Design of Capacity-Approaching Irregular Low-Density Parity-Check Codes

Thomas J. Richardson, M. Amin Shokrollahi, *Member, IEEE*, and Rüdiger L. Urbanke

***Abstract*—We design low-density parity-check (LDPC) codes that perform at rates extremely close to the Shannon capacity. The codes are built from highly irregular bipartite graphs with carefully chosen degree patterns on both sides. Our theoretical analysis of the codes is based on [1]. Assuming that the underlying communication channel is symmetric, we prove that the probability densities at the message nodes of the graph possess a certain symmetry. Using this symmetry property we then show that, under the assumption of no cycles, the message densities always converge as the number of iterations tends to infinity. Furthermore, we prove a stability condition which implies an upper bound on the fraction of errors that a belief-propagation decoder can correct when applied to a code induced from a bipartite graph with a given degree distribution.**

Our codes are found by optimizing the degree structure of the underlying graphs. We develop several strategies to perform this optimization. We also present some simulation results for the codes found which show that the performance of the codes is very close to the asymptotic theoretical bounds.

***Index Terms*—Belief propagation, irregular low-density parity-check codes, low-density parity-check codes, turbo codes.**

I. Introduction

In this paper we present *irregular* low-density parity-check (LDPC) codes which exhibit a performance extremely close to the best possible as determined by the Shannon capacity formula. For the binary-input additive white Gaussian noise (BI-AWGN) channel, the best code of rate one-half presented in this paper has a threshold within 0.06 dB from capacity, and simulation results show that our best LDPC code of length one million achieves a bit-error probability of 10^{-6} less than 0.13 dB away from capacity, surpassing the best (turbo) codes known so far.

LDPC codes possess several other distinct advantages over turbo codes. First, (belief-propagation) decoding for LDPC codes is fully parallelizable and can potentially be accomplished at significantly greater speeds. Second, as indicated in an earlier paper [1], very low complexity decoders that closely approximate belief propagation in performance may be (and have been) designed for these codes. Third, LDPC decoding is verifiable in the sense that decoding to a correct codeword is a detectable event. One practical objection to LDPC codes has been that their encoding complexity is high. One way to get around this problem is to slightly modify the construction of codes from bipartite graphs to a cascade of such graphs, see [2], [24], [3]. An alternative solution for practical purposes, which does not require cascades, is presented in [4].

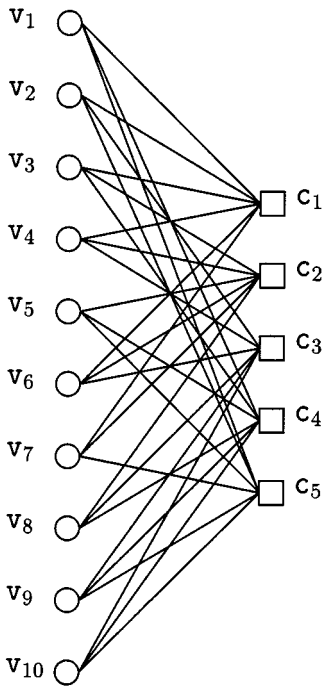

Fig. 1. A $(3, 6)$-regular LDPC code of length 10 and rate one-half. There are 10 variable nodes and five check nodes. For each check node check node c_i the sum (over GF (2)) of all adjacent variable nodes is equal to zero.

Let us recall some basic notation. As originally suggested by Tanner [5], LDPC codes are well represented by bipartite graphs in which one set of nodes, the *variable nodes*, corresponds to elements of the codeword and the other set of nodes, the *check nodes*, corresponds to the set of parity-check constraints which define the code. *Regular* LDPC codes are those for which all nodes of the same type have the same degree. For example, a $(3, 6)$-regular LDPC code has a graphical representation in which all variable nodes have degree 3 and all check nodes have degree 6. The bipartite graph determining such a code is shown in Fig. 1. *Irregular* LDPC codes were introduced in [2], [24] and were further studied in [6]–[8]. For such an irregular LDPC code, the degrees of each set of nodes are chosen according to some distribution. Thus, an irregular LDPC code might have a

Manuscript received May 28, 1999; revised September 1, 2000. The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Sorrento, Italy, June 2000.

T. J. Richardson was with Bell Labs, Lucent Technologies, Murry Hill, NJ 07974 USA. He is now with Flarion Technologies, Bedminster, NJ 07921 USA (e-mail: richardson@flarion.com).

M. A. Shokrollahi was with Bell Labs, Lucent Technologies, Murry Hill, NJ 07974 USA. He is now with Digital Fountain, San Francisco, CA 94110 USA (e-mail: amin@digitalfountain.com)

R. Urbanke was with Bell Labs. He is now with EPFL, LTHC-DSC, CH-1015 Lausanne (e-mail: rudiger.urbanke@epfl.ch).

Communicated by F. R. Kschishang, Associate Editor for Coding Theory.

Publisher Item Identifier S 0018-9448(01)00738-6.

graphical representation in which half the variable nodes have degree 3 and half have degree 5, while half the constraint nodes have degree 6 and half have degree 8.

For a given length and a given degree distribution, we define an *ensemble* of codes by choosing edges, i.e., the connections between variable and check nodes, randomly. More precisely, we enumerate the edges emanating from the variable nodes in some arbitrary order and proceed in the same way with the edges emanating from the check nodes. Assume that the number of edges is E. Then a code (a particular instance of this ensemble) can be identified with a permutation on E letters. By definition, all elements in this ensemble are equiprobable. In practice, the edges are not chosen entirely randomly since certain potentially unfortunate events in the graph construction can be easily avoided.

In [1], an asymptotic analysis of LDPC codes under message-passing decoding was presented. To briefly recall the main results let us assume that we have the following setup.

[Channel] We are given an ordered family of binary-input memoryless channels parameterized by a real parameter δ such that if $\delta_1 < \delta_2$ then the channel with parameter δ_2 is a *physically degraded* version of the channel with parameter δ_1, see [1]. Furthermore, each channel in this family is *output-symmetric*,[1] i.e.,

$$p(y|x=1) = p(-y|x=-1). \tag{1}$$

[Ensemble] We say that a polynomial $\gamma(x)$ of the form

$$\gamma(x) := \sum_{i\geq 2} \gamma_i x^{i-1}$$

is a *degree distribution* if $\gamma(x)$ has nonnegative coefficients and $\gamma(1) = 1$. Note that we associate the coefficient γ_i to x^{i-1} rather than x^i. We will see that this notation, which was introduced in [2], leads to very elegant and compact descriptions of the main results. Given a degree distribution pair (λ, ρ) associate to it a sequence of code *ensembles* $\mathcal{C}^n(\lambda, \rho)$, where n is the length of the code and where

$$\lambda(x) := \sum_{i=2}^{d_{\mathsf{v}}} \lambda_i x^{i-1}$$

$(\rho(x) := \sum_{i=2}^{d_{\mathsf{c}}} \rho_i x^{i-1})$ specifies the variable (check) node degree distribution. More precisely, λ_i (ρ_i) represents the fraction of *edges* emanating from variable (check) nodes of degree i. For example, for the $(3, 6)$-regular code we have $\lambda(x) := x^2$ and $\rho(x) := x^5$. The maximum *variable degree* and *check degree* is denoted by d_{v} and d_{c}, respectively. Assume that the code has n variable nodes. The number of *variable nodes of degree* i is then

$$n \frac{\lambda_i/i}{\sum_{j\geq 2} \lambda_j/j} = n \frac{\lambda_i/i}{\int_0^1 \lambda(x)\,dx}$$

and so E, the total number of edges emanating from all variable nodes, is equal to

$$n \sum_{i\geq 2} \frac{\lambda_i/i}{\int_0^1 \lambda(x)\,dx} = n \frac{1}{\int_0^1 \lambda(x)\,dx}.$$

In the same manner, assuming that the code has m check nodes, E can also be expressed as

$$m \frac{1}{\int_0^1 \rho(x)\,dx}.$$

Equating these two expressions for E, we conclude that

$$m = n \frac{\int_0^1 \rho(x)\,dx}{\int_0^1 \lambda(x)\,dx}.$$

Generically, assuming that all these check equations are linearly independent, we see that the *design rate* is equal to

$$r(\lambda, \rho) := \frac{n-m}{n} = 1 - \frac{\int_0^1 \rho(x)\,dx}{\int_0^1 \lambda(x)\,dx}$$

as was first shown in [2].

[Message-Passing Decoder] Select a *message-passing* decoder. By definition, messages only contain *extrinsic* information, i.e., the message emitted along an edge e does not depend on the incoming message along the same edge. Further, the decoder fulfills the following *symmetry conditions*. Flipping the sign of all incoming messages at a variable node results in a flip of the sign of all outgoing messages. The symmetry condition at a check node is slightly more involved. Let e be an edge emanating from a check node c. Then flipping the sign of i incoming messages arriving at node c, excluding the message along edge e, results in a change of the sign of the outgoing message along edge e by $(-1)^i$. In all these cases, only the sign is changed, the reliability remains unchanged. Finally, we generally require that the decoder be asymptotically monotonic with respect to the channel parameter. Roughly speaking, this means that in the limit of infinitely long codes, the probability of error of the decoder is nondecreasing in the channel parameter given a fixed number of iterations. In the case of the belief-propagation decoder, this property is a direct consequence of the decoder's asymptotic optimality and the fact that we consider families of channels which can be ordered by physical degradation (see Section III). For many other decoders of interest the monotonicity property can be proved directly, and it seems to hold for virtually all decoders of interest. In this paper, we are interested almost exclusively in the belief-propagation decoder, so we shall implicitly assume monotonicity throughout.

Under the above assumptions, with (λ, ρ) and the channel family fixed, there exists a threshold δ^*, i.e., a maximum channel parameter, with the following properties. For any $\epsilon > 0$ and $\delta < \delta^*$, there exists a length $n(\epsilon, \delta)$ and a number $\ell(\epsilon, \delta)$ such that almost every[2] code in $\mathcal{C}^n(\lambda, \rho)$, $n > n(\epsilon, \delta)$, has bit error probability smaller than ϵ assuming that transmission takes place over the channel with parameter δ and that $\ell(\epsilon, \delta)$ iterations of message-passing decoding are performed. Conversely, for any fixed upper bound on the number of iterations, if the transmission takes place over a channel with parameter $\delta > \delta^*$, then almost every[2] code in $\mathcal{C}^n(\lambda, \rho)$ has bit-error

[1]It is reassuring to note that *linear* binary codes are known to be capable of achieving capacity on binary-input memoryless output-symmetric channels, see [9].

[2]More precisely, the fraction of codes for which the above statement is true converges exponentially fast (in n) to one.

probability larger than some constant $\gamma = \gamma(\delta)$, where γ does *not* depend on the number of iterations performed.[3]

The main steps taken to arrive at these conclusions are the following. The first shows that if one fixes the number of iterations, then the performance of the various realizations of the graph and channel concentrate around their expected value, where this concentration is exponential in the length of the code. (The exponent may in general depend on the degree distribution pair (λ, ρ), the chosen message-passing decoder, and the channel parameter.)[4] Therefore, in order to characterize the performance of (almost all) sufficiently long codes it suffices to determine their average performance. Unfortunately, it does not seem to be an easy task to determine this average performance for finite-length ensembles. In the limit of very long codes, however, the average performance can be determined as follows. One first observes that with probability $1-O(\frac{1}{n})$ the decoding neighborhood of a given variable node is *tree-like*, i.e., it does not contain any repetitions/cycles. When the decoding neighborhood is a tree, the performance of the decoder is fairly straightforward to determine because all involved random variables are *independent*. Moreover, under the above mentioned symmetry assumptions of the channel and the decoder, one can assume that the all-one codeword was transmitted, i.e., the conditional bit-error probability is independent of the transmitted codeword. By convention, we choose the messages in such a way that, under the all-one codeword assumption, *positive* messages signify *correct* messages whereas *negative* messages indicate errors. This is, e.g., the case for belief-propagation decoders where the messages are log-likelihood ratios of the form $\log \frac{p(y|x=1)}{p(y|x=-1)}$. The distribution of the messages initially emitted is determined by the channel and it has an associated probability of error. Under the above *independence assumption*, we now track the evolution of the message distributions as they progress up the tree, i.e., toward the root. In particular, one is interested in the evolution of the error probability as a function of the iteration number. The threshold is then defined as the "worst" channel parameter such that the message distribution evolves in such a way that its associated probability of error converges to zero as the number of iterations tends to infinity. This procedure of tracking the evolution of the message distribution is termed *density evolution*.

In [1], an efficient numerical procedure was developed to implement density evolution for the important case of belief-propagation decoders and to therewith efficiently compute the associated threshold to any desired degree of accuracy. Also in [1], threshold values and simulation results were given for a variety of noisy channel models but the class of LDPC codes considered was largely restricted to *regular* codes. In this paper, we present results indicating the remarkable performance that can be achieved by properly chosen *irregular* codes.

The idea underlying this paper is quite straightforward. Assume we are interested in transmission over a particular family of binary-input memoryless output-symmetric channels using a particular message-passing decoder. Since any given pair (λ, ρ) of degree distributions has an associated threshold, call it $\delta^*(\lambda, \rho)$, it is natural to search for those pairs that maximize this threshold.[5] This was accomplished, with great success, in the case of the erasure channel [2], [24], [10], [11]. For most other memoryless channels of interest the situation is much more complicated and new methods must be brought to bear on the optimization problem. Fig. 2 compares the performance of an instance of the $(3, 6)$-regular LDPC ensemble (which is the best regular ensemble) with the performance of an instance of the best irregular LDPC ensemble we found in our search and with the performance of an instance of the standard parallel concatenated ensemble introduced by Berrou, Glavieux, and Thitimajshima [12]. All three codes have rate one-half and their performance under belief-propagation decoding over the BIAWGNC is shown for a code word length of 10^6. Also shown is the Shannon limit and the threshold value of our best LDPC ensemble $(\sigma^* = 0.9718)$. It is evident from this figure that considerable benefit can be derived from optimizing degree distribution pairs. For $n = 10^6$ and a bit-error probability of 10^{-6}, our best LDPC code is only 0.13 dB away from capacity. This even surpasses the performance of turbo codes. Even more impressive, the threshold, which indicates the performance for infinite lengths, is a mere 0.06 dB away from the Shannon capacity limit.

The empirical evidence presented in Fig. 2 together with the results presented in Section II beg the question of whether LDPC codes under *belief-propagation* decoding can achieve capacity over a given binary-input memoryless output-symmetric channel.[6] The only definitive results in this direction are those of [2], [24], [15], which give explicit sequences of degree distribution pairs whose thresholds over the binary erasure channel (BEC) converge to the Shannon capacity limit. The following theorem, due to Gallager, imposes, at least for the binary symmetric channel (BSC), a necessary condition on LDPC codes that would achieve capacity: their maximum check degree d_c must tend to infinity.[7] Although this result bounds the performance of LDPC codes away from capacity, the gap is extremely small and the gap converges to zero exponentially fast in d_c. Hence, although of great theoretical interest, the theorem does not impose a significant practical limitation.[8]

Theorem 1 [13, p. 37]: Let $C \in \mathcal{C}^n(\lambda, \rho)$ be an LDPC code of rate r. Let C be used over a BSC with crossover probability δ and assume that each codeword is used with equal probability. If $r > 1 - h(\delta)/h(p^*)$, where $h(\cdot)$ is the binary entropy function and

$$p^* = \frac{1+(1-2\delta)^{d_c}}{2}$$

[3]We conjecture that actually the following much stronger statement is true—namely, that *all* codes in a given LDPC ensemble have bit-error probability of at least γ *regardless of their length* and *regardless of how many iterations* are performed.

[4]In our proofs, however, the obtained exponent depends only on the degree distribution pair (λ, ρ).

[5]We may also optimize degree distribution pairs under various constraints. For example, the larger the degrees used, the larger the code size needs to be in order to approach the predicted asymptote. Therefore, it is highly desirable to look for the best degree distribution pair with some *a priori* bound on the size of the degrees.

[6]In the case of *maximum-likelihood* decoding this was answered in the affirmative by Gallager and McKay, see [13], [14].

[7]We conjecture that a similar statement (and proof) can be given for continuous channels.

[8]In fact, a similar theorem holds also for the erasure channel [15, Theorem 1], and yet, there are capacity-achieving sequences of degree distributions for that channel.

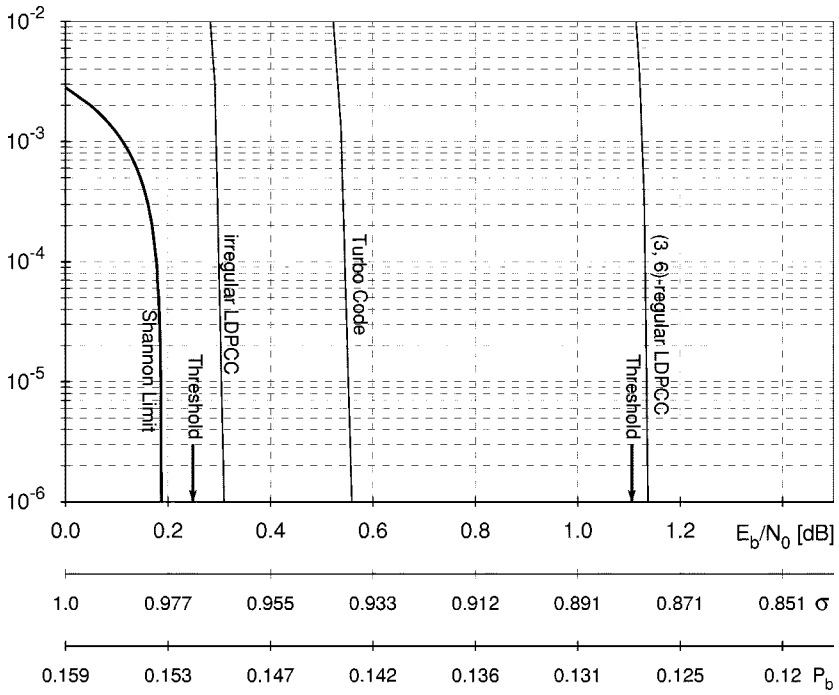

Fig. 2. Comparison of the error rates achieved by a $(3, 6)$-regular LDPC code, turbo code, the and optimized irregular LDPC code. All codes are of length 10^6 and of rate one-half. The bit-error rate for the BIAWGNC is shown as a function of E_b/N_0 (in decibels), the standard deviation σ, as well as the raw input bit-error probability P_b.

then the (block or bit)-error probability is bounded away from zero by a constant which is independent of n.

Discussion: Note that the capacity of the BSC is equal to $1 - h(\delta)$. Since $p^* > 1/2$ for any finite d_c, we have $h(p^*) < 1$ and, therefore, $1 - h(\delta)/h(p^*) < 1 - h(\delta)$. A quick calculation shows that the *gap* to capacity is well approximated by $\frac{h(\delta)}{2 \ln 2}(1 - 2\delta)^{2d_c}$ which decreases to zero exponentially fast in d_c. Gallager stated the above theorem for (d_v, d_c)-regular codes. An examination of the proof reveals, however, that it remains valid for the case of irregular codes with d_c interpreted as the maximum check node degree. The key to the proof lies in the realization that the entropy of the received word can be bounded as follows: given a received word, describe it by specifying the value of rn systematic bits (which is equivalent to specifying a codeword) plus the value of the $(1 - r)n$ parities (which will specify the coset). Since a parity is one with probability $1 - p^*$, which is strictly less than one-half, the entropy of the received word is strictly less than n bits, which gives rise to the stated upper bound. Details of the proof can be found in [13, p. 39].[9] $\diamond$

The outline of this paper is as follows. We start by presenting tables of some very good degree distribution pairs in Section II. Although we focus mostly on the BIAWGNC and rate one-half, we also give a few examples for different channels and rates. We discuss some simulation results that show that the promised performance can be closely achieved for reasonably long codes. In Section III, we describe and study properties of density evolution. Under the assumption that the input distribution arises from a symmetric channel, we show that the message distributions satisfy a certain *symmetry* condition which is invariant under density evolution. Many simplifications arise in this symmetric channel case and they afford us considerable insight into the nature of density evolution. In particular, we will derive a *stability condition* which gives rise to an upper bound on the threshold for the case of general binary-input memoryless output-symmetric channels. We also show that the threshold can, at least partially, be characterized by the (non)existence of *fixed points* of density evolution. Finally, in Section IV, we describe the numerical optimization techniques which were used to generate the tables in Section II. Throughout the paper we will motivate many definitions and statements for general binary-input memoryless output-symmetric channels with their counterpart for the BEC channel. We will see that despite the simplicity of the BEC model many of its iterative decoding properties are shared by the general class of channels considered in this paper.

II. Capacity-Approaching Degree Distribution Paris

A. Optimization Results

Using numerical optimization techniques described in some detail in Section IV, we searched for good degree distribution pairs of rate one-half with various upper bounds on the maximum variable degree d_v. The result of our search for the BIAWGNC is summarized in Tables I and II. Table I contains those degree distribution pairs with $d_v = 4, \ldots, 12$, whereas Table II contains degree distribution pairs with $d_v = 15, 20, 30$, and 50. In each table, columns correspond to one particular degree distribution pair. For each degree distribution pair, the coefficients of λ and ρ are given as well as the threshold σ^*,[10] the

[9] We note that a slightly sharper bound can be given if we replace d_c with the average degree of the $(1 - r)$-fraction of highest degree nodes. However, since the improvement is usually only slight and since the exact size of the gap is not significant in practice, we leave the details to the interested reader.

[10] We assume standard pulse-amplitude modulation (PAM) signaling $y = x + z$ with $x = \pm 1$ and $z \propto N(0, \sigma^2)$.

TABLE I
GOOD DEGREE DISTRIBUTION PAIRS OF RATE ONE-HALF FOR THE BIAWGNC WITH MAXIMUM VARIABLE NODE DEGREES $d_v = 4, 5, 6, 7, 8, 9, 10, 11,$ and 12. FOR EACH DEGREE DISTRIBUTION PAIR THE THRESHOLD VALUE σ^*, THE CORRESPONDING $(\frac{E_b}{N_0})^*$ IN DECIBELS, AND $p^* = Q(1/\sigma^*)$ (i.e., THE INPUT BIT-ERROR PROBABILITY OF A HARD-DECISION DECODER) ARE GIVEN. ALSO LISTED IS λ_2, THE MAXIMUM STABLE VALUE OF λ_2 FOR THE GIVEN $\rho'(1)$ AND FOR $\sigma = \sigma^*$

d_v	4	5	6	7	8	9	10	11	12
λ_2^*	0.38364	0.34648	0.34043	0.31571	0.30166	0.28321	0.27165	0.26269	0.25522
λ_2	0.38354	0.32660	0.33241	0.31570	0.30013	0.27684	0.25105	0.23882	0.24426
λ_3	0.04237	0.11960	0.24632	0.41672	0.28395	0.28342	0.30938	0.29515	0.25907
λ_4	0.57409	0.18393	0.11014				0.00104	0.03261	0.01054
λ_5		0.36988							0.05510
λ_6			0.31112						
λ_7				0.43810					
λ_8					0.41592				0.01455
λ_9						0.43974			
λ_{10}							0.43853		0.01275
λ_{11}								0.43342	
λ_{12}									0.40373
ρ_5	0.24123								
ρ_6	0.75877	0.78555	0.76611	0.43810	0.22919	0.01568			
ρ_7		0.21445	0.23389	0.56190	0.77081	0.85244	0.63676	0.43011	0.25475
ρ_8						0.13188	0.36324	0.56989	0.73438
ρ_9									0.01087
σ^*	0.9114	0.9194	0.9304	0.9424	0.9497	0.9540	0.9558	0.9572	0.9580
$(\frac{E_b}{N_0})^*_{dB}$	0.8085	0.7299	0.6266	0.5153	0.4483	0.4090	0.3927	0.3799	0.3727
p^*	0.1369	0.1384	0.1412	0.1443	0.1462	0.1473	0.1477	0.1481	0.1483

corresponding value of E_b/N_0 in decibels, and finally the raw bit-error probability p^* of the input if it were quantized to 1 bit, i.e., $p^* = Q(1/\sigma^*)$, where $Q(x)$ is the well-known Q-function. In Section III-E, we will show that, given the channel and $\rho'(1)$, there is a *maximum stable* value of λ_2, call it λ_2^*. More precisely, we will show that for any degree distribution pair with λ_2 strictly larger than λ_2^* the probability of bit error cannot converge to zero, regardless of how many iterations are performed. This value λ_2^* is also listed in the tables. As required, we can see that for every listed degree distribution $\lambda_2 < \lambda_2^*$, and the two values are fairly close.

The results are quite encouraging. Compared to regular LDPC codes for which the highest achievable threshold for the BIAWGNC is $\sigma^* = 0.88$, irregular LDPC codes have substantially higher thresholds. The threshold increases initially rapidly with d_v and for the largest investigated degree $d_v = 50$, the threshold value is only 0.06 dB away from capacity!

It is quite tempting to conjecture that the threshold will converge to the ultimate (Shannon) limit (which, up to the precision given, is equal to $\sigma_{\text{opt}} = 0.9787$ for rate one-half codes) as d_v tends to infinity. Unfortunately, as of this moment, we only have this empirical evidence to support this conjecture.

Although in this paper we focus mainly on the BIAWGNC and binary codes of rate one-half, the following examples show that the same techniques can be used to find very good degree distribution pairs for other memoryless channels and different rates. We note that, for a particular rate, degree distribution pairs that were optimized for the BIAWGNC are usually very good degree distribution pairs for a large class of channels, including the binary-input Laplace channel (BILC) and the BSC. Nevertheless, optimizing a degree distribution pair for a particular channel will generally give even better results.

Example 1 [BIAWGNC; $r = \frac{8}{9}$]: In this example, we consider codes of rate $r = \frac{8}{9}$ for the BIAWGNC channel. For this rate, the Shannon bound for σ is $\sigma_{\text{opt}} = 0.528936$. We found the following degree distribution pair which has a theoretical threshold of $\sigma^* = 0.5183$:

$$\lambda(x) := 0.1575x + 0.3429x^2 + 0.0363x^5 + 0.0590x^6 + 0.2790x^8 + 0.1253x^9$$

and

$$\rho(x) := 0.8266x^{34} + 0.1345x^{35} + 0.0087x^{70} + 0.0302x^{71}.$$

Allowing higher degrees would almost certainly result in degree distribution pairs with larger thresholds. •

Example 2 [BSC; $r = \frac{1}{2}$]: The ultimate threshold, i.e., the Shannon limit, for the BSC and rate one-half is $\delta_{\text{opt}} = 0.110028$. The best degree distribution pair we have found so far has $\delta^* = 0.106$ and is given by

$$\begin{aligned}\lambda(x) := {} & 0.157581x + 0.164953x^2 + 0.0224291x^3 \\ & + 0.045541x^4 + 0.0114545x^5 + 0.0999096x^6 \\ & + 0.0160667x^7 + 0.00258277x^8 + 0.00454797x^9 \\ & + 0.000928767x^{10} + 0.0188361x^{11} + 0.0648277x^{12} \\ & + 0.0206867x^{13} + 0.000780516x^{14} + 0.0383603x^{15} \\ & + 0.0419398x^{16} + 0.0023117x^{19} + 0.00184157x^{20} \\ & + 0.0114194x^{22} + 0.0116636x^{28} + 0.0850183x^{39} \\ & + 0.01048x^{40} + 0.0169308x^{55} + 0.0255644x^{56} \\ & + 0.0364086x^{70} + 0.0869359x^{74}\end{aligned}$$

TABLE II
GOOD DEGREE DISTRIBUTION PAIRS OF RATE ONE-HALF FOR THE BIAWGNC WITH MAXIMUM VARIABLE NODE DEGREES $d_v = 15, 20, 30$, AND 50. FOR EACH DEGREE DISRRIBUTION PAIR THE THRESHOLD VALUE σ^*, THE CORRESPONDING $(\frac{E_b}{N_0})^*$ IN dB AND $p^* = Q(1/\sigma^*)$ (i.e., THE INPUT BIT-ERROR PROBABILITY OF A HARD-DECISION DECODER) ARE GIVEN. ALSO LISTED IS λ_2^*, THE MAXIMUM STABLE VALUE OF λ_2 FOR THE GIVEN $\rho'(1)$ AND FOR $\sigma = \sigma^*$

d_v	15	20	30	50
λ_2^*	0.24446	0.23261	0.21306	0.18379
λ_2	0.23802	0.21991	0.19606	0.17120
λ_3	0.20997	0.23328	0.24039	0.21053
λ_4	0.03492	0.02058		0.00273
λ_5	0.12015			
λ_6		0.08543	0.00228	
λ_7	0.01587	0.06540	0.05516	0.00009
λ_8		0.04767	0.16602	0.15269
λ_9		0.01912	0.04088	0.09227
λ_{10}			0.01064	0.02802
λ_{14}	0.00480			
λ_{15}	0.37627			0.01206
λ_{19}		0.08064		
λ_{20}		0.22798		
λ_{28}			0.00221	
λ_{30}			0.28636	0.07212
λ_{50}				0.25830
ρ_8	0.98013	0.64854	0.00749	
ρ_9	0.01987	0.34747	0.99101	0.33620
ρ_{10}		0.00399	0.00150	0.08883
ρ_{11}				0.57497
σ^*	0.9622	0.9649	0.9690	0.9718
$(\frac{E_b}{N_0})^*_{dB}$	0.3347	0.3104	0.2735	0.2485
p^*	0.1493	0.1500	0.1510	0.1517

and $\rho(x) := 0.25x^9 + 0.75x^{10}$. Here we allowed a maximum variable node degree of 75 and a maximum check node degree of 11. •

Example 3 [BILC; $r = \frac{1}{2}$]: Consider the binary-input Laplace channel (BILC) with continuous output alphabet and additive noise. More precisely, the channel is modeled by $y := x+z$, where $x \in \{\pm 1\}$ and where z is a random variables with probability density

$$p_{\mathrm{BIL}}(z) := \frac{1}{2\lambda} e^{-\frac{|z|}{\lambda}}.$$

The ultimate threshold for the BILC of rate $\frac{1}{2}$ (as given by the Shannon formula) is equal to $\lambda_{\mathrm{opt}} = 0.752$. We found the following degree distribution pair which has a threshold above $\lambda^* = 0.74$:

$$\begin{aligned}\lambda(x) := {} & 0.0869518x + 0.26831x^2 + 0.0928357x^3 \\ & + 0.0214918x^4 + 0.0120479x^5 + 0.00416287x^6 \\ & + 0.010689x^7 + 0.00271656x^8 + 0.00478356x^9 \\ & + 0.000976879x^{10} + 0.0198119x^{11} + 0.0681859x^{12} \\ & + 0.0229726x^{13} + 0.0782138x^{14} + 0.00315161x^{15} \\ & + 0.0033345x^{16} + 0.00251733x^{17} + 0.00243145x^{19} \\ & + 0.00224413x^{20} + 0.00372345x^{21} + 0.012011x^{22} \\ & + 0.00763323x^{24} + 0.0615922x^{25} + 0.0122678x^{28} \\ & + 0.00113751x^{53} + 0.00565326x^{54} + 0.0178079x^{55} \\ & + 0.0268887x^{56} + 0.000395817x^{59} + 0.00229811x^{64} \\ & + 0.0144568x^{73} + 0.126305x^{74}\end{aligned}$$

and $\rho(x) := 0.25x^9 + 0.75x^{10}$. The maximum variable node and check node degrees are again 75 and 11, respectively. •

B. Simulation Results

The concentration results proved in [1] guarantee that for sufficiently large block lengths almost every code in the given ensemble will have vanishing probability of bit error for channels with parameters below the calculated threshold. Nevertheless, the lengths required by the proofs are far beyond any practical values and one might expect that medium-sized codes will deviate significantly from the predicted performance. Given that the maximum possible number of loop-free iterations grows only logarithmically in the block length, it seems *a priori* doubtful that simulation results for practical lengths can closely match the predicted asymptotic performance. For regular LDPC codes, however, it was demonstrated in [1] that the actual convergence is much faster and that realistically sized block codes already perform close to the asymptotic value.

For irregular codes, finite-length effects not only include the deviation of the input variance from its mean and a nonzero probability of small loops but also the deviation of a given support tree from its average, i.e., for a given node the fraction of neighbors of this node with a given degree might deviate from its expected value. This effect is expected to influence the finite-length performance more severely for larger d_v and d_c. Also, when operating very close to a degree distribution pair's threshold, it will require a large number of iterations (one thousand or more) to reach target bit-error probabilities of, say, 10^{-5}. (In the limit the number of iterations converges to ∞.) Fortunately, however, a small margin from a degree distribution pair's threshold is typically enough to drastically reduce the required number of iterations.

Simulation results show that even in the irregular case the actual convergence of the performance of finite-length codes to the asymptotic performance is much faster than predicted by the bounds appearing in the analysis. Fig. 3 shows the performance of particular LDPC codes. The chosen lengths start at one thousand and go up to one million. More precisely, the lengths presented are 10^3, 10^4, 10^5, and 10^6. The maximum variable degrees appearing are 9, 20, 50, and 50, respectively. In each case, we let the decoder run for enough iterations to get the best possible performance. (The number of practically useful iterations is a function of length and, since our interest here is in the question of parameter design, we shall not address this issue.) For length 10^3, the error rates are given for systematic bits. (A specific encoder was constructed.) For length 10^4 and above, the error rate is given over all of the bits in the codeword. These graphs were not chosen entirely randomly. The degree-two nodes were made loop-free for lengths less than 10^6 and, in the length 10^3 case, all of them correspond to nonsystematic bits. The length 10^6 graph was randomly constructed except that double edges and loops with two variable nodes were avoided. For shorter lengths some small loop removal was performed. We note that, particularly for small lengths, better

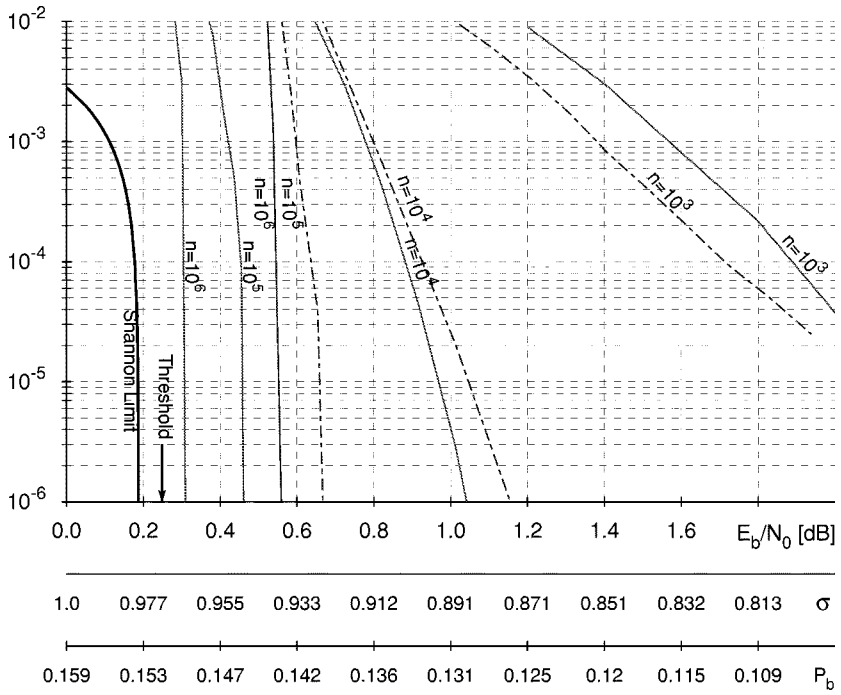

Fig. 3. Comparison between bit-error rates achieved by turbo codes (dashed curves) and LDPC codes (solid curves) of lengths $n = 10^3, 10^4, 10^5$, and 10^6. All codes are of rate one-half. Observe that longer LDPC codes outperform turbo codes and that the gap becomes the more significant the larger n is chosen. For short lengths, it appears that the structure of turbo codes gives them an edge over LDPC codes despite having a lower threshold.

performance can be achieved by using degree distribution pairs with smaller values of d_v even though such degree distribution pairs have a smaller threshold.[11] We see that for $n = 10^6$, the actual performance is quite close to the predicted asymptotic performance and that for longer lengths these codes handily beat the best known turbo code of the same length. At one million bits, the code is less than 0.13 dB away from capacity at bit-error probabilities of 10^{-6}. Given that LDPC codes have slightly lower complexity, are fully parallelizable, and allow many different decoding algorithms with a far ranging tradeoff between performance and complexity, LDPC codes can be considered serious competitors to turbo codes.

Although we spent a considerable amount of computing time on the optimization it is clear that any given degree distribution pair can be further (slightly) improved given enough patience.[12]

III. Analytic Properties of Density Evolution

In this section, we will study and exhibit some analytic properties of density evolution for belief-propagation decoders. Without loss of generality, we will assume that the all-one codeword was transmitted and that the messages are represented as log-likelihood ratios. Under the *independence assumption* we will then give a compact mathematical description of the evolution of the message densities as they proceed up the tree. In doing so we largely follow [1]. We will show that for output-symmetric channels the received message distribution is *symmetric* and that this symmetry is preserved under belief-propagation decoding. Further, we will discuss a *stability condition* of density evolution which stems from analyzing the convergence behavior of density evolution under the assumption that small error probabilities have already been achieved in the evolution process. Finally, we will give a partial characterization of the threshold in terms of the (non)existence of fixed points of the density evolution recursion. For each topic, we will motivate our definitions/statements by considering the equivalent definitions/statements for the simple case of the BEC. Quite surprisingly, given the uncharacteristically simple nature of the BEC, we will find that many results extend to the general case.

A. Belief Propagation

The main result of this section, stated in Theorem 2, is an explicit recursion which connects the distributions of messages passed from variable nodes to check nodes at two consecutive iterations of *belief propagation*. In the case of the BEC, this task has been accomplished in [2], [24], [10] where it was shown that the expected fraction of erasure messages which are passed in the ℓth iteration, call it x_ℓ, evolves as

$$x_\ell = x_0 \lambda(1 - \rho(1 - x_{\ell-1})), \qquad \ell \geq 1.$$

For general binary-input memoryless output-symmetric channels, the situation is much more involved since one has to keep track of the evolution of general distributions, which usually cannot be parameterized by a single parameter.

Let us begin by recalling the *belief-propagation* algorithm. We will use the standard binary PAM map $0 \mapsto 1$, $1 \mapsto -1$ throughout. At each iteration, messages are passed along the edges of the graph from variable nodes to their incident check nodes and back. The messages are typically real-valued but they

[11]The degree distribution pairs presented here are those giving the highest threshold under a maximum degree constraint. For small graphs it is not always best to pick the degree distribution pair with the highest threshold. When designing degree distribution pairs for small graphs, it can be advantageous to look for the highest possible threshold under an appropriate constraint on the allowed number of iterations.

[12]Indeed this has been accomplished in [16].

can also take on the values $\pm\infty$, reflecting the situation where some bits are known with absolute certainty.

Generically, messages which are sent in the ℓth iteration will be denoted by $\mathsf{m}^{(\ell)}$. By $\mathsf{m}_{\mathsf{vc}}^{(\ell)}$ we denote the message sent from the variable node v to its incident check node c, while by $\mathsf{m}_{\mathsf{cv}}^{(\ell)}$ we denote the message passed from check node c to its incident variable node v. Each message represents a quantity $\ln(p^+/p^-)$, where $p^+ = p(\boldsymbol{x} = 1|y)$, $p^- = p(\boldsymbol{x} = -1|y)$, x is the random variable describing the codeword bit value associated to the variable node v, and y is the random variable describing all the information incorporated into this message. By Bayes rule we have

$$\mathsf{m} = \ln \frac{p(\boldsymbol{x}=1|y)}{p(\boldsymbol{x}=-1|y)} = \ln \frac{p(y|\boldsymbol{x}=1)}{p(y|\boldsymbol{x}=-1)}$$

since $\boldsymbol{x}$ is equally likely ± 1. The message m is the log-likelihood ratio of the random variable $\boldsymbol{x}$ (under the *independence assumption*).

As we will see shortly, to represent the updates performed by *check nodes* an alternative representation of the messages is appropriate. Let us define a map

$$\gamma\colon [-\infty, +\infty] \to \mathrm{GF}(2) \times [0, +\infty]$$

as follows. Given $x \in [-\infty, +\infty]$, $x \neq 0$, let

$$\gamma(x) := (\gamma_1(x), \gamma_2(x)) := \left(\operatorname{sgn} x, -\ln\tanh\left|\frac{x}{2}\right|\right). \quad (2)$$

Several remarks are in order. We define $-\ln(0) := +\infty$. Further, we make the following slightly unconventional *probabilistic* definition of the sign function:

$$\operatorname{sgn}(x) := \begin{cases} 0, & \text{if } x > 0 \\ 0, & \text{with probability } \frac{1}{2} \text{ if } x = 0 \\ 1, & \text{with probability } \frac{1}{2} \text{ if } x = 0 \\ 1, & \text{if } x < 0. \end{cases}$$

The sign assignment when $x = 0$ does not effect the belief-propagation algorithm but it is useful for certain definitions to follow.

Let m_0 be the log-likelihood ratio of the codeword bit $\boldsymbol{x} = \pm 1$ associated to the variable node v conditioned only on the channel observation of this bit. The update equations for the messages under *belief propagation* are then the following:

$$\mathsf{m}_{\mathsf{vc}}^{(\ell)} = \begin{cases} \mathsf{m}_0, & \text{if } \ell = 0 \\ \mathsf{m}_0 + \displaystyle\sum_{\mathsf{c}' \in C_{\mathsf{v}} \setminus \{\mathsf{c}\}} \mathsf{m}_{\mathsf{c}'\mathsf{v}}^{(\ell)}, & \text{if } \ell \geq 1 \end{cases} \quad (3)$$

$$\mathsf{m}_{\mathsf{cv}}^{(\ell)} = \gamma^{-1}\left(\sum_{\mathsf{v}' \in V_{\mathsf{c}} \setminus \{\mathsf{v}\}} \gamma\left(\mathsf{m}_{\mathsf{v}'\mathsf{c}}^{(\ell-1)}\right)\right) \quad (4)$$

where C_{v} is the set of check nodes incident to variable node v, and V_{c} is the set of variable nodes incident to check node c.[13]

It is easy to check that the belief-propagation decoder as defined above fulfills the symmetry conditions discussed in Section I. Therefore, if we restrict ourselves to output-symmetric channels, then we can assume that the all-one codeword was transmitted without loss of generality since, under these assumptions, the conditional error probability is independent of the transmitted codeword, see [1]. In the sequel, we will only be interested in output-symmetric channels and we will therefore always assume that the all-one codeword was transmitted. Note that in this case any actually observed y must fulfill $0 < p(y|\boldsymbol{x} = 1)$. It follows that the log-likelihood ratios $\log \frac{p(y|\boldsymbol{x}=1)}{p(y|\boldsymbol{x}=-1)}$ take values in the range $(-\infty, +\infty]$. A short glance at the above update equations then shows that *all* messages sent during the course of belief propagation lie in this range, i.e., that no message can ever take on the value $-\infty$.

B. Distributions

For some channels, e.g., the BEC and the BSC, the density of received log-likelihood ratios is discrete. For others, e.g., the BIAWGNC, the density is continuous. In the first case, the message densities will themselves be discrete and in the second case, the message densities will be continuous. In order to be able to treat all these cases in a uniform manner we shall work with a fairly general class of distributions. The reader only casually interested in the technical aspects of this paper is advised to think of smooth densities and their associated continuous distributions and to only skim the current section.

Let $\mathcal{F}$ denote the space of right-continuous, nondecreasing functions F defined on $\mathbb{R}$ satisfying $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to+\infty} F(x) \leq 1$. To each $F \in \mathcal{F}$ we associate a random variable z over $(-\infty, +\infty]$. The random variable z has *law* or *distribution* F, i.e.,

$$\Pr\{z \in (-\infty, x]\} = F(x).$$

The reason we allow $\lim_{x\to+\infty} F(x) \leq 1$ rather than $\lim_{x\to+\infty} F(x) = 1$ is to permit z to have some probability mass at $+\infty$, indeed

$$\Pr\{z = +\infty\} = 1 - \lim_{x\to+\infty} F(x).$$

A random variable z over $(-\infty, +\infty]$ is completely specified by its distribution $F_z \in \mathcal{F}$. Given an element $F \in \mathcal{F}$ we define $F^-(x)$ to be the left limit of F at x, i.e.,

$$F^-(x) := \lim_{y\uparrow x} F(y).$$

Note that F^- is left continuous.

We will work with "densities" over $(-\infty, +\infty]$ which, formally, can be treated as (Radon–Nikodyn) derivatives of elements of $\mathcal{F}$. The derivative, when it exists, is the density of the associated random variable z over $(-\infty, +\infty)$ although there may be an additional point mass at $+\infty$: recall $\Pr\{z = +\infty\} = 1 - F^-(\infty)$. We will use densities primarily in the following way. The Lebesgue–Stieltjes integral $\int_{\mathbb{R}} h(x)\,dF(x)$ is well-defined for, e.g., nonnegative continuous functions h and $F \in \mathcal{F}$. If f is the density corresponding to the distribution F we will write $\int_{\mathbb{R}} h(x)f(x)\,dx$ as a proxy for $\int_{\mathbb{R}} h(x)\,dF(x)$.[14]

Consider the update equations of belief propagation given in (3) and (4). Note that they consist of the following two com-

[13]The reader might wonder how $((+\infty) + (-\infty))$ is defined in (3), but it is impossible under belief propagation for a node to receive both the message $+\infty$ and the message $-\infty$. (Either message implies absolute certainty of the value of the associated bit but with opposite signs.)

[14]If $\lim_{x\to+\infty} F(x) < 1$ and $\lim_{x\to+\infty} h(x)$ exists then one could/should include the term $(1 - \lim_{x\to+\infty} F(x))(\lim_{x\to+\infty} h(x))$ in the definition of $\int_{-\infty}^{+\infty} h(x)f(x)\,dx$. For our purposes, however, it will always be the case when we use this notation that either $\lim_{x\to+\infty} F(x) = 1$ or $\lim_{x\to+\infty} h(x) = 0$, so we need not be concerned with this issue.

ponents: i) summation of messages, where these messages are either in log-likelihood ratio representation or represented as "sign-and-reliability" and ii) change of representation. We are interested in the evolution of the message distributions under the *independence assumption*. Therefore, we will now discuss how distributions evolve when independent random variables (in either representation) are summed and when the representation of such variables is changed.

Given $F, G \in \mathcal{F}$ their *convolution* $F \otimes G \in \mathcal{F}$ is defined by[15]

$$(F \otimes G)(x) := \int_{\mathbb{R}} F(x-y)\, dG(y) = \int_{\mathbb{R}} G(x-y)\, dF(y).$$

This generalizes the notion of convolution of densities for if F and G have corresponding densities f and g, respectively, then $(F \otimes G)(x)$ is the distribution corresponding to the density $f \otimes g$. We write $f \otimes g$ for arbitrary densities to indicate the density associated to the distribution $F \otimes G \in \mathcal{F}$. It is easy to check that if z_1 and z_2 are independent random variables over $(-\infty, +\infty]$ with distributions F_{z_1} and F_{z_2}, respectively, then the distribution of $z_1 + z_2$ is $F_{z_1} \otimes F_{z_2}$ (as is the case for independent random variables defined over $(-\infty, \infty)$).

Now, suppose we have a random variable z over $(-\infty, +\infty]$ with distribution F_z and we wish to describe the "distribution" of the random variable $\gamma(z) = (\gamma_1(z), \gamma_2(z))$, where $\gamma_1(x)$ and $\gamma_2(x)$ are defined as in (2). We approach this problem by assigning two connected distributions associated to $\gamma_2(z)$ under the conditions $\gamma_1(z) = 0$ and $\gamma_1(z) = 1$, respectively.

Any function G over GF $(2) \times [0, +\infty)$ can be written as

$$G(s, x) := \chi_{\{s=0\}} G^0(x) + \chi_{\{s=1\}} G^1(x)$$

where $\chi_{\{s=a\}}$ denotes the characteristic function of the set $\{s = a\}$, i.e., $\chi_{\{s=a\}} = 1$ if $s = a$ and $\chi_{\{s=a\}} = 0$ otherwise. Let $\mathcal{G}$ denote the space of functions over GF $(2) \times [0, +\infty)$ such that $G^0(x)$ and $G^1(x)$ are nondecreasing and right continuous

$$\lim_{x \to +\infty} G^0(x) \geq \lim_{x \to +\infty} G^1(x)$$

and such that $G^0(0) \geq 0$ and $G^1(0) = 0$. (The last two conditions correspond to the conditions $\lim_{x \to \infty} F(x) \leq 1$ and $\lim_{x \to -\infty} F(x) = 0$ for functions in $\mathcal{F}$.)

Given a random variable $z \in (-\infty, +\infty]$ with distribution F_z we define the "distribution" of $\gamma(z)$ as

$$\Gamma(F_z)(s, x) = \chi_{\{s=0\}} \Gamma_0(F_z)(x) + \chi_{\{s=1\}} \Gamma_1(F_z)(x) \quad (5)$$

where

$$\Gamma_0(F_z)(x) = 1 - F_z^{-}\left(-\ln \tanh \frac{x}{2}\right)$$

and

$$\Gamma_1(F_z)(x) = F_z\left(\ln \tanh \frac{x}{2}\right).$$

Thus

$$\begin{aligned}\Gamma_0(F_z)(x) &= \Pr\{\gamma_1(z) = 0,\, \gamma_2(z) \leq x\} \\ &= \Pr\{z \geq -\ln \tanh \tfrac{x}{2}\}\end{aligned}$$

and

$$\begin{aligned}\Gamma_1(F_z)(x) &= \Pr\{\gamma_1(z) = 1,\, \gamma_2(z) \leq x\} \\ &= \Pr\{z \leq \ln \tanh \tfrac{x}{2}\}.\end{aligned}$$

[15]The integral is defined for almost all x and right continuity determines the rest.

Note that $\Gamma(F_z) \in \mathcal{G}$, and, in particular

$$\lim_{x \to +\infty} \Gamma_0(F_z)(x) - \lim_{x \to +\infty} \Gamma_1(F_z)(x) = \Pr\{z = 0\}.$$

Let $G = \chi_{\{s=0\}} G^0 + \chi_{\{s=1\}} G^1$ be an element of $\mathcal{G}$. We speak of densities over GF $(2) \times [0, +\infty]$

$$g(s, x) := \chi_{\{s=0\}} g^0(x) + \chi_{\{s=1\}} g^1(x),$$

by substituting for G^0 and G^1 their associated densities. The definition is analogous to that used for $\mathcal{F}$ except that, here, g^0 has a point mass at $x = 0$ of magnitude $1 - G^0(0)$ and both g^0 and g^1 have point masses at $x = +\infty$ of magnitude $\frac{1}{2}(\lim_{x \to +\infty} F^0(x) - \lim_{x \to +\infty} F^1(x))$. For $\Gamma(F_z)$ this corresponds to assigning $\mathsf{sgn}(0)$ to $\{1, 0\}$ equally likely. We shall only be interested in densities over GF $(2) \times [0, +\infty]$ that satisfy these conditions.

The function Γ has a well-defined inverse. Given

$$G = \chi_{\{s=0\}} G^0 + \chi_{\{s=1\}} G^1 \in \mathcal{G}$$

we have

$$\begin{aligned}\Gamma^{-1}(G)(x) = \chi_{\{x>0\}} G^0\left(-\ln \tanh \frac{x}{2}\right) \\ + \chi_{\{x<0\}} G^1\left(-\ln \tanh \frac{-x}{2}\right)\end{aligned} \quad (6)$$

and

$$\Gamma^{-1}(G)(0) = \lim_{x \to +\infty} G^0(x).$$

It is easy to check that Γ^{-1}: $\mathcal{G} \to \mathcal{F}$ and that $\Gamma^{-1}(\Gamma(F)) = F$ for all $F \in \mathcal{F}$. Further, Γ and Γ^{-1} are additive operators on the spaces $\mathcal{F}$ and $\mathcal{G}$, respectively.

For convenience, although it constitutes an abuse of notation, we will apply Γ and Γ^{-1} to densities. It is implicitly understood that the notation is a representation of the appropriate operation applied to distributions.

The space $\mathcal{G}$ has a well-defined convolution. Here, the convolution of two distributions $\chi_{\{s=1\}} G^0 + \chi_{\{s=1\}} G^1$ and $\chi_{\{s=0\}} H^0 + \chi_{\{s=1\}} H^1$ is the distribution

$$\begin{aligned}\chi_{\{s=0\}}((G^0 \otimes H^0) + (G^1 \otimes H^1)) \\ + \chi_{\{s=1\}}((G^0 \otimes H^1) + (G^1 \otimes H^0))\end{aligned}$$

where, here, $\otimes$ denotes the (one-sided) convolution of standard distributions. In other words, the new convolution is a convolution over the group GF $(2) \times [0, +\infty)$. By abuse of notation, we denote this new convolution by the same symbol $\otimes$. Again, we shall allow the convolution operator to act on the densities associated to elements of $\mathcal{G}$ with the implicit understanding that the above provides the rigorous definition.

If z_1 and z_2 are independent random variables over GF $(2) \times [0, +\infty]$ with distributions $G_{z_1}, G_{z_2} \in \mathcal{G}$, respectively, then the distribution of $z_1 + z_2$ is $G_{z_1} \otimes G_{z_2}$.

Example 4: Let us give a few examples of densities of interest. By Δ_z, $z \in \mathbb{R}$, we denote the density corresponding to the distribution $\chi_{\{x \geq z\}} \in \mathcal{F}$. In other words, $\Delta_z(x) = \delta(x-z)$, where δ denotes the Dirac delta function. A special case is Δ_∞ which corresponds to the distribution $0 \in \mathcal{F}$.

The density $\Gamma(\Delta_0)$ is given by

$$\Gamma(\Delta_0)(s, x) = \frac{1}{2}\chi_{\{s=0\}} \Delta_\infty(x) + \frac{1}{2}\chi_{\{s=1\}} \Delta_\infty(x).$$

Expressed using distributions, we have $\Gamma(\chi_{\{x\geq 0\}})(s, x) = 0$.

The density $\Gamma(\Delta_\infty)$ is given by $\Gamma(\Delta_\infty)(s, x) = \chi_{\{s=0\}}\Delta_0$. Expressed using distributions, we have $\Gamma(0)(s, x) = \chi_{\{s=0\}}$.

C. Description of Density Evolution

The symbols P_ℓ and Q_ℓ will be shorthand notations for the densities of the random variables $\mathsf{m}_{\mathsf{vc}}^{(\ell)}$ and $\mathsf{m}_{\mathsf{cv}}^{(\ell)}$, respectively. We will use the notation $\int \mathrm{P}_\ell$ and $\int \mathrm{Q}_\ell$ to denote the associated distributions.

By (4), we see that the random variable describing the message passed from check node c to variable node v is the image under γ^{-1} of a sum of random variables from GF$(2)\times[0, +\infty]$. These random variables are independent by the *independence assumption*. So, the density of their sum is the convolution of their densities.

Let the graph have degree distribution pair (λ, ρ) where

$$\lambda(x) = \sum_{i\geq 2} \lambda_i x^{i-1} \quad \text{and} \quad \rho(x) = \sum_{i\geq 2} \rho_i x^{i-1}.$$

Recall that the fraction of edges connected to a variable node of degree i is λ_i, and the fraction of edges connected to a check node of degree i is ρ_i. Thus, a randomly chosen edge in the graph is connected to a check node of degree i with probability ρ_i. Therefore, with probability ρ_i the sum in (4) has $(i-1)$ terms, corresponding to the edges connecting c to all its neighbors *other than* v. We conclude that, in this case, the density of $\mathsf{m}_{\mathsf{cv}}^{(\ell)}$ is equal to $\Gamma^{-1}(\Gamma(\mathrm{P}_{\ell-1})^{\otimes(i-1)})$. Summing up over all the possibilities for the degrees of the check node c, we see that the density of the message $\mathsf{m}_{\mathsf{cv}}^{(\ell)}$ equals

$$\mathrm{Q}_\ell = \Gamma^{-1}(\rho(\Gamma(\mathrm{P}_{\ell-1}))) := \Gamma^{-1}\left(\sum_{i\geq 2} \rho_i (\Gamma(\mathrm{P}_{\ell-1}))^{\otimes(i-1)}\right). \tag{7}$$

This equation also explains the unusual definition $\sum_{i\geq 2}\rho_i x^{i-1}$ (rather than $\sum_{i\geq 2}\rho_x^i$) for $\rho(x)$.

A recursion for P_ℓ in terms of Q_ℓ is derived similarly and is quite straightforward. The density of the message passed from check node c to variable node v at round ℓ is equal to Q_ℓ. At v the incoming messages from all check nodes other than c are added to m_0, the received value for v, and the result is sent back to c. Since, by the *independence assumption* the random variables describing these messages are independent, the density of this message equals

$$\mathrm{P}_\ell = \mathrm{P}_0 \otimes \lambda(\mathrm{Q}_\ell) := \mathrm{P}_0 \otimes \sum_{i\geq 2} \lambda_i (\mathrm{Q}_\ell)^{\otimes(i-1)} \tag{8}$$

where P_0 is the density of the random variable describing the channel.

Combining (7) and (8) we obtain the desired recursion for P_ℓ in terms of $\mathrm{P}_{\ell-1}$.

Theorem 2: For a given binary-input output-symmetric memoryless channel let P_0 denote the initial message density of log-likelihood ratios, assuming that the all-one codeword was transmitted. If, for a fixed degree distribution pair (λ, ρ), P_ℓ denotes the density of the messages passed from the variable nodes to the check nodes at the ℓth iteration of belief propagation then, under the *independence assumption*

$$\mathrm{P}_\ell = \mathrm{P}_0 \otimes \lambda(\Gamma^{-1}(\rho(\Gamma(\mathrm{P}_{\ell-1})))) \tag{9}$$

where the operators Γ and Γ^{-1} are defined in (5) and (6), respectively.

Discussion: The result of the previous theorem is in complete agreement with the result for the erasure channel obtained previously in [2], [10]; in fact, it contains it as a special case. Indeed, using distributions in this case, we have $\int \mathrm{P}_\ell(x) = x_\ell \chi_{\{x\geq 0\}}$. It is easily seen that

$$\Gamma\left(\int \mathrm{P}_\ell\right)(s, y) = \chi_{\{s=0\}}(1 - x_\ell).$$

We then obtain

$$\rho\left(\Gamma\left(\int \mathrm{P}_\ell\right)\right)(s, y) = \chi_{\{s=0\}}\rho(1 - x_\ell)$$

and

$$\Gamma^{-1}\rho(\Gamma\left(\int \mathrm{P}_\ell\right))(x) = (1 - \rho(1 - x_\ell))\chi_{\{x\geq 0\}}.$$

Finally

$$\left(\int \mathrm{P}_0\right) \otimes \lambda\left(\Gamma^{-1}\rho\left(\Gamma\left(\int \mathrm{P}_\ell\right)\right)\right) = x_0\lambda(1 - \rho(1 - x_\ell))\chi_{\{x\geq 0\}}$$

and we recover

$$x_\ell = x_0\lambda(1 - \rho(1 - x_{\ell-1}))$$

which is the same as the formula proved in [2], [24], [10].

D. Symmetry

Definition 1 [Symmetry]: We call $F \in \mathcal{F}$ symmetric if

$$\int_{\mathbb{R}} h(x)\, dF(x) = \int_{\mathbb{R}} e^{-x} h(-x)\, dF(x)$$

for any function h for which the integral exists. For densities, the equivalent statement is that the density f is symmetric if $f(x) = e^x f(-x)$ for $x \in \mathbb{R}$.

Example 5: Δ_0 and Δ_∞ are symmetric. •

Our aim in this section is to prove that the density functions of messages passed from variable to check nodes during the *belief propagation* are symmetric, provided that the channel is output-symmetric.

Theorem 3: For a given binary-input memoryless output-symmetric channel let P_0 denote the initial message density of log-likelihood ratios, assuming that the all-one word was transmitted. For a fixed degree distribution pair (λ, ρ) define

$$\mathrm{P}_\ell := \mathrm{P}_0 \otimes \lambda(\Gamma^{-1}(\rho(\Gamma(\mathrm{P}_{\ell-1})))), \qquad \ell \geq 1.$$

Then P_ℓ is symmetric.

First note the following immediate corollary. For a distribution F with density f we define the *error probability operator* $\mathrm{P_e}(f) := \frac{1}{2}(F^-(0) + F(0))$.

Corollary 1: $\mathrm{P_e}(\mathrm{P}_\ell)$ converges to zero iff P_ℓ converges[16] to Δ_∞.

[16]For a definition of convergence see Section III-F

Before proceeding with the proof of Theorem 3, we exemplify the symmetry of some well-known noise densities.

Example 6 [BEC]: For the erasure channel the initial message density is $\mathrm{P}_0 = \epsilon\Delta_0 + (1-\epsilon)\Delta_\infty$. Since linear combinations of symmetric functions are symmetric, the result follows, as Δ_0 and Δ_∞ are symmetric. •

Example 7 [BSC]: For the BSC with crossover probability δ the initial message density is

$$\mathrm{P}_0(y) := \delta\Delta_{-\log\frac{1-\delta}{\delta}} + (1-\delta)\Delta_{\log\frac{1-\delta}{\delta}}.$$

For $y \neq y_0 := \log\frac{1-\delta}{\delta}$ we have $\mathrm{P}_0(\pm y) = 0$, so to show symmetry we only need to prove that $\mathrm{P}_0(y_0) = e^{y_0}\mathrm{P}_0(-y_0)$, which is straightforward. •

Example 8 [BIAWGNC]: Here the initial message density is

$$\mathrm{P}_0(y) := \sqrt{\frac{\sigma^2}{8\pi}}\, e^{(-\frac{(y-\frac{2}{\sigma^2})^2\sigma^2}{8})}.$$

The symmetry condition is then verified by

$$\begin{aligned}\mathrm{P}_0(y) &= \sqrt{\frac{\sigma^2}{8\pi}}\, e^{-\frac{(y-\frac{2}{\sigma^2})^2\sigma^2}{8}} \\ &= \sqrt{\frac{\sigma^2}{8\pi}}\, e^{-\frac{(-y-\frac{2}{\sigma^2})^2\sigma^2}{8}}\, e^y = \mathrm{P}_0(-y)e^y.\end{aligned}$$ •

Example 9 [BILC]: As a final example, the initial message density for the Laplace channel is given by

$$\mathrm{P}_0(y) = \frac{1}{2}\, e^{-\frac{2}{\lambda}\Delta_{-\frac{2}{\lambda}}} + \frac{1}{4}\, e^{\frac{y-\frac{2}{\lambda}}{2}\chi_{|y|\leq\frac{2}{\lambda}}} + \frac{1}{2}\,\Delta_{\frac{2}{\lambda}}$$

Again, it is easy to check that the symmetry condition is fulfilled. •

For the proof of Theorem 3, we proceed in several steps. We will first establish the symmetry of the initial message density under general conditions. Once this is done, we will prove that symmetry is preserved under convolutions. Using Theorem 2, it then remains to show that if F is symmetric, then $\Gamma^{-1}(\rho(\Gamma(F)))$ is symmetric as well. To do this, we first characterize symmetry in $\mathcal{G}$ and prove that, also in this representation, symmetry is preserved under convolutions. We will present the proofs of the next three results using densities. A rigorous proof can be obtained by *formally* translating the subsequent proof into the language of distributions.

Proposition 1: Consider a binary-input memoryless output-symmetric channel and let P_0 be the initial message density in log-likelihood ratio form under the all-one word assumption. Then P_0 is symmetric.

Proof: From the channel symmetry condition we obtain

$$\mathrm{L}(y) := \log\frac{p(y|\boldsymbol{x}=1)}{p(y|\boldsymbol{x}=-1)} = \log\frac{p(-y|\boldsymbol{x}=-1)}{p(-y|\boldsymbol{x}=1)} = -\mathrm{L}(-y).$$

Therefore

$$\begin{aligned}e^u\mathrm{P}_0(-u) &= e^u p(y\in L^{-1}(-u)|\boldsymbol{x}=1) \\ &= e^u p(-y\in L^{-1}(u)|\boldsymbol{x}=1) \\ &= e^u p(y\in L^{-1}(u)|\boldsymbol{x}=-1) \\ &= p(y\in L^{-1}(u)|\boldsymbol{x}=1) \\ &= \mathrm{P}_0(u).\end{aligned}$$ □

Next we will show that symmetry is preserved under convolutions.

Lemma 1: The convolution of symmetric distributions is symmetric.

Proof: Let f and g be two symmetric densities. Then

$$\begin{aligned}\int_{-\infty}^{\infty} f(x-y)g(y)\,dy &= \int_{-\infty}^{\infty} e^{x-y}f(y-x)e^y g(-y)\,dy \\ &= \int_{-\infty}^{\infty} e^x f(y-x)g(-y)\,dy \\ &= e^x\int_{-\infty}^{\infty} f(-x-y)g(y)\,dy.\end{aligned}$$ □

Consider the iteration formula of Theorem 2

$$\mathrm{P}_\ell = \mathrm{P}_0 \otimes \lambda(\Gamma^{-1}(\rho(\Gamma(\mathrm{P}_{\ell-1})))).$$

We already know that P_0 is symmetric. Hence, we can prove by induction that P_ℓ is symmetric for all ℓ if we show that $\Gamma^{-1}(\rho(\Gamma(f)))$ is symmetric when f is.

Definition 2: We say $G \in \mathcal{G}$ is symmetric if

$$\int_0^\infty h(x)\tanh(x/2)\,dG^0(x) = \int_0^\infty h(x)\,dG^1(x)$$

for any nonnegative continuous function h.

Proposition 2:

1) The function $F \in \mathcal{F}$ is symmetric if and only if $\Gamma(F) \in \mathcal{G}$ is symmetric.
2) If G, $H \in \mathcal{G}$ are symmetric distributions over $\mathrm{GF}(2)\times[0,+\infty)$, then so is their convolution $G\otimes H$.

Proof:

1) Let f be the density associated to F, i.e., $f(x) = \frac{d}{dx}F(x)$ and $\lim_{x\to+\infty}F(x) = 1$. Then we have
$$\Gamma(f)(s,r) = \chi_{\{s=0\}}\frac{f(-\ln\tanh(r/2))}{\sinh(r)} + \chi_{s=1}\frac{f(\ln\tanh(r/2))}{\sinh(r)}.$$
Using the symmetry condition for f, we see that
$$\tanh(r/2)f(-\ln\tanh(r/2)) = f(\ln\tanh(r/2))$$
so $\Gamma(f)$ is symmetric.
2) By definition, the convolution of g and h is the density u where
$$\begin{aligned}u^0 &= (g^0\otimes h^0) + (g^1\otimes h^1) \\ u^1 &= (g^0\otimes h^1) + (g^1\otimes h^0).\end{aligned}$$
We need to show that $\tanh(r/2)u^0(r) = u^1(r)$. Since g is symmetric, we have $\tanh(r/2)g^0(r) = g^1(r)$, so
$$g^0(r) + g^1(r) = g^0(r)\tfrac{2e^r}{e^r+1}.$$
Also
$$g^0(r) - g^1(r) = g^0(r)\tfrac{2}{e^r+1}.$$
We, therefore, have
$$\begin{aligned}&\frac{1}{2}((g^0+g^1)\otimes(h^0+h^1))(y) \\ &\quad= e^y\int_0^\infty g^0(r)h^0(y-r)\frac{dr}{(e^r+1)(e^{y-r}+1)} \\ &\frac{1}{2}((g^0-g^1)\otimes(h^0-h^1))(y) \\ &\quad= \int_0^\infty g^0(r)h^0(y-r)\frac{dr}{(e^r+1)(e^{y-r}+1)}.\end{aligned}$$

Since u^0 is the sum of the above two quantities and u^1 is their difference we obtain

$$u^1(y) = \frac{e^y - 1}{e^y + 1} u^0(y) = \tanh(y/2) u^0(y)$$

and we are done. □

We are now able to prove our main theorem.

Proof of Theorem 3: As outlined above, we use induction and the density evolution formula of Theorem 2. The induction is anchored by Proposition 1: P_0 is symmetric. Assume now that $\mathrm{P}_{\ell-1}$ is symmetric. By Proposition 2 part 1), $\Gamma(\mathrm{P}_\ell)$ is symmetric, and by part 2) of the same proposition $\rho(\Gamma(\mathrm{P}_\ell))$ is also symmetric. Part 1) of that proposition implies that $\Gamma^{-1}(\rho(\Gamma(\mathrm{P}_{\ell-1})))$ is symmetric. Finally, Lemma 1 shows that $\mathrm{P}_\ell = \mathrm{P}_0 \otimes \lambda(\Gamma^{-1}(\rho(\Gamma(\mathrm{P}_{\ell-1}))))$ is symmetric as well. □

E. Stability

Consider density evolution for the BEC. Recall that, for a given degree distribution pair (λ, ρ), the evolution of the expected fraction of erasure messages emitted in the ℓth iteration, call it x_ℓ, is given by

$$x_\ell = x_\ell(x_0) = x_0 \lambda(1 - \rho(1 - x_{\ell-1})), \qquad \ell \geq 0 \tag{10}$$

where x_0, the initial fraction of erasures, is equal to the erasure probability of the channel. Recall further that the threshold $x_0^*(\lambda, \rho)$ was defined as the supremum of all values x_0, $x_0 \leq 1$, such that $x_\ell(x_0)$ converges to zero as ℓ tends to infinity. It is easy to derive an upper bound on x_0^* by looking at the behavior of this recursion for small values of x_ℓ. Expanding the right-hand side of (10) into a Taylor series around zero we get

$$x_\ell = x_0 \lambda'(0) \rho'(1) x_{\ell-1} + O(x_{\ell-1}^2). \tag{11}$$

Clearly, for sufficiently small x_ℓ, the convergence behavior will be determined by the term linear in x_ℓ. More precisely, the convergence will depend on whether $x_0 \lambda'(0) \rho'(1)$ is smaller or larger than one. The precise statement is given in the following.

Theorem 4 [Stability Condition for the BEC—[15]][17]: Assume we are given a degree distribution pair (λ, ρ) and a real number x_0, $x_0 \in [0, 1]$. For $\ell \geq 1$ define

$$x_\ell(x_0) := x_0 \lambda(1 - \rho(1 - x_{\ell-1})).$$

[Necessity] If $\lambda'(0)\rho'(1) > \frac{1}{x_0}$ then there exists a constant $\xi = \xi(\lambda, \rho, x_0)$, $\xi > 0$, such that for all $\ell \in \mathbb{N}$, $x_\ell(x_0) > \xi$.

[Sufficiency] If $\lambda'(0)\rho'(1) < \frac{1}{x_0}$ then there exists a constant $\xi = \xi(\lambda, \rho, x_0)$, $\xi > 0$, such that if, for some $\ell \in \mathbb{N}$, $x_\ell(x_0) \leq \xi$ then $x_\ell(x_0)$ converges to zero as ℓ tends to infinity.

Discussion: Note that $x_0 \lambda(1 - \rho(1 - 0)) = 0$ for any initial erasure fraction x_0, so that zero is a *fixed point* of the recursion given in (10). Therefore, it is natural to think of the above condition as a *stability* condition of the fixed point at zero.

[17]The result proved in [15] does not follow the derivation as in the theorem. Rather, it uses the condition that the decoding is successful if *and only if* the inequality $x_0 \lambda(1 - \rho(1 - x)) < x$ is valid on $(0, x_0)$.

This stability condition has far-reaching implications in the case of the BEC. To name just the most important one, it immediately gives rise to the bound

$$x_0^*(\lambda, \rho) \leq \frac{1}{\lambda'(0)\rho'(1)}. \tag{12}$$

Further, it was shown in [17, Theorem 3.1] that for any sequence of capacity achieving degree distribution pairs this inequality becomes tight. ◇

Given the important role that the stability condition plays in the case of the BEC it is natural to ask whether an equivalent condition can be formulated for general binary-input memoryless output-symmetric channels. Fortunately, the answer is in the affirmative and our main result along this line is summarized in the following.

Theorem 5 [General Stability Condition]: Assume we are given a degree distribution pair (λ, ρ) and a symmetric density P_0. For $\ell \geq 1$ define

$$\mathrm{P}_\ell := \mathrm{P}_0 \otimes \lambda(\Gamma^{-1}(\rho(\Gamma(\mathrm{P}_{\ell-1})))).$$

Let

$$r := -\ln\left(\int_{\mathbb{R}} \mathrm{P}_0(x) e^{-\frac{x}{2}}\, dx\right)$$

and assume that $\int_{\mathbb{R}} e^{sx} d(\int \mathrm{P}_0)(x) < \infty$ for all s in some neighborhood of zero.

[Necessity] If $\lambda'(0)\rho'(1) > e^r$ then there exists a constant $\xi = \xi(\lambda, \rho, \mathrm{P}_0)$, $\xi > 0$, such that for all $\ell \in \mathbb{N}$, $\mathrm{P_e}(\mathrm{P}_\ell) > \xi$.

[Sufficiency] If $\lambda'(0)\rho'(1) < e^r$ then there exists a constant $\xi = \xi(\lambda, \rho, \mathrm{P}_0)$, $\xi > 0$, such that if for some $\ell \in \mathbb{N}$, $\mathrm{P_e}(\mathrm{P}_\ell(\mathrm{P}_0)) \leq \xi$ then $\mathrm{P_e}(\mathrm{P}_\ell)$ converges to zero as ℓ tends to infinity.

Discussion: As for the BEC, the stability condition gives rise to an upper bound on the threshold. Assume that the channel family is parameterized by the real parameter δ and assume further that there exists a unique number $\overline{\delta}$ such that

$$\int_{\mathbb{R}} \mathrm{P}_0^{\overline{\delta}}(x) e^{-\frac{x}{2}}\, dx = \frac{1}{\lambda'(0)\rho'(1)}$$

where $\mathrm{P}_0^\delta(x)$ is the message density of the received values corresponding to the channel with parameter δ. Then $\delta^*(\lambda, \rho) \leq \overline{\delta}$. We note that for some codes, e.g., *cycle codes*, the stability condition determines the threshold exactly, see [18], [19] for some specific examples. ◇

In this paper, we will only prove the necessity of the stated stability condition. Demonstrating sufficiency is quite involved and the proof can be found in [20]. Before venturing into the proof of the necessity of the stability condition let us calculate the stability condition explicitly for various channels.

Example 10 [BEC]: For the BEC (see Example 6) we have

$$e^{-r} = \int_{\mathbb{R}} [\epsilon \Delta_0 + (1 - \epsilon)\Delta_\infty] e^{-x/2}\, dx = \epsilon.$$

Therefore, the stability condition[18] reads

$$\lambda'(0)\rho'(1) < \frac{1}{\epsilon}$$

as stated above and as previously obtained in [15]. •

Example 11 [BSC]: For the BSC (see Example 7) we have

$$\begin{aligned} e^{-r} &= \int_{\mathbb{R}} \left[\delta\Delta_{-\log\frac{1-\delta}{\delta}} + (1-\delta)\Delta_{\log\frac{1-\delta}{\delta}}\right] e^{-x/2}\,dx \\ &= 2\sqrt{\delta(1-\delta)}. \end{aligned}$$

It follows that the stability condition[19] for the BSC is given by

$$\lambda'(0)\rho'(1) < \frac{1}{2\sqrt{\delta(1-\delta)}}.$$ •

Example 12 [BIAWGNC]: For the BIAWGNC (see Example 8) we have

$$e^{-r} = \int_{\mathbb{R}} \sqrt{\frac{\sigma^2}{8\pi}}\, \exp\left(-\frac{\left(x-\frac{2}{\sigma^2}\right)\sigma^2}{8}\right) e^{-x/2}\,dx = e^{-\frac{1}{2\sigma^2}}.$$

Thus, the stability condition[20] reduces to

$$\lambda'(0)\rho'(1) < e^{\frac{1}{2\sigma^2}}.$$ •

Example 13 [BILC]: For the BILC (see Example 9) we have

$$\begin{aligned} e^{-r} &= \int_{\mathbb{R}} \left[\frac{1}{2}\,e^{-\frac{2}{\lambda}}\Delta_{-\frac{2}{\lambda}} + \frac{1}{4}\,e^{\frac{x-\frac{2}{\lambda}}{2}}\chi_{|x|\le\frac{2}{\lambda}} + \frac{1}{2}\,\Delta_{\frac{2}{\lambda}}\right] e^{-x/2}\,dx \\ &= e^{-\frac{1}{\lambda}}\,\frac{\lambda+1}{\lambda}. \end{aligned}$$

The stability condition for the BILC can, therefore, be written as

$$\lambda'(0)\rho'(1) < e^{\frac{1}{\lambda}}\,\frac{\lambda}{1+\lambda}.$$ •

The reader might have noticed that for all of the above examples e^{-r} is equal to the constant which appears in the Bhattacharyya bound. This is no coincidence. The constant r is simply the exponent in the Chernoff bound when we ask for the probability that the sum of n independent samples with common density P_0 is negative.

Proof of Theorem 5: Recall that in the case of the BEC we observed that zero was a fixed point of the recursion and by linearizing the recursion around this fixed point we were able to analyze its stability. For the general case we will proceed along the same lines. Recall further that we deal with distributions of log-likelihood ratios. From Corollary 1 we know that a zero probability of error corresponds to the density Δ_∞. Clearly, if $\mathrm{P}_\ell = \Delta_\infty$ for some $\ell \ge 0$ then $\mathrm{P}_{\ell+i} = \Delta_\infty$ for any $i \ge 0$, so that Δ_∞ is indeed a fixed point of density evolution. To analyze local convergence to this fixed point we shall again consider a linearization of density evolution about this fixed point.

To that effect, consider a density $Q_0 = 2\epsilon\Delta_0 + (1-2\epsilon)\Delta_\infty$. Note that this density is symmetric and that

$$\mathrm{P_e}(2\epsilon\Delta_0 + (1-2\epsilon)\Delta_\infty) = \epsilon.$$

After a complete iteration of density evolution this density will evolve to

$$Q_1 = 2\epsilon\lambda'(0)\rho'(1)\mathrm{P}_0 + (1-2\epsilon\lambda'(0)\rho'(1))\Delta_\infty + O(\epsilon^2).$$

More generally, if we consider n iterations of density evolution we see that the density Q_0 will evolve to

$$Q_n = 2\epsilon(\lambda'(0)\rho'(1))^n\mathrm{P}_0^{\otimes n} + (1-2\epsilon(\lambda'(0)\rho'(1))^n)\Delta_\infty + O(\epsilon^2).$$

We are interested in the error probability associated to Q_n, i.e., we are interested in $\mathrm{P_e}(Q_n)$. To this end note that if $\int_{\mathbb{R}} e^{sx} d(\int \mathrm{P}_0)(x) < \infty$ for all s in some neighborhood of zero then

$$r := -\lim_{n\to\infty} \frac{1}{n}\log \mathrm{P_e}(\mathrm{P}_0^{\otimes n}) \tag{13}$$

is well-defined, see [21]. Therefore, if we assume that $\lambda'(0)\rho'(1) > e^r$ then there exists an integer n such that

$$(\lambda'(0)\rho'(1))^n \mathrm{P_e}(\mathrm{P}_0^{\otimes n}) > 1.$$

It then follows that for this n

$$\begin{aligned} \mathrm{P_e}(Q_n) &= 2\epsilon(\lambda'(0)\rho'(1))^n \mathrm{P_e}(\mathrm{P}_0^{\otimes n}) + O(\epsilon^2) \\ &> 2\epsilon + O(\epsilon^2) \\ &> \epsilon, \qquad \text{if } \epsilon \le \xi \end{aligned}$$

where ξ is a positive constant depending only on (λ, ρ) and P_0. Now assume that for some iteration ℓ we have $\mathrm{P_e}(\mathrm{P}_\ell) =: \epsilon$, $\epsilon \le \xi$. We claim that then

$$\mathrm{P_e}(\mathrm{P}_{\ell+n}) \ge \mathrm{P_e}(Q_n) > 2\epsilon + O(\epsilon^2) > \epsilon$$

a contradiction, since as shown in [1], the error probability is a nonincreasing function in the number of iterations. This will show that if $\lambda'(0)\rho'(1) > e^r$ then $\mathrm{P_e}(\mathrm{P}_\ell) > \xi$ for some suitable positive constant ξ for all $\ell \in \mathbb{N}$.

To show that $\mathrm{P_e}(\mathrm{P}_{\ell+n}) \ge \mathrm{P_e}(Q_n)$ we argue as follows. Consider a random tree of depth $2n+1$ with variable nodes at the leaves and a variable node at the root. Let the leaf nodes have observations which correspond to samples from the density Q_0 and let the internal variable nodes have observations which correspond to samples from the density P_0. The density of the message emitted at the root node is then Q_n. Now compare this to the scenario where we use the same setup but where we let the leaf nodes take observations which correspond to samples from the density P_ℓ. In this case, the density of the message emitted at the root node will be $\mathrm{P}_{\ell+n}$. Note that in both cases the estimate of the root node message is a maximum-likelihood (ML) estimate. In Appendix B we show in Lemma 4 that one can think of the samples from the density P_ℓ as *physically degraded* samples from the density Q_0. We claim that this implies $\mathrm{P_e}(\mathrm{P}_{\ell+n}) \ge \mathrm{P_e}(Q_n)$. To see this, assume to the contrary that $\mathrm{P_e}(\mathrm{P}_{\ell+n}) < \mathrm{P_e}(Q_n)$. In words, we can improve our estimate emitted at the root nodes in the first scenario by adding noise to the observations corresponding to the leaf nodes and then applying an ML estimate to these new observations rather than by applying an ML estimate to the original observations. This contradicts the well-known fact that for a uniform prior ML estimates have minimum probability of error.

[18]Equivalently, we get

$$\epsilon^*(\lambda, \rho) \le \frac{1}{\lambda'(0)\rho'(1)}$$

which is nontrivial only if $\lambda'(0)\rho'(1) \ge 1$.

[19]Formulated as an upper bound, this gives

$$\delta^*(\lambda, \rho) \le \tfrac{1}{2}\left(1 - \sqrt{1 - \frac{1}{(\lambda'(0)\rho'(1))^2}}\right)$$

which is well-defined if $\lambda'(0)\rho'(1) \ge 1$.

[20]This gives rise to

$$\sigma^*(\lambda, \rho) \le \frac{1}{\sqrt{2\ln(\lambda'(0)\rho'(1))}}$$

which is well-defined if $\lambda'(0)\rho'(1) > 1$.

It remains to prove that

$$r := -\ln\left(\int_{\mathbb{R}} \mathrm{P}_0(x) e^{-\frac{x}{2}}\, dx\right)$$

as stated. A general large-deviation principle (see, e.g., [21]) implies that $e^r = \inf_{s<0} g(s)$, where $g(s) = \int_{\mathbb{R}} \mathrm{P}_0(x) e^{sx}\, dx$. Since P_0 is symmetric we have

$$\begin{aligned} e^{-r} &= \inf_{s<0} g(s) \\ &= \inf_{s<0} \int_{\mathbb{R}} \mathrm{P}_0(x) e^{sx}\, dx \\ &= \inf_{s<0} \tfrac{1}{2} \int_{\mathbb{R}} \mathrm{P}_0(x) e^{-x/2} \left[e^{-x(s+\frac{1}{2})} + e^{x(s+\frac{1}{2})}\right] dx \\ &= \int_{\mathbb{R}} \mathrm{P}_0(x) e^{-x/2}\, dx. \end{aligned}$$

□

We close this subsection by posing two fundamental open questions.

- Is it generally true that for any sequence of capacity-achieving degree distribution pairs the stability condition becomes tight as is the case for the BEC?
- Is it possible to formulate higher order stability conditions and to show that each of them becomes tight for any sequence of capacity-achieving degree distribution pairs as is again the case for the BEC, [17, Theorem 3.1].

F. Fixed Points of Density Evolution

The main result of this section states that density evolution for belief propagation always converges to a fixed point.

Consider again the example of the BEC and its associated density evolution recursion given in (10). In this case, we have the following complete characterization of the threshold in terms of fixed points of (10).

Theorem 6 [Fixed-Point Characterization of the Threshold for the BEC—[2]][21]: For a given degree distribution pair (λ, ρ) let $f(x, y) := y\lambda(1-\rho(1-x))$. For any $x_0 \in [0, 1]$ and $\ell \geq 1$ define $x_\ell(x_0) := f(x_{\ell-1}, x_0)$. Define

$$x_0^*(\lambda, \rho) := \sup\{0 \leq x_0 \leq 1 \colon x_\ell(x_0) \xrightarrow{\ell\to\infty} 0\}.$$

[Sufficiency] For any $x_0 \in [0, 1]$, $x_\ell(x_0)$ converges to a solution of $x = f(x, x_0)$ with $x \in [0, x_0]$. Therefore, if $x \neq f(x, x_0)$ for all $x \in (0, x_0]$, then $x_\ell(x_0)$ converges to zero as ℓ tends to infinity.[22]

[Necessity] If there exists an x, $x \in [0, x_0]$, such that $x = f(x, x_0)$, then $x_\ell(x_0) \geq x$ for all $\ell \geq 1$.[23]

[Fixed-Point Characterizations of the Threshold]

i) $x_0^*(\lambda, \rho) := \sup\{x_0 \in [0, 1] \colon x = f(x, x_0)$ has no solution x in $(0, x_0]\}$.
ii) $x_0^*(\lambda, \rho) := \inf\{x_0 \in [0, 1] \colon x = f(x, x_0)$ has a solution x in $(0, x_0]\}$.

Proof: First note that for any $x_0 \in [0, 1]$, $x_\ell(x_0)$ is nonincreasing and therefore converges to a point, call it x. Clearly, x fulfills $x = f(x, x_0)$ and $x \leq x_0$. Therefore, if no fixed point with $0 < x \leq x_0$ exists then $x_\ell(x_0)$ converges to zero as ℓ tends to infinity, which proves the first assertion. To prove the second assertion, note that if x is a fixed point with $x \leq x_0$ then

$$x_1(x_0) = f(x_0, x_0) \geq f(x, x_0) = x$$

where we have used the fact that $f(x, y)$ is a nondecreasing function in its first argument for $x, y \in [0, 1]$. By finite induction it then follows that

$$x_\ell(x_0) = f(x_{\ell-1}, x_0) \geq f(x, x_0) = x, \qquad \text{for all } \ell \geq 1.$$

It remains to discuss the characterization of the threshold in terms of fixed points. First note that, as remarked earlier, $x = 0$ is a fixed point of the recursion for any $0 \leq x_0 \leq 1$ and that $x = 1$ is a fixed point for $x_0 = 1$. It follows that both characterizations are well-defined. Finally, the equivalence of the two characterizations follows from the fact that $f(x, y)$ is *strictly* increasing in y for $0 < x \leq 1$ and $0 \leq y \leq 1$. □

At first it might appear that this behavior is special to the BEC case since this is the only (known) belief-propagation decoder for which density evolution has a one-dimensional description. So it is quite surprising that at least the sufficiency part has a complete analog in the general case.

In order to generalize the above results we must generalize various notions used in the proof. One crucial ingredient in the above argument is the *monotonicity* of $x_\ell(x_0)$. Since $x_\ell(x_0)$ is a sequence of *real numbers*, this monotonicity guarantees the convergence of $x_\ell(x_0)$ to some fixed point of the recursion. For the general case of density evolution, we have to assert the convergence of distributions. We will now show that there exists a large *family* of monotonicity conditions which we will later use to prove convergence.

Theorem 7: Let P_0 and g be symmetric densities on $(-\infty, +\infty]$ and for a given degree distribution pair (λ, ρ) define

$$\mathrm{P}_\ell := \mathrm{P}_0 \otimes \lambda(\Gamma^{-1}(\rho(\Gamma(\mathrm{P}_{\ell-1})))).$$

Then $\mathrm{P}_e(\mathrm{P}_\ell \otimes g)$ is a nonincreasing function of ℓ.

Proof: The message of which P_ℓ is the density represents a conditional probability of a particular bit value. Assume that an independent observation of the same bit value is available to the decoder and assume that this independent observation is obtained by passing the bit through a channel $p(\,\cdot|\cdot\,)$ which fulfills the symmetry condition and has $p(y|\boldsymbol{x} = 1) = g(y)$. By Lemma 3 and under the assumption that the all-one codeword was transmitted, the conditional density of the bit log-likelihood ratios, conditioned on all information incorporated in P_ℓ and the independent observation, has density $\mathrm{P}_\ell \otimes g$. Since the new density corresponds again to a maximum *a posteriori* (MAP) estimate conditioned on information which is nondecreasing in ℓ, the stated monotonicity condition follows. □

In the above theorem we can use *any* symmetric density for g. It will prove useful in the sequel to consider the family of densities

$$g_z(x) := \frac{1}{1+e^z}\,\Delta_{-z} + \frac{e^z}{1+e^z}\,\Delta_z. \tag{14}$$

[21]Although the exact statement differs from the one given in [15], the result is nevertheless an easy consequence of the statements given there.

[22]Hence, $x_0^*(\lambda, \rho) \geq x_0$.

[23]Therefore, if $x > 0$ then $x_0^*(\lambda, \rho) \leq x_0$.

Clearly, $g_z(x)$ is a symmetric density for any $0 \leq z \leq \infty$. If for any symmetric density f we define $\mathcal{P}_z(f)$ by

$$\mathcal{P}_z(f) := \mathrm{P_e}(f \otimes g_z)$$

then we immediately have the following.

Corollary 2: For a given symmetric density P_0 and a given degree distribution pair (λ, ρ) define

$$\mathrm{P}_\ell := \mathrm{P}_0 \otimes \lambda(\Gamma^{-1}(\rho(\Gamma(\mathrm{P}_{\ell-1})))).$$

Then $\mathcal{P}_z(\mathrm{P}_\ell)$ is a nonincreasing function of ℓ for every $0 \leq z \leq \infty$.

We will now show that a symmetric density f is uniquely determined by the "basis" of functionals $\{\mathcal{P}_z(f)\colon z \geq 0\}$.

Lemma 2 [Basic Lemma]: A symmetric density f is uniquely determined by $\{\mathcal{P}_z(f)\colon z \geq 0\}$.

Proof: Let $F \in \mathcal{F}$ be the distribution associated to f. Let z be a point of continuity of F. Then

$$\mathcal{P}_z(f) := \frac{1}{1+e^z} F(z) + \frac{e^z}{1+e^z} F(-z). \tag{15}$$

It is not too hard to check that $F(z)+e^z F(-z)$ is differentiable at z, and that the symmetry condition implies that the derivative is $e^z F(-z)$. In particular, $\frac{d}{dz}((1+e^z)\mathcal{P}_z(f)) = e^z F(-z)$ almost everywhere. Since $\mathcal{P}_0(f) = \frac{1}{2}(F(0)+F^-(0))$, we recover $F(0)$. Since F is symmetric, the proof is now complete. □

Finally, to generalize the above convergence results for the BEC we must provide a precise definition of the notion of "convergence" for general symmetric densities. In particular, convergence to Δ_∞ must be well-defined. Let $\mathcal{F}'$ denote the set of all right-continuous nondecreasing functions $F(x)$ with $\lim_{x\to-\infty} F(x) \geq 0$ and $\lim_{x\to\infty} F(x) \leq 1$. Note that $\mathcal{F} \subset \mathcal{F}'$. We say that a sequence $F_k \in \mathcal{F}'$, $k = 1, 2, \ldots$, converges to $F \in \mathcal{F}'$ if the sequence converges pointwise at all points of continuity of F. Convergence in this sense implies weak convergence of the associated Lebesgue–Stieltjes measure. That is, if $F_l \to F$ then $\int h(x)\, dF_l(x) \to \int h(x)\, dF(x)$ for any suitable function h. An important property of the space $\mathcal{F}'$ is that it is sequentially compact. Given any infinite sequence F_l from $\mathcal{F}'$, then, by the Helley selection principle, see [22, Theorem 25.9], there exists a subsequence which converges to some element of $\mathcal{F}'$.

Let $\mathcal{F}^s$ denote the subset of $\mathcal{F}'$ consisting of symmetric functions. Note that $\mathcal{F}^s$ is sequentially compact since, i.e., if $F \in \mathcal{F}'$ is a limit point of F_l, $l = 1, 2, \ldots$ with $F_\ell \in \mathcal{F}^s$ then $F \in \mathcal{F}^s$, since from the definition of symmetry given in Definition 1 we see that weak convergence immediately implies symmetry of the limit. We say that a sequence of symmetric densities P_ℓ converges to P if the corresponding distributions converge, i.e., if $F_\ell(x) := \int_{-\infty}^{x+} \mathrm{P}_\ell(x)\, dx$ converges to $F(x) := \int_{-\infty}^{x+} P(x)\, dx$.

Theorem 8 [Partial Fixed-Point Characterization of the Threshold for General Channels]: For a given degree distribution pair (λ, ρ) and for a given symmetric density P_0 define

$$\mathrm{P}_\ell := \mathrm{P}_0 \otimes \lambda(\Gamma^{-1}(\rho(\Gamma(\mathrm{P}_{\ell-1})))).$$

[Sufficiency] For any symmetric density P_0, the sequence of densities P_ℓ converges to a symmetric density P_∞ which is a fixed-point solution to (9). Therefore, if there does not exist a symmetric density P such that $0 < \mathcal{P}_z(\mathrm{P}) \leq \mathcal{P}_z(\mathrm{P}_0)$ for all $z \geq 0$ and such that P is a fixed point of (9), then $\mathrm{P_e}(\mathrm{P}_\ell)$ converges to zero as ℓ tends to infinity, or, equivalently, $\mathrm{P}_\infty = \Delta_\infty$.

Proof: Consider the sequence P_ℓ of message densities. By sequential compactness of $\mathcal{F}^s$, some subsequence converges to a symmetric limit density, call if P_∞. If the original sequence P_ℓ does not converge to P_∞ then there must exist another subsequence which converges to a distinct symmetric limit $\mathrm{P}'_\infty \neq \mathrm{P}_\infty$. It follows from the basis lemma that, for some z, $\mathcal{P}_z(\mathrm{P}'_\infty) \neq \mathcal{P}_z(\mathrm{P}_\infty)$. But this is a contradiction since $\mathcal{P}_z(\mathrm{P}_\ell)$ is a monotonic function for every z by Theorem 7 and cannot, therefore, possess two distinct limits. We now conclude that P_ℓ converges to P_∞.

The density P_∞ is a fixed point of (9) since the update equations are continuous under our notion of convergence. Furthermore, since the sequence $\mathcal{P}_z(\mathrm{P}_\ell)$ is monotonically nonincreasing for each $z \geq 0$, we have $\mathcal{P}_z(\mathrm{P}_\infty) \leq \mathcal{P}_z(\mathrm{P}_0)$. If $\mathcal{P}_z(P) = 0$ for any $z > 0$ and if P is a symmetric density then $P = \Delta_\infty$. We conclude that if there does not exist a fixed-point P of (9) satisfying $0 < \mathcal{P}_z(\mathrm{P}) \leq \mathcal{P}_z(\mathrm{P}_0)$ then $\mathrm{P}_\infty = \Delta_\infty$. □

IV. Optimization

In this section, we briefly describe the optimization techniques that we used to obtain degree distribution pairs with large thresholds.

The following general remarks apply to any numerical optimization technique. First, formally, the threshold is defined as the supremum of all channel parameters for which the probability of error under density evolution converges to zero. By Corollary 1, this is equivalent to requiring that the message distribution converges to Δ_∞. In practice, we can verify at best that the probability of error reaches a value below a prescribed ϵ. From Theorem 5 we know that if we choose ϵ small enough then this automatically implies convergence to zero probability of error. In practice, the issue of convergence is not of great concern since we always allow a finite (but small) probability of error.

Secondly, in order to perform the computations we need to quantize the quantities involved. This quantization leads to a quantization error and this error might accumulate over the course of many iterations, rendering the computations useless. This problem can be circumvented in the following way. By carefully performing the quantization one can ensure that the quantized density evolution corresponds to the exact density evolution of a quantized message-passing scheme. Since belief propagation is optimal, such a quantized version is suboptimal and, hence, the reported thresholds can be thought of as lower bounds on the actual thresholds.

A. Local Optimization

To find good degree distribution pairs we started with the following simple hill-climbing approach. Fix a small target error probability ϵ and a maximum number of iterations m. Start with a given degree distribution pair and determine the maximum *admissible* channel parameter, i.e., the maximum channel parameter such that the error probability after m iterations is below ϵ. Now apply a small change to the degree distribution pair and

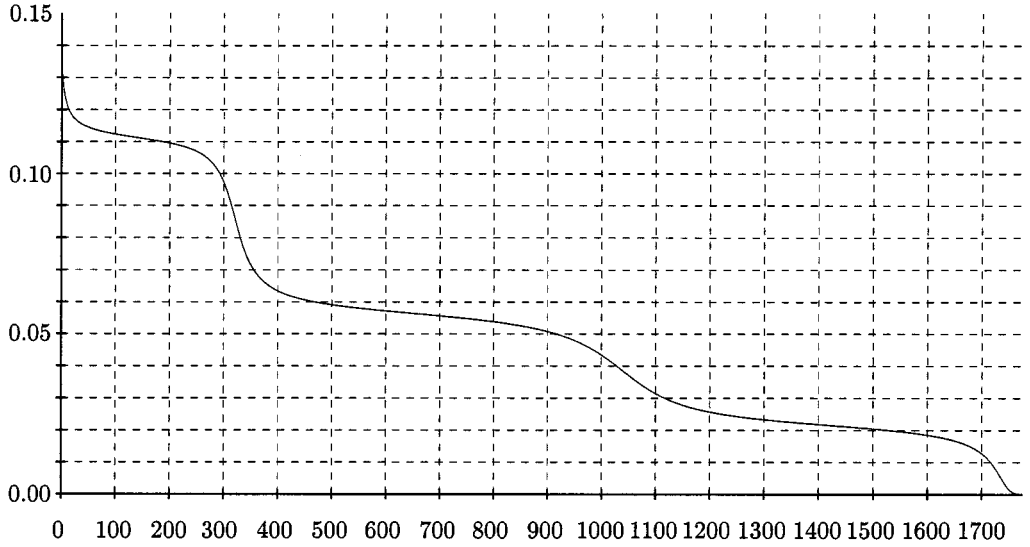

Fig. 4. Evolution of the bit-error probability under density evolution as a function of the iteration number.

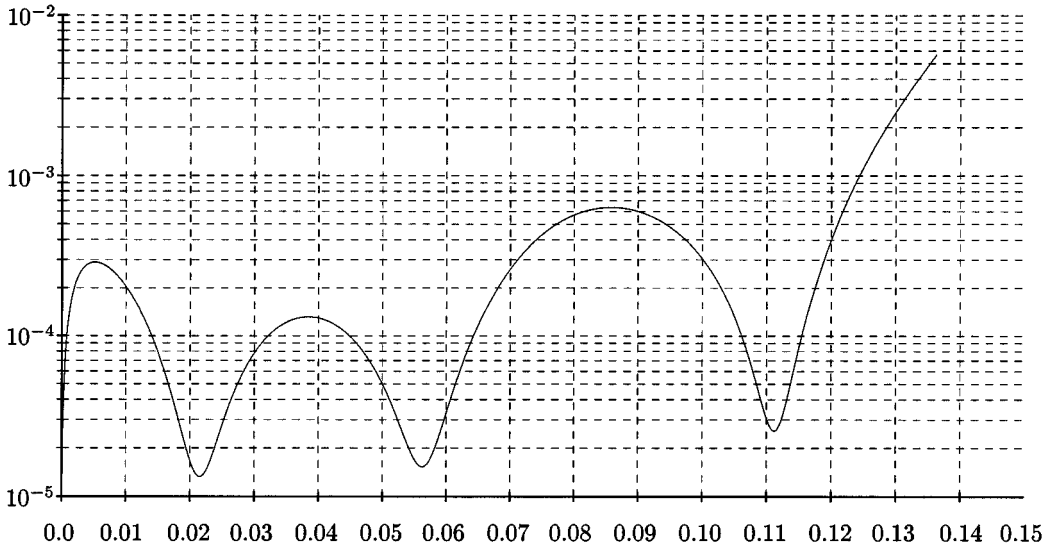

Fig. 5. The decrease of the bit-error probability as a function of the current bit-error probability.

check if it has either a larger admissible channel parameter or at least a smaller target error probability after m iterations. If so, declare the new degree distribution pair to be the currently best degree distribution pair, otherwise keep the original degree distribution pair. The same basic step is then repeated a large number of times.

The search for good degree distribution pairs can be substantially accelerated by appropriately limiting the search space. We found, for example, that very good degree distribution pairs exist with only a few nonzero terms. In particular, it suffices to allow two or three nonzero check node degrees (and these degrees can be chosen consecutively) and to limit the nonzero variable node degrees to $2, 3$, the maximum such degree d_l, and, possibly, a few well-chosen degrees in-between.

A further substantial savings in running time can be achieved as follows. For a particular degree distribution pair (λ, ρ), Fig. 4 shows the evolution of the bit-error probability as a function of the iteration number. An even clearer picture emerges if we plot the *decrease* of the bit-error probability as a function of the current bit error probability p_b. This is shown in Fig. 5.

As can be seen in these figures, after an initial swift decrease in the bit-error probability the procedure almost comes to a halt at $p_b = 0.111$ with decreases in bit-error probability of only 2.6×10^{-5} per iteration. The convergence then speeds up again until it hits another low at $p_b = 0.056$ and then later again at $p_b = 0.022$. At these three *critical points*, the outgoing message distribution is *almost* a fixed point of the equation system corresponding to one iteration. Indeed, if the parameter σ were

slightly increased then the iteration could not overcome any of those points and one can verify that there arise corresponding fixed points of density evolution.

Provided that the fixed points are stable, the message distributions at these points are *continuous* functions of the degree distribution pair. Hence, a small change in the degree distribution pair causes only small changes in the associated fixed-point distributions. Furthermore, if the fixed points are stable, then this affords a certain memorylessness to the density evolution process because they serve as local attractors. Small perturbations to the path will not matter once the domain of convergence of the fixed point is entered and, once the fixed point is found, the path that leads to it is irrelevant.

In practice, we observe that the points at which the density evolution gets stuck are indeed stable fixed points. The fixed point theorem in Section III-F shows that fixed points which are limits of density evolution must be at least marginally stable. The above considerations suggest the following scheme. Assume we determine the critical points (near fixed points, or likely fixed points for a slightly worse initial distribution) for a particular degree distribution pair and we would like to determine the merit of a particular small change of the degree distribution pair. Rather than starting with the initial distribution and then checking if (and how fast) this initial distribution converges to Δ_∞, one can memorize the distributions at the critical points of the original degree distribution pair and then determine how the proposed change affects the speed of convergence locally at these points. Once a promising change has been found, the merit of this change can be verified by starting with the initial degree distribution pair. Typically, only a few iterations are necessary at each critical point to determine if the change of the degree distribution pair improves the convergence or not. This has to be compared to hundreds of iterations or even thousands of iterations which are necessary if one starts with the initial distribution.

In the optimization scheme we just described we made use of the distributions at the "critical points" to find promising changes of the degree distribution pair. The following schemes extend this idea even further; the resulting algorithms are reminiscent of the algorithms used in the BEC case. For simplicity, we will only describe the optimization of the variable node degree distribution. The extension to the check node degree distribution and to joint optimization should be quite apparent.

Assume that we are given a degree distribution pair (λ, ρ), a particular channel parameter σ, and a target probability of error ϵ. Let $\{p_\ell\}_{\ell=0}^m$ be the sequence of error probabilities of the belief-propagation algorithm. More precisely, p_0 is the initial error probability, p_ℓ is the probability of error after the ℓth iteration, and $p_m \leq \epsilon < p_{m-1}$. Assume that we want to find a new degree distribution $\tilde{\lambda}$ which achieves the target probability of error in fewer iterations or achieves a lower target in the same number of iterations.

Define a matrix $A_{\ell,j}, 1 \leq \ell \leq m, 2 \leq j \leq d_r$. The entry $A_{\ell,j}$ is the error probability which results if we run the belief-propagation decoder for $(\ell-1)$ steps assuming that the variable node degree distribution is λ followed by one step in which we assume that the variable node degree distribution is a singleton with all its mass on the degree j. Note that the actual error probability after the ℓth iteration, p_ℓ, can be expressed in terms of $A_{\ell,j}$ as

$$p_\ell = \sum_{j=2}^{d_l} A_{\ell,j}\lambda_j.$$

Let us define a function $p(t)$ for $t \in [0, m]$ by linearly interpolating the p_ℓ, setting $p(\ell) = p_\ell$. Define

$$L(\lambda) := \int_{p_m}^{p_0} \left(-\frac{dp}{dt}(x)\right)^{-1} dx.$$

We interpret L as the number of iterations required to take the initial probability of error p_0 down to p_m. Using the expression above, we can write down the gradient of $L(\lambda)$ with respect to λ. In particular, for a perturbation h we can compute

$$D_h L(\lambda) := \frac{d}{d\eta} L(\lambda + \eta h)|_{\eta=0}$$

as

$$D_h L = \int_{p_m}^{p_0} \left(\frac{dp}{dt}(x)\right)^{-2} D_h\left(\frac{dp}{dt}(x)\right) dx.$$

Returning to the discrete representation this is equivalent to

$$D_h L = \sum_{j=2}^{d_l} h_j \left(\sum_{\ell=1}^{m} \frac{A_{\ell,j} - p_\ell}{p_{\ell-1} - p_\ell}\right).$$

Thus, we observe that the gradient of $L(\lambda)$ is given by

$$\frac{d}{d\lambda_j} L(\lambda) = \sum_{\ell=1}^{m} \frac{A_{\ell,j} - p_\ell}{p_{\ell-1} - p_\ell}.$$

There are two ways we can exploit this expression. One is to use the (negative) gradient direction to do hill climbing, and the other is to globally optimize the linearized approximation of L. In either case, we must incorporate the constraints on λ.

Let $\tilde{\lambda}$ be an alternative degree distribution. Clearly, $\tilde{\lambda}$ has to be a probability mass function, i.e.

$$\sum_{j=2}^{d_l} \tilde{\lambda}_j = 1 \tag{16}$$

and, further, it has to correspond to a code of equal rate, i.e.

$$\sum_{j=2}^{d_l} \frac{\tilde{\lambda}_j}{j} = \sum_{j=2}^{d_l} \frac{\lambda_j}{j}. \tag{17}$$

Let h be the negative gradient direction of L. If we set $\tilde{\lambda} = \lambda + \eta h$ (for positive η) then the above constraints may not be satisfied. However, among degree distributions satisfying the constraints the one closest to $\lambda + \eta h$ in Euclidean distance can be easily computed by alternating projections. Two projections are required: the first is orthogonal projection of h onto the subspace determined by $\sum_j h_j = 0$ (total probability constraint) and $\sum_j \frac{1}{j} h_j = 0$ (rate constraint), and the second projection sets $\eta h_j = -\lambda_j$ if, prior to the projection, $\eta h_j + \lambda_j < 0$. Note that an alternative interpretation is to project the gradient direction h onto the convex polytope of admissible directions. One

can then compute the maximum step size η for which the constraints remain satisfied and then recompute the projection at that point. In this way, one can easily walk along the projected gradient direction to look for an improved degree distribution.

Let us now consider the second way to exploit the gradient expression for L. Let

$$\tilde{p}_\ell := \sum_{j=2}^{d_l} A_{\ell,j}\tilde{\lambda}_j.$$

Then we have

$$L(\tilde{\lambda}) \simeq \sum_{\ell=1}^{m} \frac{\tilde{p}_\ell - p_\ell}{p_{\ell-1} - p_\ell}. \tag{18}$$

This approximation is valid as long as $\tilde{\lambda}$ does not differ too much from λ, i.e., assuming that the message distributions corresponding to λ and $\tilde{\lambda}$ are not too different, if

$$\max_\ell \frac{|p_\ell - \tilde{p}_\ell|}{p_{\ell-1} - p_\ell} < \delta \tag{19}$$

where $\delta \ll 1$, and if

$$\tilde{p}_\ell < p_{\ell-1}, \qquad 1 \le \ell \le m. \tag{20}$$

Recall that we want to minimize $L(\tilde{\lambda})$. Since the right-hand side of (18) is (up to a constant) a linear function in the degree distribution and since the constraints stated in (16), (17), (19), and (20) are also linear, this can be (approximately) accomplished by means of a linear program. The same procedure is then applied repeatedly in an attempt to converge to a good degree distribution. Since both approaches are local optimizations it is appropriate to repeat the optimization with various initial conditions.

B. Global Optimization

The code design problem as described above belongs to the class of nonlinear constraint satisfaction problems with continuous space parameters. Many general algorithms for solving such problems have been developed. We experimented with an algorithm called Differential Evolution (DE) [23] that has already been successfully applied to the design of good erasure codes [11]. DE is a robust optimizer for multivariate functions. We will not describe the details here, suffice it to say that the algorithm is in part a hill climbing algorithm and in part a genetic algorithm.

Our goal is to maximize the cost function which we define to be the threshold value for the channel. Since such optimizers, and DE in particular, operate best in a continuous parameter space of not too large dimension, and since frequent function evaluations are required in the optimization, we found it convenient to let the parameter space be a continuous space of small dimension. To accomplish this, we introduced *fractional phantom distributions*. Let the polynomials λ and ρ take on the general form $\sum_i \lambda_i x^{i-1}$ (similarly for ρ), where now both the λ_i *and the degree* i could take any positive real value. The real degree distribution is obtained from this phantom distribution as $\sum_i(\lambda_{i1}x^{\lfloor i\rfloor-1} + \lambda_{i2}x^{\lceil i\rceil-1})$, where λ_{i1} and λ_{i2} are uniquely determined via the equations

$$\lambda_{i1} + \lambda_{i2} = \lambda_i, \quad \text{and} \quad \frac{\lambda_{i1}}{\lfloor i\rfloor} + \frac{\lambda_{i2}}{\lceil i\rceil} = \frac{\lambda_i}{i}.$$

This way, we are guaranteed to obtain a degree distribution which respects the rate-constraints for the code.

By allowing fractional degrees we, in effect, force the program to choose (close to) optimal degrees. This results in a significant reduction of the dimensionality of the parameter space, hence the running time, and also in the sparsity of the degree distributions obtained.

Appendix A
Channel Equivalence Lemma

We say that two binary-input memoryless output-symmetric channels are *equivalent* if they have the same density of log-likelihood ratios. It is often convenient to pick one representative from each equivalence class. This can be done as shown in the following lemma.

Lemma 3 [Channel Equivalence Lemma]: Let $\mathrm{P}(y)$ be a symmetric density. The binary-input memoryless output-symmetric channel $p(\cdot|\cdot)$ with $p(y|\boldsymbol{x}=1) = \mathrm{P}(y)$ (and, hence, by symmetry $p(y|\boldsymbol{x}=-1) = \mathrm{P}(-y)$) has an associated density of log-likelihood ratios equal to $\mathrm{P}(y)$.

Proof:

$$\log\frac{p(y|\boldsymbol{x}=1)}{p(y|\boldsymbol{x}=-1)} = \log\frac{p(y|\boldsymbol{x}=1)}{p(-y|\boldsymbol{x}=1)} = \log\frac{\mathrm{P}(y)}{\mathrm{P}(-y)} = \log\frac{\mathrm{P}(y)}{\mathrm{P}(y)e^{-y}} = y. \qquad \square$$

Appendix B
The Erasure Decomposition Lemma

Lemma 4 [Erasure Decomposition Lemma]: Let p be a binary-input memoryless output-symmetric channel. Let P denote its associated distribution of log-likelihood ratios $\log\frac{p(y|x=1)}{p(y|x=-1)}$. Then the channel p can be represented as the concatenation of an erasure channel with erasure probability $2\mathrm{P_e}(P)$ and a ternary-input memoryless output-symmetric channel q, i.e., p is a physical degraded version of an erasure channel.

Proof: Recall from Lemma 3 that, without loss of generality, we may assume that $p(y|\boldsymbol{x}=1) = P(y)$ and hence, by symmetry, $p(y|\boldsymbol{x}=-1) = P(-y)$. Let ε denote an erasure and let 1, -1 denote bit values. Let q denote a channel whose input alphabet is $\{-1, \varepsilon, 1\}$. Further, let q have real output y and set

$$q(y|\boldsymbol{x}=\varepsilon) = \frac{1}{2\epsilon}e^{-|y|}P(|y|)$$

$$q(y|\boldsymbol{x}=1) = \frac{1}{1-2\epsilon}(1-e^{-y})\chi_{\{y\ge 0\}}P(y)$$

and

$$q(y|\boldsymbol{x}=-1) = \frac{1}{1-2\epsilon}(1-e^{y})\chi_{\{y\le 0\}})P(-y).$$

It is easy to check that these quantities are well-defined densities. Let $\tilde{q}$ denote the concatenation of the erasure channel with the channel q, i.e., the output of the erasure channel is fed into q.

Note that for equally likely inputs the output of the erasure channel, and hence the input to q, have probabilities $(1-2\epsilon)/2$, 2ϵ, and $(1-2\epsilon)/2$, respectively. Then we have

$$\begin{aligned}\tilde{q}(y|\boldsymbol{x}=1) &= 2\epsilon\frac{1}{2\epsilon}\,e^{-|y|}P(|y|) \\ &\quad + (1-2\epsilon)\frac{1}{1-2\epsilon}\,(1-e^{-y})\chi_{\{y\geq 0\}}P(y) \\ &= P(y) = p(y|\boldsymbol{x}=1).\end{aligned}$$

and

$$\begin{aligned}\tilde{q}(y|\boldsymbol{x}=-1) &= 2\epsilon\frac{1}{2\epsilon}\,e^{-|y|}P(|y|) \\ &\quad + (1-2\epsilon)\frac{1}{1-2\epsilon}\,(1-e^{-y})\chi_{\{y\leq 0\}}P(-y) \\ &= P(-y) = p(y|\boldsymbol{x}=-1).\end{aligned}$$

□

ACKNOWLEDGMENT

The authors wish to thank Sae-Young Chung and Igal Sason as well as the reviewers for their detailed comments and many helpful suggestions.

REFERENCES

[1] T. Richardson and R. Urbanke, "The capacity of low-density parity-check codes under message-passing decoding," *IEEE Trans. Inform. Theory*, vol. 47, pp. 599–618, Feb. 2000.

[2] M. Luby, M. Mitzenmacher, A. Shokrollahi, D. Spielman, and V. Stemann, "Practical loss-resilient codes," in *Proc. 29th Annu. ACM Symp. Theory of Computing*, 1997, pp. 150–159.

[3] D. Spielman, "Linear-time encodeable and decodable error-correcting codes," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1723–1731, Nov. 1996.

[4] T. Richardson and R. Urbanke, "Efficient encoding of low-density parity-check codes," *IEEE Trans. Inform. Theory*, vol. 47, pp. 638–656, Feb. 2000.

[5] M. Tanner, "A recursive approach to low complexity codes," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 533–547, Sept. 1981.

[6] M. Luby, M. Mitzenmacher, A. Shokrollahi, and D. Spielman, "Analysis of low density codes and improved designs using irregular graphs," in *Proc. 30th Annu. ACM Symp. Theory of Computing*, 1998, pp. 249–258.

[7] ——, "Improved low-density parity-check codes using irregular graphs and belief propagation," in *Proc. 1998 IEEE Int. Symp. Information Theory*, Cambridge, MA, Aug. 1998, p. 117.

[8] D. MacKay, S. Wilson, and M. Davey, "Comparison of constructions of irregular Gallager codes," *IEEE Trans. Commun.*, vol. 47, pp. 1449–1454, Oct. 1999.

[9] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[10] M. Luby, M. Mitzenmacher, and A. Shokrollahi, "Analysis of random processes via and-or tree evaluation," in *Proc. 9th Annu. ACM-SIAM Symp. Discrete Algorithms*, 1998, pp. 364–373.

[11] A. Shokrollahi and R. Storn, "Design of efficient erasure codes with differential evolution," in *Proc. Int. Symp. Information Theory*, Sorrento, Italy, June 2000.

[12] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding," in *Proc. Int. Communications Conf. (ICC'93)*, Geneve, Switzerland, May 1993, pp. 1064–1070.

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

[14] D. J. C. MacKay, "Good error correcting codes based on very sparse matrices," *IEEE Trans. Information Theory*, vol. 45, pp. 399–431, Mar. 1999.

[15] A. Shokrollahi, "New sequences of linear time erasure codes approaching the channel capacity," in *Proc. 13th Conf. Applied Algebra, Error Correcting Codes, and Cryptography (Lecture Notes in Computer Science)*. Berlin, Germany: Springer Verlag, 1999, pp. 65–76.

[16] S.-Y. Chung, J. G. D. Forney, T. Richardson, and R. Urbanke, "On the design of low-density parity-check codes within 0.0045 dB of the Shannon limit," *IEEE Commun. Lett.*, vol. 5, pp. 58–60, Feb. 2001.

[17] A. Shokrollahi, "Capacity-achieving sequences," in *Codes, Systems, and Graphical Models (IMA Volumes in Mathematics and Its Applications)*, B. Marcus and J. Rosenthal, Eds. New York: Springer-Verlag, 2000, vol. 123, pp. 153–166.

[18] L. Decreusefond and G. Zémor, "On the error-correcting capabilities of cycle codes of graphs," *Comb., Probab., Computing*, no. 6, pp. 27–38, 1997.

[19] G. Horn, "Iterative decoding and pseudocodewords," Ph.D. dissertation, Dept. Elect. Eng., Calif. Inst. Technol., Pasedena, CA, May 1999.

[20] T. Richardson and R. Urbanke, "A proof of the stability condition for LDPC codes," paper, in preparation.

[21] A. Shwartz and A. Weiss, *Large Deviations for Performance Analysis*. London, U.K.: Chapman & Hall, 1995.

[22] P. Billingsley, *Probability and Measure*, 3rd ed. New York: Wiley, 1995.

[23] K. Price and R. Storn, "Differential evolution—A simple and efficient heuristic for global optimization over continuous spaces," *J. Global Optimiz.*, vol. 11, pp. 341–359, 1997.

[24] M. Luby, M. Mitzenmacher, A. Shokrollahi, D. Spielman, and V. Stemann, "Practical loss-resilient codes," *IEEE Trans. Inform. Theory*, vol. 47, pp. 569–584, Feb. 2001.