\ell,m) = q^{m\ell}\mathbb{P}_0\{\mathrm{rank}(\underline{Z}) = s\}$ where $\mathbb{P}_0$ denotes probability with respect to a uniformly random matrix $\underline{Z}$. Assume without loss of generality that $\ell, m \geq s$. If $z_1, \dots, z_\ell$ be the lines of $\underline{Z}$, then the first s lines are independent with probability $(1-q^{-\ell})(1-q^{-\ell+1})\cdots(1-q^{-\ell+s}) \geq 1 - sq^{-\ell+s}$. then the space .

$$A(s,\ell,m) \geq q^{m\ell}\,\mathbb{P}\{z_{s+1}\dots z_\ell \in \quad (z_1\dots z_s),\, \mathrm{rank}(z_1\dots z_s) = s\} \geq q^{m\ell}\, q^{-(\ell-s)(m-s)}(1 - sq^{-\ell+s})\,. \tag{28}$$

On the other hand $\mathbb{P}_0\{\mathrm{rank}(\underline{Z}) = s\}$ is upper bounded by summing over all subsets of s lines (there are $\binom{\ell}{s} \leq 2^\ell$ such subsets), the probability that such lines are independent and that the other lines are in the span generated by these. Such an upper bound is at most 2^ℓ larger than the above lower bound. By taking $N \to \infty$ with $\ell = N\lambda = N - m$, $\lambda \in (0,1)$ and $\omega \in (0, \min(1, (1-\lambda)/\lambda$ we get

$$H(\underline{Z}) = \log A(s,\ell,m) = N\ell\,(\omega + \omega^2\lambda) + O(N)\,. \tag{29}$$

Therefore $I(\underline{X};\underline{Y}) = H(\underline{Y}) - H(\underline{Z}) = N\ell(1 - \lambda - \omega + \omega^2\lambda) + O(N)$ whence the thesis follows. □

References

[1] R. Ahlswede, N. Cai, S.-Y. R. Li, and R. W. Yeung, *Network information flow*, IEEE Trans. Inform. Theory, 46 (2000), pp. 1204–1216.

[2] P. A. Chou, Y. Wu, and K. Jain, *Practial network coding*, in Proc. of the Allerton Conf. on Commun., Control, and Computing, Monticello, IL, USA, 2003.

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

[4] C. Fragouli and E. Soljanin, *Network Coding Fundamentals*, vol. 2 of Foundations and Trends in Networking, NOW, Delft, Holland, 2007.

[5] R. G. Gallager, *Low-Density Parity-Check Codes*, MIT Press, Cambridge, MA, USA, 1963.

[6] T. Ho, R. Kötter, M. Medard, D. R. Karger, and M. Effros, *The benefits of coding over routing in a randomized setting*, in Proc. of the IEEE Int. Symposium on Inform. Theory, Yokohama, Japan, 2003, p. 442.

[7] R. Kötter and F. R. Kschischang, *Coding for errors and erasures in random network coding*. Submitted, Nov. 2007.

[8] R. Kötter and M. Medard, *An algebraic approach to network coding*. Submitted, Feb. 2004.

[9] S.-Y. R. Li, R. W. Yeung, and N. Cai, *Linear network coding*, IEEE Trans. Inform. Theory, 49 (2003), pp. 371–381.

[10] M. Luby, M. Mitzenmacher, A. Shokrollahi, D. A. Spielman, and V. Stemann, *Practical loss-resilient codes*, in Proc. of the 29th annual ACM Symposium on Theory of Computing, 1997, pp. 150–159.

[11] V. Rathi and R. Urbanke, *Density evolution, threshold and the stability condition for non-binary LDPC codes*, IEE Proc. Commun., 152 (2005), pp. 1069–1074.

[12] T. Richardson and R. Urbanke, *Efficient encoding of low-density parity-check codes*, IEEE Trans. Inform. Theory, 47 (2001), pp. 638–656.

[13] ———, *Modern Coding Theory*, Cambridge University Press, 2007. In preparation.

[14] D. Silva, R. Kötter, and F. R. Kschischang, *A rank-metric approach to error control in random network coding*. Submitted, Nov. 2007.

[15] H. S. Wilf, *Generatingfunctionology*, Academic Press, 2 ed., 1994.

[16] R. W. Yeung, S.-Y. R. Li, N. Cai, and Z. Zhang, *Network Coding Theory: Multiple Sources*, vol. 2 of Foundations and Trends in Communications and Information Theory, NOW, Delft, Holland, 2005.

[17] ———, *Network Coding Theory: Single Sources*, vol. 2 of Foundations and Trends in Communications and Information Theory, NOW, Delft, Holland, 2005.