The price of certainty: "waterslide curves" and the gap to capacity

Anant Sahai
Pulkit Grover

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2008-1
http://www.eecs.berkeley.edu/Pubs/TechRpts/2008/EECS-2008-1.html

January 1, 2008

Acknowledgement

Years of conversations with colleagues in the Berkeley Wireless Research Center have helped motivate this investigation and informed the perspective here. Cheng Chang participated in the research discussions, especially as regards the AWGN channel. Sae-Young Chung gave valuable feedback at an early stage of this research and Hari Palaiyanur caught many typos in early drafts of this manuscript. Funding support from NSF CCF 0729122, NSF ITR 0326503, NSF CNS 0403427, and gifts from Sumitomo Electric.

The price of certainty: “waterslide curves” and the gap to capacity

Anant Sahai and Pulkit Grover
Wireless Foundations, Department of EECS
University of California at Berkeley, CA-94720, USA
{sahai, pulkit}@eecs.berkeley.edu

Abstract

The classical problem of reliable point-to-point digital communication is to achieve a low probability of error while keeping the rate high and the **total power consumption** small. Traditional information-theoretic analysis uses explicit models for the communication channel to study the power spent in transmission. The resulting bounds are expressed using ‘waterfall’ curves that convey the revolutionary idea that unboundedly low probabilities of bit-error are attainable using only finite transmit power. However, practitioners have long observed that the decoder complexity, and hence the total power consumption, goes up when attempting to use sophisticated codes that operate close to the waterfall curve.

This paper gives an explicit model for power consumption at an idealized decoder that allows for extreme parallelism in implementation. The decoder architecture is in the spirit of message passing and iterative decoding for sparse-graph codes, but is further idealized in that it allows for more computational power than is currently known to be implementable. Generalized sphere-packing arguments are used to derive lower bounds on the decoding power needed for any possible code given only the gap from the Shannon limit and the desired probability of error. As the gap goes to zero, the *energy per bit* spent in decoding is shown to go to infinity. This suggests that to optimize total power, the transmitter should operate at a power that is strictly above the minimum demanded by the Shannon capacity.

The lower bound is plotted to show an unavoidable tradeoff between the average bit-error probability and the total power used in transmission and decoding. In the spirit of conventional waterfall curves, we call these ‘waterslide’ curves. The bound is shown to be order optimal by showing the existence of codes that can achieve similarly shaped waterslide curves under the proposed idealized model of decoding.

The price of certainty: “waterslide curves” and the gap to capacity

Note: A preliminary version of this work with weaker bounds was submitted to ITW 2008 in Porto [1].

I. Introduction

As digital circuit technology advances and we pass into the era of billion-transistor chips, it is clear that the fundamental limit on practical codes is not any nebulous sense of “complexity” but the concrete issue of power consumption. At the same time, the proposed applications for error-correcting codes continue to shrink in the distances involved. Whereas earlier “deep space communication” helped stimulate the development of information and coding theory [2], [3], there is now an increasing interest in communication over much shorter distances ranging from a few meters [4] to even a few millimeters in the case of inter-chip and on-chip communication [5].

The implications of power-consumption beyond transmit power have begun to be studied by the community. The common thread in [6]–[10] is that the power consumed in processing the signals can be a substantial fraction of the total power. In [11], it is observed that within communication networks, it is worth developing cross-layer schemes to reduce the time that devices spend being active. In [9], an information-theoretic formulation is considered. When the transmitter is in the ‘on’ state, its circuit is modeled as consuming some fixed power in addition to the power radiated in the transmission itself. Therefore, it makes sense to shorten the overall duration of a packet transmission and to satisfy an average transmit-power constraint by bursty signalling that does not use all available degrees of freedom. In [7], the authors take into account a peak-power constraint as well, as they study the optimal constellation size for uncoded transmission. A large constellation requires a smaller ‘on’ time, and hence less circuit power. However, a larger constellation requires higher power to maintain the same spacing of constellation points. An optimal constellation has to balance between the two, but overall this argues for the use of higher rates. However, none of these really tackle the role of the decoding complexity itself.

In [12], the authors take a more receiver-centric view and focus on how to limit the power spent in sampling the signal at the receiver. They point out that empirically for ultrawideband systems aiming for moderate probabilities of error, this sampling cost can be larger than the decoding cost! They introduce the ingenious idea of adaptively puncturing the code at the receiver rather than at the transmitter. They implicitly argue for the use of longer codes whose rates are further from the Shannon capacity so that the decoder has the flexibility to adaptively puncture as needed and thereby save on total power consumption.

In [4], the authors study the impact of decoding complexity using the metric of coding gain. They take an empirical point of view using power-consumption numbers for certain decoder implementations at moderately low probabilities of error. They observe that it is often better to use no coding at all if the communication range is low enough.

In this paper, we take an asymptotic approach to see if considering decoding power has any fundamental implications as the average probability of bit error tends to zero. In Section II, we give an asymptotic formulation of what it should mean to approach capacity when we must consider the power spent in decoding in addition to that spent in transmission. We next consider whether classical approaches to encoding/decoding such as dense linear block codes and convolutional codes can satisfy our stricter standard of approaching capacity and argue that they cannot. Section III then focuses our attention on iterative decoding by message passing and defines the system model for the rest of the paper.

Section IV derives general lower bounds to the complexity of iterative decoders for BSC and AWGN channels in terms of the number of iterations required to achieve a desired probability of error at a given transmit power. These bounds can be considered iterative-decoding counterparts to the classical sphere-packing bounds (see e.g. [13], [14]) and are derived by generalizing the delay-oriented arguments of [15], [16] to the decoding neighborhoods in iterative decoding. These bounds are then used to show that it is in principle possible for iterative decoders to be a part of a weakly capacity-achieving communication system. However, the power spent by our model of an iterative decoder must go to infinity as the probability of error tends to zero and so this style of decoding rules out a strong sense of capacity-achieving communication systems.

We discuss related work in the sparse-graph-code context in Section V and make precise the notion of gap to capacity before evaluating our lower-bounds on the number of iterations as the gap to capacity closes. We conclude in Section VI with some speculation and point out some interesting questions for future investigation.

II. Certainty-achieving codes

Consider a classical point-to-point AWGN channel with no fading. For uncoded transmission with BPSK signaling, the probability of bit-error is an exponentially decreasing function of the transmitted energy per symbol. To approach certainty (make the probability of bit-error very small), the transmitted energy per symbol must go to infinity. If the symbols each carry a small number of bits, then this implies that the transmit *power* is also going to infinity since the number of symbols per second is a nonzero constant determined by the desired rate of R bits per second.

Shannon's genius in [17] was to recognize that while there was no way to avoid having the transmitted *energy* go to infinity and still approach certainty, this energy could be amortized over many bits of information. This meant that the transmitted *power* could be kept finite and certainty could be approached by paying for it using end-to-end delay (see [16] for a review) and whatever implementation complexity is required for the encoding and decoding. For a given channel and transmit power P_T, there is a maximum rate $C(P_T)$ that can be supported. Turned around, this classical result is traditionally expressed by fixing the desired rate R and looking at the required transmit power. The resulting "waterfall curves" are shown[1] in Figure 1. These sharp curves are distinguished from the more gradual "waterslide curves" of uncoded transmission.

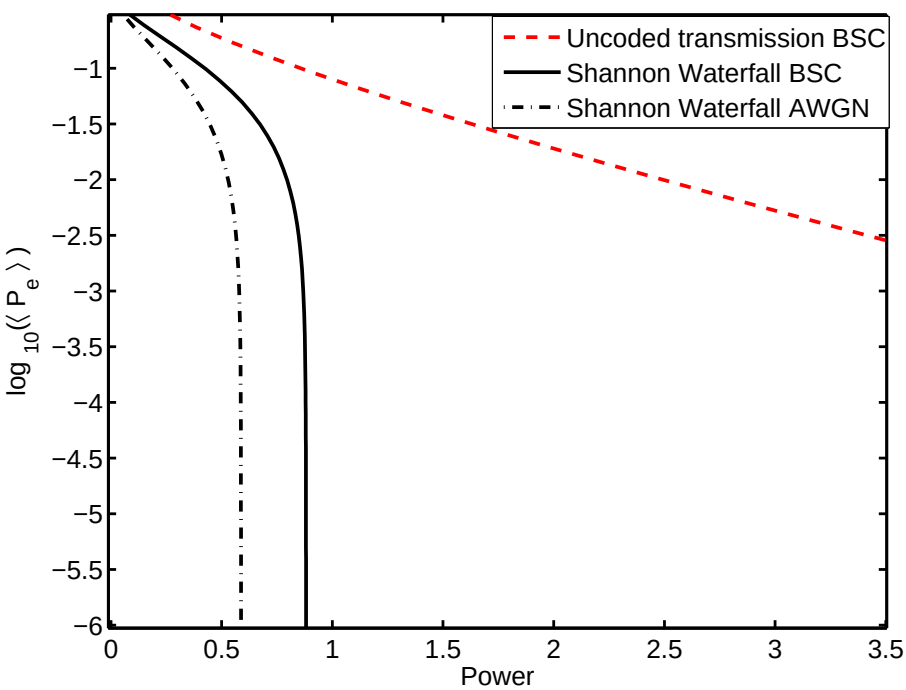

Fig. 1. The Shannon waterfalls: plots of $\log(\langle P_e \rangle)$ vs required SNR (in dB) for a fixed rate-$1/3$ code transmitted using BPSK over an AWGN channel with hard decisions at the detector. A comparison is made with the rate-$1/3$ repetition code: uncoded transmission with the same bit repeated three times. Also shown is the waterfall curve for the average power constrained AWGN channel.

Traditionally, a family of codes was considered capacity achieving if it could support arbitrarily low probabilities of error at transmit powers arbitrarily close to that predicted by capacity. The complexity of the encoding and decoding steps was considered to be a separate and qualitatively distinct performance metric. This makes sense

[1]Since the focus of this paper is on average bit error probability, these curves combine the results of [17], [18] and adjust the required capacity by a factor of the relevant rate-distortion function $1 - h_b(\langle P_e \rangle)$.

when the communication is long-range, since the "exchange rate" between transmitter power and the power that ends up being delivered to the receiver is very poor due to distance-induced attenuation.

In light of the advances in digital circuits and the need for shorter-range communication, we propose a new way of formalizing what it means for a coding approach to be "capacity achieving" using the single natural metric: power.

A. Definitions

Assume the traditional information-theoretic model (see e.g. [13], [19]) of fixed-rate discrete-time communication with k total information bits, m channel uses, and the rate of $R = \frac{k}{m}$ bits per channel use. As is traditional, the rate R is held constant while k and m are allowed to become asymptotically large. $\langle P_{e,i} \rangle$ is the average probability of bit error on the i-th message bit and $\langle P_e \rangle = \frac{1}{k} \sum_i \langle P_{e,i} \rangle$ is used to denote the overall average probability of bit error. No restrictions are assumed on the codebooks aside from those required by the channel model. The channel model is assumed to be indexed by the power used in transmission. The encoder and decoder are assumed to be physical entities that consume power according to some model that can be different for different codes.

Let $\xi_T P_T$ be the actual power used in transmission and let P_C and P_D be the power consumed in the operation of the encoder and decoder respectively. ξ_T is the exchange rate (total path-loss) that connects the power spent at the transmitter to the received power P_T that shows up at the receiver. In the spirit of [10], we assume that the goal of the system designer is to minimize some weighted combination $P_{total} = \xi_T P_T + \xi_C P_C + \xi_D P_D$ where the vector $\vec{\xi} > 0$. The weights can be different depending on the application[2] and ξ_T is tied to the distance between the transmitter and receiver as well as the propagation environment.

For any rate R and average probability of bit error $\langle P_e \rangle > 0$, we assume that the system designer will minimize the weighted combination above to get optimized $P_{total}(\vec{\xi}, \langle P_e \rangle, R)$ as well as constituent $P_T(\vec{\xi}, \langle P_e \rangle, R), P_C(\vec{\xi}, \langle P_e \rangle, R)$, and $P_D(\vec{\xi}, \langle P_e \rangle, R)$.

Definition 1: The *certainty* of a particular encoding and decoding system is the reciprocal of the average probability of bit error.

Definition 2: An encoding and decoding system at rate R bits per second is *weakly certainty achieving* if $\liminf_{\langle P_e \rangle \to 0} P_T(\vec{\xi}, \langle P_e \rangle, R) < \infty$ for all weights $\vec{\xi} > 0$.

If an encoder/decoder system is not weakly certainty achieving, then this means that it does not deliver on the revolutionary promise of the Shannon waterfall curve from the perspective of transmit power. Instead, such codes encourage system designers to pay for certainty using unbounded transmission power.

Definition 3: An encoding and decoding system at rate R bits per second is *strongly certainty achieving* if $\liminf_{\langle P_e \rangle \to 0} P_{total}(\vec{\xi}, \langle P_e \rangle, R) \neq \infty$ for all weights $\vec{\xi} > 0$.

A strongly certainty-achieving system would deliver on the full spirit of Shannon's vision: that certainty can be approached at finite total power just by accepting longer end-to-end delays and amortizing the total energy expenditure over many bits. The general distinction between strong and weak certainty-achieving systems relates to how the decoding power $P_D(\vec{\xi}, \langle P_e \rangle, R)$ varies with the probability of bit-error $\langle P_e \rangle$ for a fixed rate R. Does it have waterfall or waterslide behavior? For example, it is clear that uncoded transmission has very simple encoding/decoding[3] and so $P_D(\vec{\xi}, \langle P_e \rangle, R)$ has a waterfall behavior.

Definition 4: A {weakly|strongly} certainty-achieving system at rate R bits per second is also {*weakly*|*strongly*} *capacity achieving* if

$$\liminf_{\xi_C, \xi_D \to \vec{0}} \liminf_{\langle P_e \rangle \to 0} P_T(\vec{\xi}, \langle P_e \rangle, R) = C^{-1}(R) \tag{1}$$

where $C^{-1}(R)$ is the minimum transmission power that is predicted by the Shannon capacity of the channel model.

[2]For example, in an RFID application, the power used by the tag is actually supplied wirelessly by the reader. If the tag is the decoder, then it is natural to make ξ_D even larger than ξ_T in order to account for the inefficiency of the power transfer from the reader to the tag. One-to-many transmission of multicast data is another example of an application that can increase ξ_D. The ξ_D in that case should be increased in proportion to the number of receivers that are listening to the message.

[3]All that is required is the minimum power needed to sample the received signal and threshold the result.

This sense of capacity achieving makes explicit the sense in which we should consider encoding and decoding to be *asymptotically free*, but not actually free. The traditional approach of modeling encoding and decoding as being actually free can be recovered by swapping the order of the limits in (1).

Definition 5: An encoding and decoding system is considered *traditionally capacity achieving* if

$$\liminf_{\langle P_e \rangle \to 0} \liminf_{\xi_C, \xi_D \to \vec{0}} P_T(\vec{\xi}, \langle P_e \rangle, R) = C^{-1}(R). \tag{2}$$

where $C^{-1}(R)$ is the minimum transmission power that is predicted by the Shannon capacity of the channel model.

By taking the limit $(\xi_C, \xi_D) \to 0$ for a fixed probability of error, this traditional approach makes it impossible to capture any fundamental tradeoff with complexity in an asymptotic sense.

The conceptual distinction between the new (1) and old (2) senses of capacity-achieving systems parallels Shannon's distinction between zero-error capacity and regular capacity [20]. If $C(\epsilon, d)$ is the maximum rate that can be supported over a channel using end-to-end delay d and average probability of error ϵ, then traditional capacity $C = \lim_{\epsilon \to 0} \lim_{d \to \infty} C(\epsilon, d)$ while zero-error capacity $C_0 = \lim_{d \to \infty} \lim_{\epsilon \to 0} C(\epsilon, d)$. When the limits are taken together in some balanced way, then we get concepts like anytime capacity [16], [21]. It is known that $C_0 < C_{any} < C$ in general and so it is natural to wonder whether any codes are capacity achieving in the new stricter sense of Definition 4.

B. Are classical codes capacity achieving?

1) Dense linear block codes with nearest-neighbor decoding: Dense linear fixed-block-length codes are traditionally capacity achieving under ML decoding [13]. To understand whether they are weakly certainty achieving, we need a model for the encoding and decoding power. Let m be the block length of the code. Each codeword symbol requires mR operations to encode and it is reasonable to assume that each operation consumes some energy. Thus, the encoding power is $O(m)$. Meanwhile, a straightforward implementation of ML (nearest-neighbor) decoding has complexity exponential in the block-length and thus it is reasonable to assume that it consumes an exponential amount of power as well.

The probability of error for ML decoding drops exponentially with m with an exponent that is bounded above by the sphere-packing exponent $E_{sp}(R)$ [13]. An exponential reduction in the probability of error is thus paid for using an exponential increase in decoding power. Consequently, it is easy to see that the certainty return on investments in decoding power is only polynomial. Meanwhile, the certainty return on investments in transmit power is exponential even for uncoded transmission. So no matter what the values are for $\xi_D > 0$, in the high-certainty limit of very low probabilities of error, an optimized communication system built using dense linear block codes will be investing ever increasing amounts in transmit power.

A plot of the resulting waterslide curves for both transmit power and decoding power are given in Figure 2. Following tradition, the horizontal axes in the plots are given in normalized SNR units for power. Notice how the optimizing system invests heavily in additional transmit power to approach low probabilities of error.

2) Convolutional codes under Viterbi decoding: For convolutional codes, there are two decoding algorithms, and hence two different analyses. (See [22], [23] for details) For Viterbi decoding, the complexity per-bit is exponential in the constraint length RL_c bits. The error exponents with the constraint length of L_c channel uses are upper-bounded in [24], and this bound is given parametrically by

$$E_{conv}(R, P_T) = E_0(\rho, P_T) \ ; \ R = \frac{E_0(\rho, P_T)}{\rho} \tag{3}$$

where E_0 is the Gallager function [13] and $\rho > 0$. The important thing here is that just as in dense linear block codes, the certainty return on investments in decoding power is only polynomial, albeit with a better polynomial than linear block-codes since $E_{conv}(R, P_T)$ is higher than the sphere-packing bound for block codes [13]. Thus, an optimized communication system built using Viterbi decoding will also be investing ever increasing amounts in transmit power. Viterbi decoding is not weakly certainty achieving.

A plot of the resulting waterslide curves for both transmit power and decoding power is given in Figure 3. Notice that the performance in Figure 3 is better than that of Figure 2. This reflects the superior error exponents of convolutional codes with respect to their computational parameter — the constraint length.

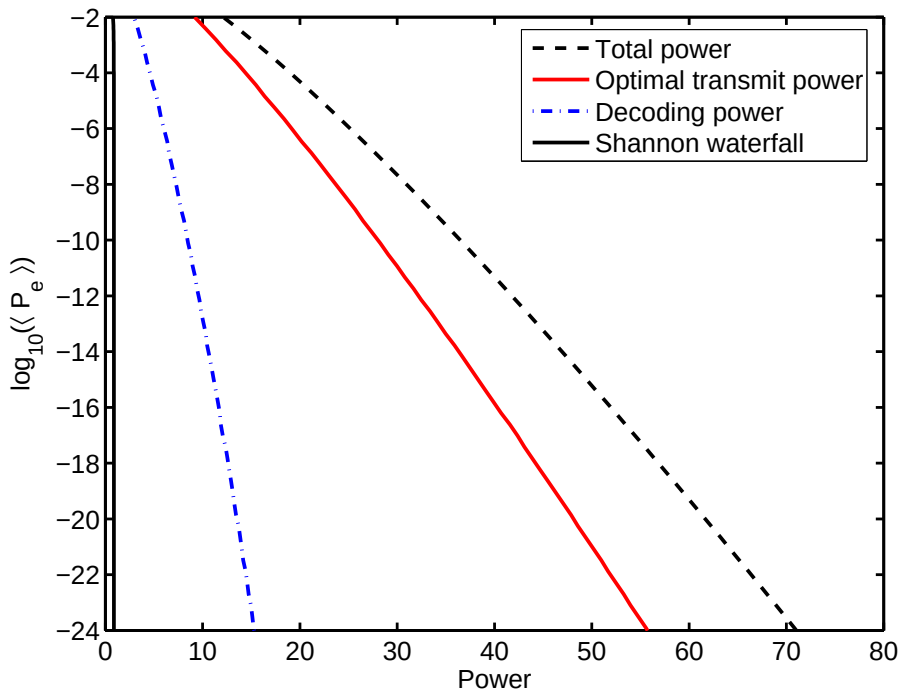

Fig. 2. The waterslide curves for transmit power, decoding power, and the total power for dense linear block-codes of rate $R = 1/3$ under brute-force ML decoding. It is assumed that the normalized energy required per operation at the decoder is $E = 0.3$ and that it takes $2^{mR} \times mR$ operations per channel output to decode using nearest-neighbor search for a block length of m channel uses.

3) Convolutional codes under magical sequential decoding: For convolutional codes with sequential decoding, it is shown in [25] that the average number of guesses must increase to infinity if the message rate exceeds the *cut-off rate*, $E_0(1)$. However, below the cut-off rate, the average number of guesses is finite. Each guess at the decoder costs $L_c R$ multiply-accumulates and we assume that this means that average decoding power also scales as $O(L_c)$ since at least one guess is made for each received sample.

For simplicity, let us ignore the issue of the cut-off rate and further assume that the decoder magically makes just one guess and always gets the ML answer. The convolutional coding error exponent (3) still applies, and so the system's certainty gets an exponential return for investments in decoding power. It is now no longer obvious how the optimized-system will behave in terms of transmit power.

For the magical system, the encoder power and decoder power are both linear in the constraint-length. Group them together with the path-loss and normalize units to get a single effective term γL_C. The goal now is to minimize

$$P_T + \gamma L_c \tag{4}$$

over P_T and L_C subject to the probability of error constraint that $\ln \frac{1}{\langle P_e \rangle} = E_{conv}(R, P_T)\frac{L_c}{R}$. Since we are interested in the limit of $\ln \frac{1}{\langle P_e \rangle} \to \infty$, it is useful to turn this around and use Lagrange multipliers. A little calculation reveals that the optimizing values of P_T and L_c must satisfy the balance condition

$$E_{conv}(R, P_T) = \gamma L_C \frac{\partial E_{conv}(R, P_T)}{\partial P_T} \tag{5}$$

and so (neglecting integer-effects) the optimizing constraint-length is either 1 (uncoded transmission) or

$$L_c = \frac{1}{\gamma} E_{conv}(R, P_T) / \frac{\partial E_{conv}(R, P_T)}{\partial P_T}. \tag{6}$$

To get ever lower values of $\langle P_e \rangle$, the transmit power P_T must therefore increase unboundedly unless the ratio $E_{conv}(R, P_T)/\frac{\partial E_{conv}(R,P_T)}{\partial P_T}$ approaches infinity for some finite P_T. Since the convolutional coding error exponent

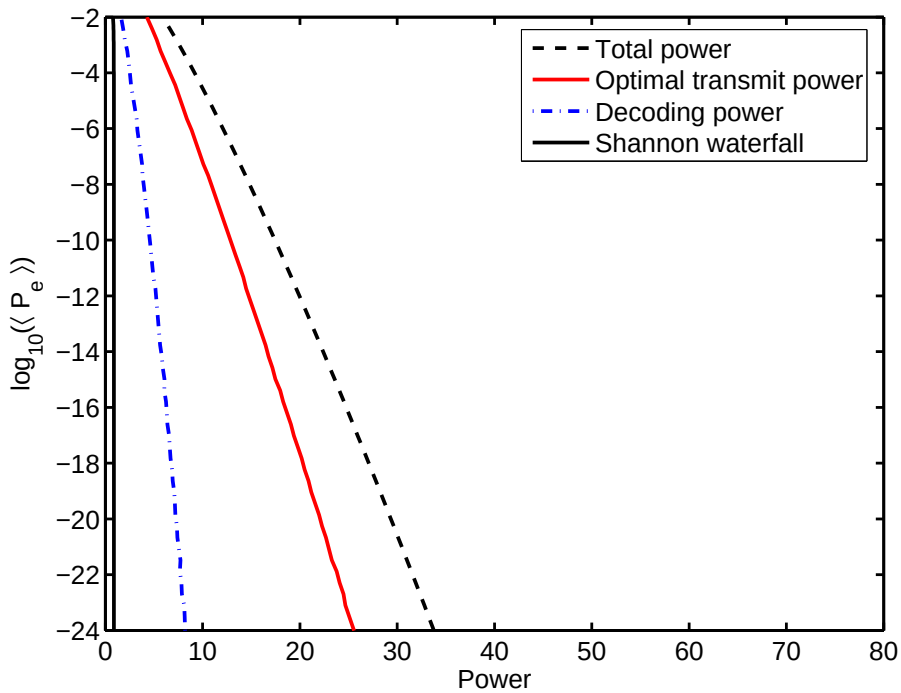

Fig. 3. The waterslide curves for transmit power, decoding power, and the total power for convolutional codes of rate $R = 1/3$ used with Viterbi decoding. It is assumed that the normalized energy required per operation at the decoder is $E = 0.3$ and that it takes $2^{L_c R} \times L_c R$ operations per channel output to decode using Viterbi search for a constraint length of L_c channel uses.

(3) does not go to infinity at a finite power, this requires $\frac{\partial E_{conv}(R,P_T)}{\partial P_T}$ to approach zero. For AWGN style channels, this only occurs[4] as P_T approaches infinity and thus the gap between R and the capacity gets large.

The resulting plots for the waterslide curves for both transmit power and encoding/decoding power are given in Figure 4. Although these plots are much better than those in Figure 3, the surprise is that even such a magical system that attains an error-exponent with investments in decoding power is unable to be weakly certainty achieving at any rate. Instead, the optimizing transmit power goes to infinity.

4) Dense linear block codes with magical syndrome decoding: It is well known that linear codes can be decoded by looking at the syndrome of the received codeword [13]. Suppose that we had a magical syndrome decoder that could use a free lookup table to translate the syndrome into the ML corrections to apply to the received codeword. The complexity of the decoding would just be the complexity of computing the syndrome. For a dense random linear block code, the parity-check matrix is itself typically dense and so the per-channel-output complexity of computing each bit of the syndrome is linear in the block-length. This gives rise to behavior like that of magical sequential decoding above and is illustrated in Figure 5.

From the above discussion, it seems that in order to have even a weakly certainty-achieving system, the certainty-return for investments in encoding/decoding power must be faster than exponential!

III. Parallel iterative decoding: a new hope

The unrealistic magical syndrome decoder suggests a way forward. If the parity-check matrix were sparse, then it would be possible to compute the syndrome using a constant number of operations per received symbol. If the probability of error dropped with block-length, that would give rise to an infinite-return on investments in decoder

[4]There is a slightly subtle issue here. Consider random codes for a moment. The convolutional random-coding error exponent is flat at $E_0(1, P_T)$ for rates R below the computational cutoff rate. However, that flatness with rate R is not relevant here. For any fixed constellation, the $E_0(1, P_T)$ is a strictly monotonically increasing function of P_T, even though it asymptotes at a non-infinite value. This is not enough since the derivative with transmit power still tends to zero only as P_T goes to infinity.

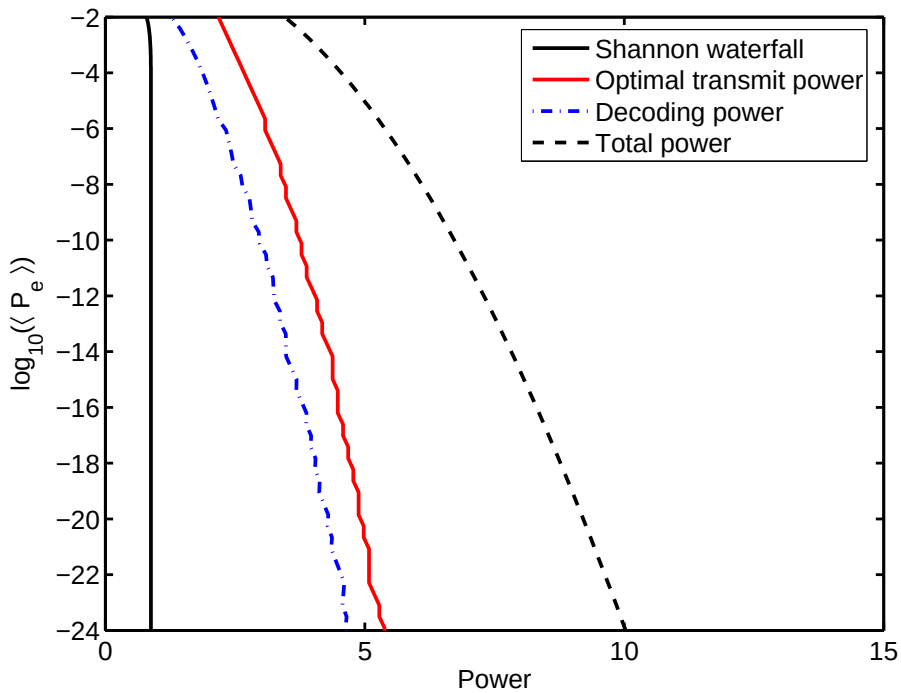

Fig. 4. The waterslide curves for transmit power, decoding power, and the total power for convolutional codes of rate $R = 1/3$ used with "magical" sequential decoding. It is assumed that the normalized energy required per operation at the decoder is $E = 0.3$ and that the decoding requires just $L_c R$ operations per channel output.

power. This suggests looking in the direction of LDPC codes [26]. While magical syndrome decoding is unrealistic, many have observed that message-passing decoding gives good results for such codes while being implementable [27].

Upon reflection, it is clear that parallel iterative decoding based on message passing holds out the potential for *super-exponential* improvements in probability of error with decoding power. This is because messages can reach an exponential-sized neighborhood in only a small number of iterations, and large-deviations thinking suggests that there is the possibility for an exponential reduction in the probability of error with neighborhood size. In fact, exactly this sort of double-exponential reduction in the probability of error under iterative decoding has been shown to be possible for regular LDPCs [28, Theorem 5].

To make all this precise, we need to fix our model of the problem and of an implementable decoder. Consider a point-to-point communication link. An information sequence $\mathbf{B}_1^k$ is encoded into 2^{mR} codeword symbols $\mathbf{X}_1^m$, using a possibly randomized encoder. The observed channel output is $\mathbf{Y}_1^m$. The information sequences are assumed to consist of iid fair coin tosses and hence the rate of the code is $R = k/m$. Following tradition, both k and m are considered to be very large. We ignore the complexity of doing the encoding under the hope that encoding is simpler than decoding.[5]

Two channel models are considered: the BSC and the power-constrained AWGN channel. The true channel is always denoted P. The underlying AWGN channel has noise variance σ_P^2 and the average received power is denoted P_T so the received SNR is $\frac{P_T}{\sigma_P^2}$. Similarly, we assume that the BSC has crossover probability p. We consider the BSC to have resulted from BPSK modulation followed by hard-decision detection on the AWGN channel and so $p = \mathcal{Q}\left(\sqrt{\frac{P_T}{\sigma_P^2}}\right)$.

For maximum generality, we do not impose any *a priori* structure on the code itself. Instead, inspired by [30]–

[5]For certain LDPC-codes, it is shown in [29] that encoding can be made to have complexity linear in the block-length for a certain model of encoding. In our context, linear complexity means that the complexity per data bit is constant and thus this does not require power at the encoder that grows with either the block length or the number of decoder iterations. We have not yet verified if the complexity of encoding is linear under our computational model.

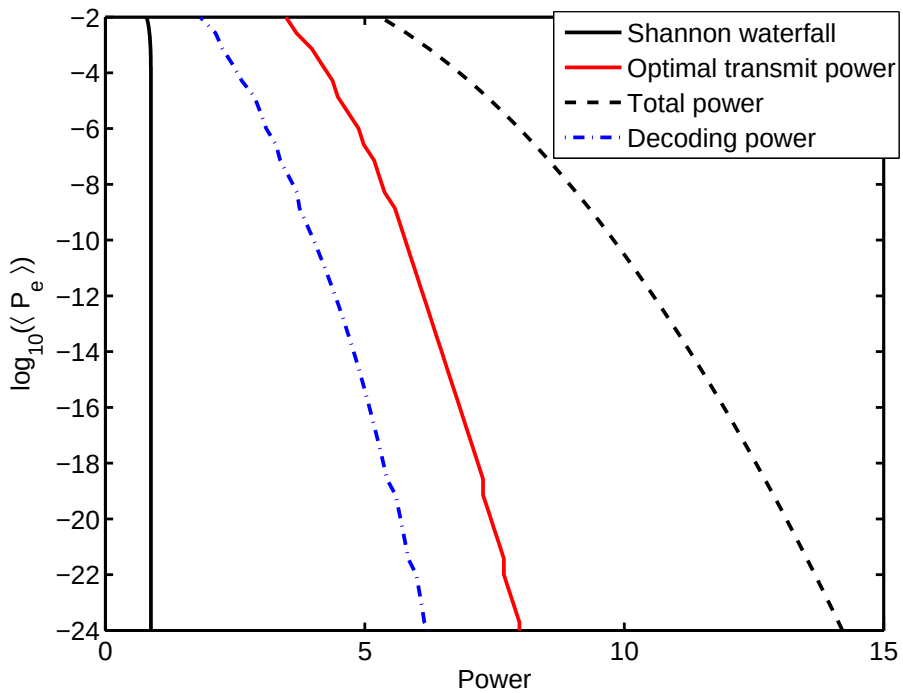

Fig. 5. The waterslide curves for transmit power, decoding power, and the total power for dense linear block-codes of rate $R = 1/3$ under magical syndrome decoding. It is assumed that the normalized energy required per operation at the decoder is $E = 0.3$ and that the decoding requires just $(1 - R)mR$ operations per channel output to compute the syndrome.

[33], we focus on the parallelism of the decoder and the energy consumed within it. We assume that the decoder is physically made of computational nodes that pass messages to each other in parallel along physical (and hence unchanging) wires. A subset of nodes are designated 'message nodes' in that each is responsible for decoding the value of a particular message bit. Another subset of nodes (not necessarily disjoint) has members that are each initialized with at most one observation of the received channel-output symbols. There may be additional computational nodes that are just there to help decode.

The implementation technology is assumed to dictate that each computational node is connected to at most $\alpha + 1 > 2$ other nodes[6] with bidirectional wires. No other restriction is assumed on the topology of the decoder. In each iteration, each node sends (possibly different) messages to all its neighboring nodes. **No restriction is placed on the size or content of these messages except for the fact that they must depend on the information that has reached the computational node in previous iterations.** If a node wants to communicate with a more distant node, it has to have its message relayed through other nodes. No assumptions are made regarding the presence or absence of cycles in this graph. The neighborhood size at the end of l iterations is denoted by $n \leq \alpha^{l+1}$. We assume $m \gg n$. Each computational node is assumed to consume a fixed E_{node} joules of energy at each iteration.

Let the average probability of bit error of a code be denoted by $\langle P_e \rangle_P$ when it is used over channel P. The goal is to derive a lower bound on the neighborhood size n as a function of $\langle P_e \rangle_P$ and R. This then translates to a lower bound on the number of iterations which can in turn be used to lower bound the required decoding power.

Throughout this paper, we allow the encoding and decoding to be randomized with all computational nodes allowed to share a common pool of common randomness. We use the term 'average probability of error' to refer to the probability of bit error averaged over the channel realizations, the messages, the encoding, and the decoding.

[6]In practice, this limit could come from the number of metal layers on a chip. $\alpha = 1$ would just correspond to a big ring of nodes and is uninteresting for that reason.

IV. Lower bounds on decoding complexity: iterations and power

In this section, lower bounds are stated on the computational complexity for iterative decoding as a function of the gap from capacity. These bounds reveal that the decoding neighborhoods must grow unboundedly as the system tries to approach capacity. We assume the decoding algorithm is implemented using the iterative technology described in Section III. The resulting bounds are then optimized numerically to give plots of the optimizing transmission and decoding powers as the average probability of bit error goes to zero. For transmit power, it is possible to evaluate the limiting value as the system approaches certainty. However, decoding power is shown to diverge to infinity for the same limit. This shows that the lower bound does not rule out weakly capacity-achieving schemes, but strongly capacity-achieving schemes are impossible using Section III's model of iterative decoding.

A. Lower bounds on the probability of error in terms of decoding neighborhoods

The main bounds are given by theorems that capture a local sphere-packing effect. These can be turned around to give a family of lower bounds on the neighborhood size n as a function of $\langle P_e \rangle_P$. This family is indexed by the choice of a hypothetical channel G and the bounds can be optimized numerically for any desired set of parameters.

Theorem 1: Consider a BSC with crossover probability $p < \frac{1}{2}$. Let n be the maximum size of the decoding neighborhood of any individual bit. The following lower bound holds on the average probability of bit error.

$$\langle P_e \rangle_P \geq \sup_{C^{-1}(R) < g \leq \frac{1}{2}} \frac{h_b^{-1}(\delta(G))}{2} 2^{-nD(g||p)} \left(\frac{p(1-g)}{g(1-p)} \right)^{\epsilon\sqrt{n}} \tag{7}$$

where $h_b(\cdot)$ is the usual binary entropy function, $D(g||p) = g \log_2 \left(\frac{g}{p}\right) + (1-g) \log_2 \left(\frac{1-g}{1-p}\right)$ is the usual KL-divergence, and

$$\delta(G) = 1 - \frac{C(G)}{R} \tag{8}$$

$$\text{where } C(G) = 1 - h_b(g)$$

$$\text{and } \epsilon = \sqrt{\frac{1}{K(g)} \log_2 \left(\frac{2}{h_b^{-1}(\delta(G))} \right)} \tag{9}$$

$$\text{where } K(g) = \inf_{0<\eta<1-g} \frac{D(g+\eta||g)}{\eta^2}. \tag{10}$$

Proof: See Appendix I. ■

Theorem 2: For the AWGN channel and the decoder model in Section III, let n be the maximum size of the decoding neighborhood of any individual message bit. The following lower bound holds on the average probability of bit error.

$$\langle P_e \rangle_P \geq \sup_{\sigma_G^2 :\, C(G)<R} \frac{h_b^{-1}\left(\delta(G)\right)}{2} \exp\left(-nD(\sigma_G^2||\sigma_P^2) - \sqrt{n} \left(\frac{3}{2} + 2\ln\left(\frac{2}{h_b^{-1}(\delta(G))} \right) \right) \left(\frac{\sigma_G^2}{\sigma_P^2} - 1 \right) \right) \tag{11}$$

where $\delta(G) = 1 - C(G)/R$, the capacity $C(G) = \frac{1}{2}\log_2\left(1 + \frac{P_T}{\sigma_G^2}\right)$, and the KL divergence $D(\sigma_G^2||\sigma_P^2) = \frac{1}{2}\left[\frac{\sigma_G^2}{\sigma_P^2} - 1 - \ln\left(\frac{\sigma_G^2}{\sigma_P^2}\right)\right]$.

The following lower bound also holds on the average probability of bit error

$$\langle P_e \rangle_P \geq \sup_{\sigma_G^2 > \sigma_P^2 \mu(n) :\, C(G)<R} \frac{h_b^{-1}\left(\delta(G)\right)}{2} \exp\left(-nD(\sigma_G^2||\sigma_P^2) - \frac{1}{2}\phi(n, h_b^{-1}\left(\delta(G)\right)) \left(\frac{\sigma_G^2}{\sigma_P^2} - 1 \right) \right), \tag{12}$$

where

$$\mu(n) = \frac{1}{2}\left(1 + \frac{1}{T(n)+1} + \frac{4T(n)+2}{nT(n)(1+T(n))}\right) \tag{13}$$

$$\text{where } T(n) = -W_L(-\exp(-1)(1/4)^{1/n}) \tag{14}$$

$$\text{and } W_L(x) \text{ solves } x = W_L(x)\exp(W_L(x)) \tag{15}$$

$$\text{while satisfying } W_L(x) \leq -1 \ \forall x \in [-\exp(-1), 0],$$

and

$$\phi(n,y) = -n(W_L\left(-\exp(-1)(\frac{y}{2})^{\frac{2}{n}}\right)+1). \tag{16}$$

The $W_L(x)$ is the transcendental Lambert W function [34] that is defined implicitly by the relation (15) above.

Proof: See Appendix II. ■

The expression (12) is better for plotting bounds when we expect n to be moderate while (11) is more easily amenable to asymptotic analysis as n gets large.

B. Joint optimization of the weighted total power

Consider the total energy spent in transmission. For transmitting k bits at rate R, the number of channel uses is $m = k/R$. If each transmission has power $\xi_T P_T$, the total energy used in transmission is $\xi_T P_T m$.

At the decoder, let the number of iterations be l. Assume that each node consumes E_{node} joules of energy in each iteration. The number of computational nodes can be lower bounded by the number m of received channel outputs.

$$E_{dec} \geq E_{node} \times m \times l. \tag{17}$$

This gives a lower bound of $P_D \geq E_{node} l$ for decoder power. There is no lower bound on the encoder complexity and so the encoder is considered free. This results in the following bound for the weighted total power

$$P_{total} \geq \xi_T P_T + \xi_D E_{node} \times l. \tag{18}$$

Using $l \geq \frac{\log_2(n)}{\log_2(\alpha)}$ as the natural lower bound on the number of iterations given a desired maximum neighborhood size,

$$\begin{aligned} P_{total} &\geq \xi_T P_T + \frac{\xi_D E_{node} \log_2(n)}{\log_2(\alpha)} \\ &\propto \frac{P_T}{\sigma_P^2} + \gamma \log_2(n) \end{aligned} \tag{19}$$

where $\gamma = \frac{\xi_D E_{node}}{\sigma_P^2 \xi_T \log_2(\alpha)}$ is a constant that summarizes all the technology and environmental terms. The neighborhood size n itself can be lower bounded by plugging the desired average probability of error into Theorems 1 and 2.

It is clear from (19) that for a given rate R bits per channel use, if the transmit power P_T is extremely close to that predicted by the channel capacity, then the value of n would have to be extremely large. This in turn implies that there are a large number of iterations and thus it would require high power consumption at the decoder. Therefore, the optimized encoder has to transmit at a power larger than that predicted by the Shannon limit in order to decrease the power consumed at the decoder. Also, from (7), as $\langle P_e \rangle \to 0$, the required neighborhood size $n \to \infty$. This implies that for any fixed value of transmit power, the power consumed at the decoder diverges to infinity as the probability of error converges to zero. Hence the total power consumed must diverge to infinity as the probability of error converges to zero. This immediately rules out the possibility of having a strongly certainty-achieving code using this model of iterative decoding. The price of certainty is infinite power. The only question that remains is whether the optimal transmitter power can remain bounded or not.

The optimization can be performed numerically once the exchange rate ξ_T is fixed, along with the technology parameters $E_{node}, \alpha, \xi_C, \xi_D$. Figures 6 and 7 show the total-power waterslide curves for iterative decoding assuming the lower bounds.[7] These plots show the effect of changing the relative cost of decoding. The waterslide curves become steeper as decoding becomes cheaper and the plotted scale is chosen to clearly illustrate the double-exponential relationship between decoder power and probability of error.

Figure 8 fixes the technology parameters and breaks out the optimizing transmit power and decoder power as two separate curves. It is important to note that only the weighted total power curve is a true bound on what a real system could achieve. The constituent P_T and P_D curves are merely indicators of what the qualitative behaviour

[7]The order-of-magnitude choice of $\gamma = 0.3$ was made using the following numbers. The energy cost of one iteration at one node $E_{node} \approx 1$pJ (optimistic extrapolation from the reported values in [4], [12]), path-loss $\xi_T \approx 86dB$ corresponding to a range in the tens of meters, thermal noise energy per sample $\sigma_P^2 \approx 4 \times 10^{-21}$J from kT with T around room temperature, and computational node connectivity $\alpha = 4$.

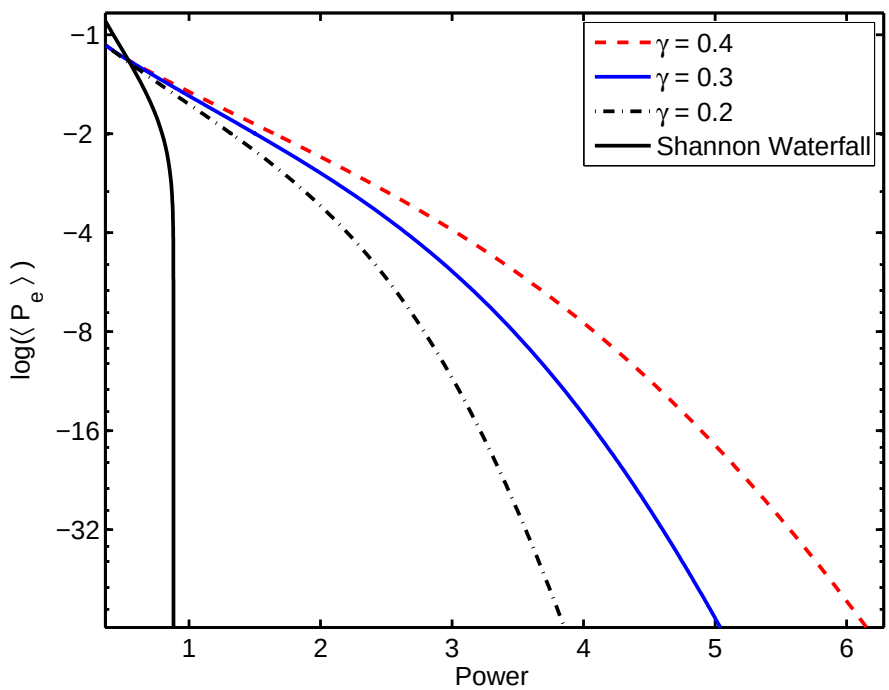

Fig. 6. The BSC Waterslides: plots of $\log(\langle P_e \rangle)$ vs bounds on required total power for any fixed rate-$1/3$ code transmitted over an AWGN channel using BPSK modulation and hard decisions. $\gamma = \xi_D E_{node}/(\xi_T \sigma_P^2 \log_2(\alpha))$ denotes the normalized energy per node per iteration in SNR units. Total power takes into account the transmit power as well as the power consumed in the decoder. The Shannon limit is a universal lower bound for all γ.

would be if the true tradeoff behaved like the lower bound.[8] The optimal transmit power approaches a finite limit as the probability of error approaches 0. This limit can be calculated directly by examining (7) for the BSC and (11) for the AWGN case.

To compute this limit, recall that the goal is to optimize $\frac{P_T}{\sigma_P^2} + \gamma \log_2(n)$ over P_T so as to satisfy a probability of error constraint $\langle P_e \rangle$, where the probability of error is tending to zero. Instead of constraining the probability of error to be small, it is just as valid to constrain $\gamma \log\log \frac{1}{\langle P_e \rangle}$ to be large. Now, take the logarithm of both sides of (7) (or similarly for (11)). It is immediately clear that the only order n term is the one that multiplies the divergence. Since $n \to \infty$ as $\langle P_e \rangle \to 0$, this term will dominate when a second logarithm is taken. Thus, we know that the bound on the double logarithm of the certainty $\gamma \log\log \frac{1}{\langle P_e \rangle} \to \gamma \log_2(n) + \gamma \log f(R, \frac{P_T}{\sigma_P^2})$ where $f(R, \frac{P_T}{\sigma_P^2}) = D(G||P)$ is the divergence expression involving the details of the channel. It turns out that G approaches $C^{-1}(R)$ when $\langle P_e \rangle \to 0$ since the divergence is maximized there.

Optimizing for $\zeta = \frac{P_T}{\sigma_P^2}$ by taking derivatives and setting to zero gives:

$$f(R, \zeta) / \frac{\partial f(R, \zeta)}{\partial \zeta} = \gamma. \tag{20}$$

It turns out that this has a unique root $\zeta(R, \gamma)$ for all rates R and technology factors γ for both the BSC and the AWGN channel.

The key difference between (5) and (20) is that no term that is related to the neighborhood-size or number of iterations has survived in (20). This is a consequence of the double-exponential[9] reduction in the probability of

[8]This doesn't mean that the bound is useless however. A lower bound on the transmit power can be computed once any implementable scheme exists. Simply look up where the bounded total power matches the implementable scheme. This will immediately give rise to lower bounds on the optimal transmit and decoding powers.

[9]In fact, it is easy to verify that anything faster than double-exponential will also work.

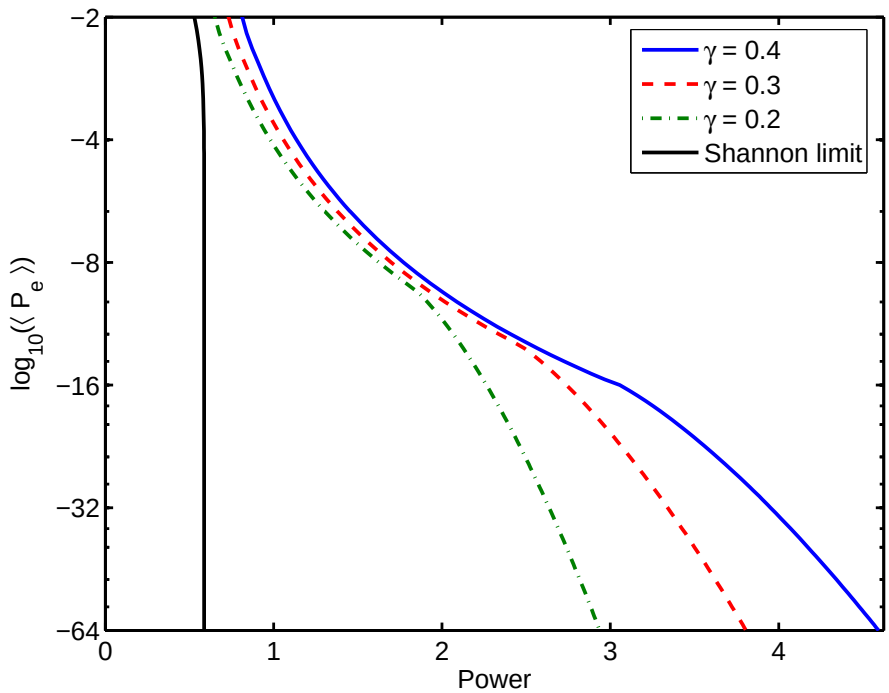

Fig. 7. The AWGN Waterslides: plots of $\log(\langle P_e \rangle)$ vs bounds on required total power for any fixed rate-$1/3$ code transmitted over an AWGN channel. The initial segment where all the waterslide curves almost coincide illustrates the looseness of the bound since that corresponds to the case of $n = 1$ or when the bound suggests that uncoded transmission could be optimal. However, the probability of error is too optimistic for uncoded transmission.

error with the number of iterations and the fact that the transmit power shows up in the outer and not the inner exponential.

To see if iterative decoding allows weakly capacity-achieving codes, we take the limit of $\xi_D \to 0$ which implies $\gamma \to 0$. (20) then suggests that we need to solve $f(R,\zeta)/\frac{\partial f(R,\zeta)}{\partial \zeta} = 0$ which implies that either the numerator is zero or the denominator becomes infinite. For AWGN or BSC channels, the slope of the error exponent $f(R,\zeta)$ is monotonically decreasing as the SNR $\zeta \to \infty$ and so the unique solution is where $f(R,\zeta) = D(C^{-1}(R)||P_T) = 0$. This occurs at $P_T = C^{-1}(R)$ and so the lower bounds of this section do not rule out weakly capacity-achieving codes.

In the other direction, as the γ term gets large, the $P_T(R,\gamma)$ increases. This matches the intuition that as the relative cost of decoding increases, more power should be allocated to the transmitter. This effect is plotted in Figure 9. Notice that it becomes insignificant when γ is very small (long-range communication) but becomes non-negligible whenever the γ exceeds 0.1.

Figure 10 illustrates how the effect varies with the desired rate R. The penalty for using low-rate codes is quite significant and this gives further support to the lessons drawn from [7], [9] with some additional intuition regarding why it is fundamental. The error exponent governing the probability of error as a function of the neighborhood size is limited by the sphere-packing bound at rate 0 – this is finite and the only way to increase it is to pay more transmit power. However, the decoding power is proportional to the number of received samples and this is larger at lower rates.

Finally, the plots were all made assuming that the neighborhood size n could be chosen arbitrarily and the number of iterations could be a real number rather than being restricted to integer values. This is fine when the desired probability of error is low, but it turns out that this integer effect cannot be neglected when the tolerable probability of error is high. This is particularly significant when γ is large. To see this, it is useful to consider the boundary between when uncoded transmission is optimal and when coding might be competitive. This is done in Figure 11 where the minimum $\gamma \log_2(\alpha)$ power required for the first decoder iteration is instead given to the

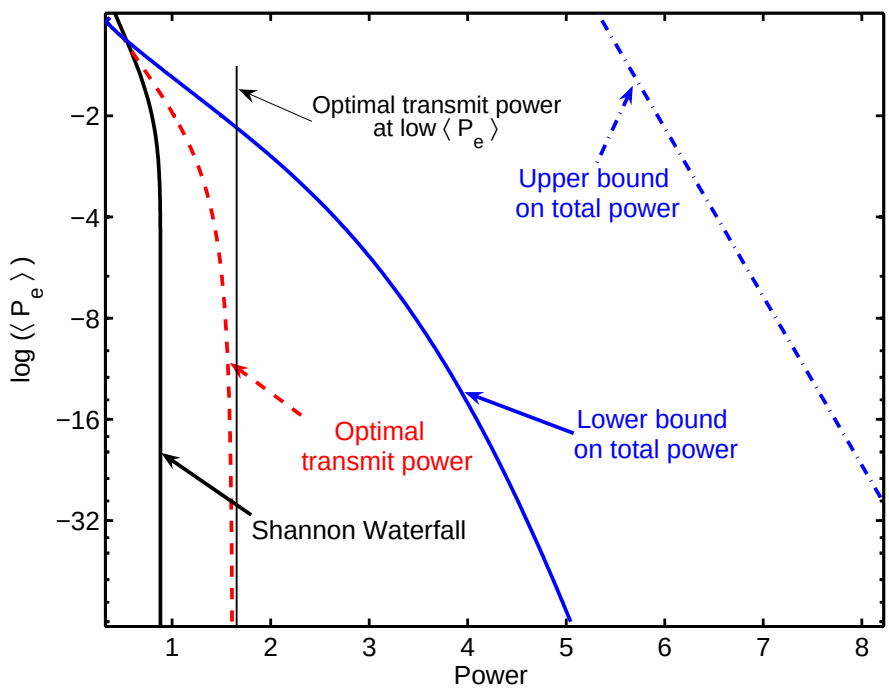

Fig. 8. The BSC Waterslide curve for $\gamma = 0.3$, $R = 1/3$. An upper bound (from Section IV-C) that is parallel to the lower bound is also shown along with the heuristically optimal transmit power. This transmit power is larger than that predicted by the Shannon limit for small probabilities of error. This suggests that the transmitter has to make accommodations for the decoder complexity in order to minimize the total power consumption.

transmitter. Once $\gamma > 10$, it is hard to beat uncoded transmission unless the desired probability of error is very low indeed.

C. Upper bounds on complexity

It is unclear how tight the lower bounds given earlier in this section are. The most shocking aspect of the lower bounds is that they predict a double exponential improvement in probability of error with the number of iterations. This is what is leading to the potential for weakly capacity-achieving codes. To see the order-optimality of the bound in principle, we will "cheat" and exploit the fact that our model for iterative decoding in Section III does not limit either the size of the messages or the computational power of each node in the decoder. This allows us to give upper bounds on the number of iterations required for a given performance.

Theorem 3: There exists a code of rate $R < C$ such that the required neighborhood size to achieve $\langle P_e \rangle$ average probability of error is upper bounded by

$$n \leq \frac{\log_2\left(\frac{1}{\langle P_e \rangle}\right)}{E_r(R)} \tag{21}$$

where $E_r(R)$ is the random-coding error exponent for the channel [13]. The required number of iterations to achieve this neighborhood size is bounded above by

$$l - 2 \leq 2\frac{\log_2(n)}{\log_2(\alpha)}. \tag{22}$$

Proof: This "code" is basically an abuse of the definitions. We simply use a rate-R random code of length n from [13] where each code symbol is drawn iid. Such random codes if decoded using ML decoding satisfy

$$\langle P_e \rangle_P \leq \langle P_e \rangle_{\text{block}} \leq \exp(-nE_r(R)). \tag{23}$$

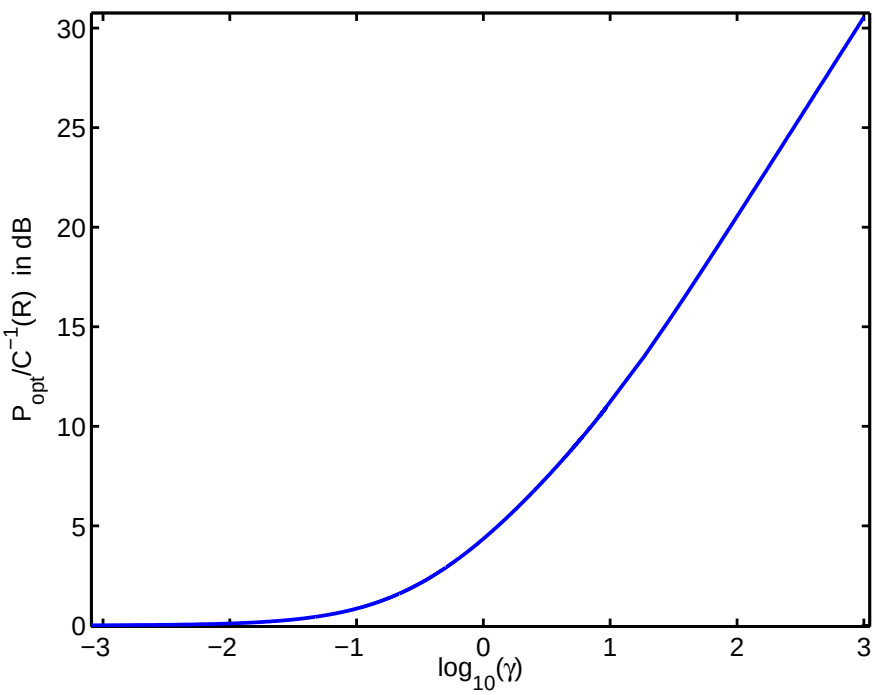

Fig. 9. The impact of γ on the heuristically predicted optimum transmit power for the BSC used at $R = \frac{1}{3}$. The plot shows the gap from the Shannon prediction in a factor sense.

The decoder for each bit needs at most n channel-output symbols to decode the block (and hence any particular bit).

Now it is enough to show an upper bound on the number of iterations, l. Consider a regular tree structure imposed on the code with a branching factor of α and thus overall degree $\alpha+1$. Since the tree would have α^d nodes in it at depth d, a required depth of $d = \frac{\log_2(n)}{\log_2(\alpha)} + 1$ is sufficient to guarantee that everything within a block is connected.

Designate some subset of computational nodes as responsible for decoding the individual message bits. At each iteration, the "message" transmitted by a node is just the complete list of its own observation plus all the messages that node has received so far. Because the diameter of a tree is no more than twice its depth, at the end of $2d$ iterations, all the nodes will have received all the values of received symbols in the neighborhood. They can then each ML decode the whole block, with average error probability given by (23). The result follows. ■

For both the AWGN channel and BSC, this bound recovers the basic behavior that is needed to have the probability of error drop doubly-exponentially in the number of iterations. For the BSC, it is also clear that since $E_r(R) = D(C^{-1}(R)||p)$ for rates R in the neighborhood of capacity, the upper and lower bounds essentially agree on the asymptotic neighborhood size when $\langle P_e \rangle \to 0$. The only difference comes in the number of iterations. This is at most a factor of 2 and so has the same effect as a slightly different ξ_D in terms of the shape of the curves and optimizing transmit power.

We note here that this upper bound points to the fact that the decoding model of Section III is too powerful rather than being overly constraining. It allows free computations at each node and unboundedly large messages. This suggests that the lower bounds are relevant, but it is unclear whether they are actually attainable with any implementable code. We delve further into this in Section VI.

V. The gap to capacity and related work

Looking back at our bounds of Theorems 1 and 2, they seem to suggest that a certain minimum number ($\log_\alpha f(R, P_T)$) of iterations are required and after that, the probability of error can drop doubly exponentially with additional iterations. This parallels the result of [28, Theorem 5] for regular LDPCs that essentially implies that regular LDPCs can be considered weakly *certainty-achieving* codes. However, our bounds above indicate that

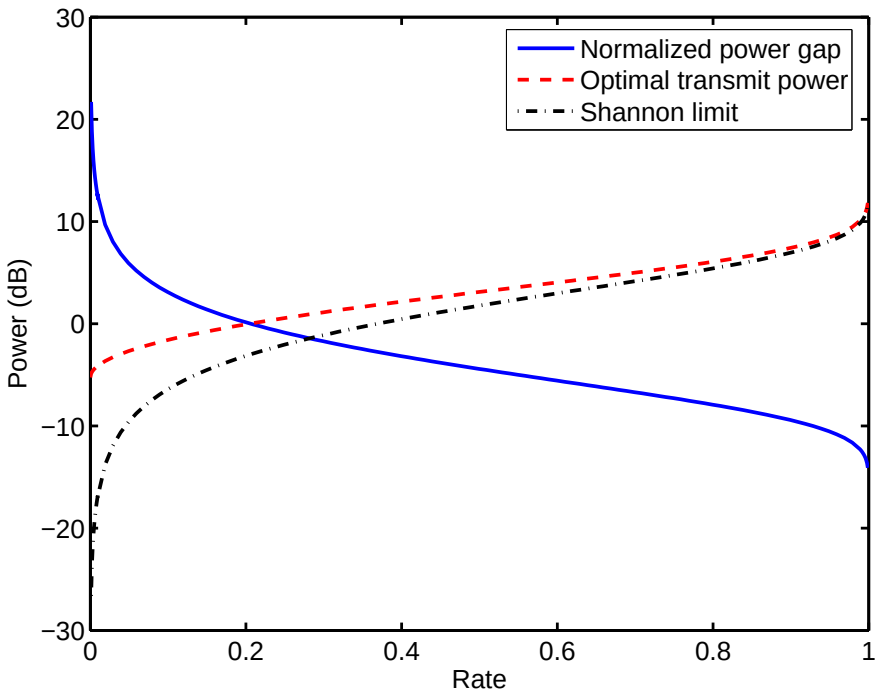

Fig. 10. The impact of rate R on the heuristically predicted optimum transmit power for $\gamma = 0.3$. The plot shows the Shannon minimum power, our predictions, and the ratio of the difference between the two to the Shannon minimum. Notice that the predicted extra power is very substantial at low data rates.

iterative decoding might be compatible with weakly *capacity-achieving* codes as well. Thus, it is interesting to ask how the complexity behaves if we operate very close to capacity. Following tradition, denote the difference between the channel capacity $C(P)$ and the rate R as the $gap = C(P) - R$.

Since our bounds are general, it is interesting to compare them with the existing specialized bounds in the vicinity of capacity. After first reviewing a trivial bound in Section V-A to establish a baseline, we review some key results in the literature in Section V-B. Before we can give our results, we take another look at the waterfall curve in Figure 1 and notice that there are a number of ways to approach the Shannon limit. We discuss our approach in Section V-C before giving our lower bounds to the number of iterations in Section V-D.

A. The trivial bound for the BSC

Given a crossover probability p, it is important to note that there exists a semi-trivial bound on the neighborhood size that only depends on the $\langle P_e \rangle$. Since there is at least one configuration of the neighborhood that will decode to an incorrect value for this bit, it is clear that

$$\langle P_e \rangle \geq p^n. \tag{24}$$

This implies that the number of computational iterations for a code with maximum decoding degree $\alpha + 1$ is lower bounded by $\frac{\log\log\frac{1}{\langle P_e \rangle} - \log\log\frac{1}{p}}{\log\alpha}$. This trivial bound does not have any dependence on the capacity and so does not capture the fact that the complexity should increase inversely as a function of gap as well.

B. Prior work

There is a large literature relating to codes that are specified by sparse graphs. The asymptotic behavior as these codes attempt to approach Shannon capacity is a central question in that literature. For regular LDPC codes, a result in Gallager's Ph.D. thesis [26, Pg. 40] shows that the average degree of the graph (and hence the average number of operations per iteration) must diverge to infinity in order for these codes to approach capacity even under ML

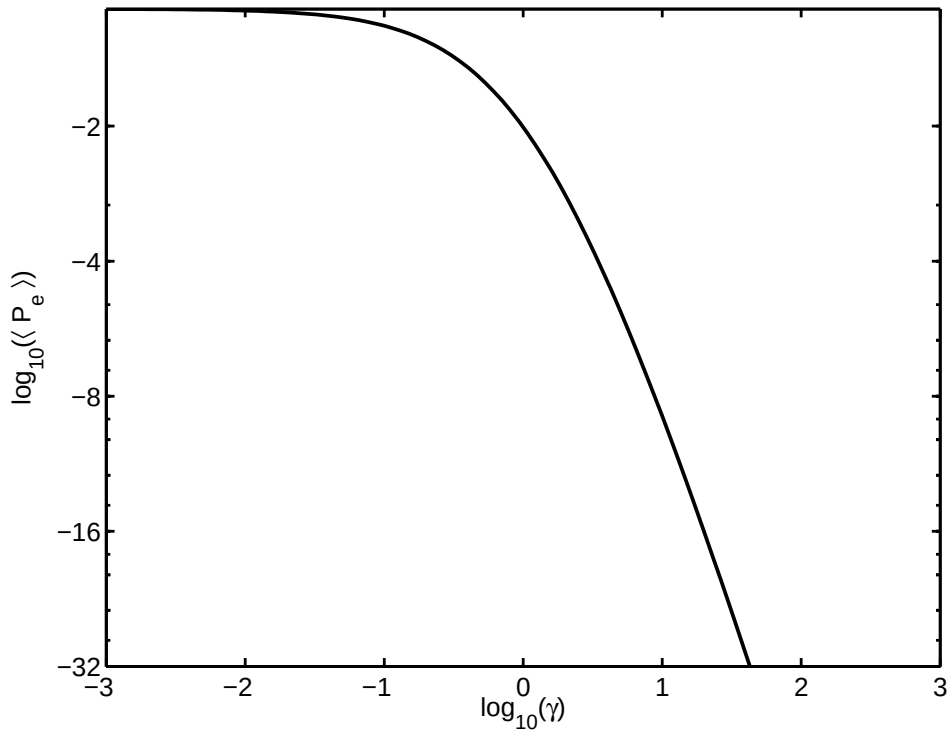

Fig. 11. The probability of error below which coding could potentially be useful. This plot assumes an AWGN channel used with BPSK signaling and hard-decision detection, target message rate $R = \frac{1}{3}$, and an underlying iterative-decoding architecture with $\alpha = 3$. This plot shows what probability of error would be achieved by uncoded transmission (repetition coding) if the transmitter is given extra power beyond that predicted by Shannon capacity. This extra power corresponds to that required to run one iteration of the decoder. Once γ gets large, there is effectively no point in doing coding.

decoding. It turns out that it is not hard to specialize our Theorem 1 to regular LDPC codes and have it become tighter along the way. Such a modified bound would show that as the gap *from Gallager's rate bound* converges to zero, the number of iterations must diverge to infinity. However, it would permit double-exponential improvements in the probability of error as the number of iterations increased.

More recently, in [35, Pg. 69] and [36], Khandekar and McEliece conjectured that for all sparse-graph codes, the number of iterations must scale either multiplicatively as

$$\Omega\left(\log_2\left(\frac{1}{\langle P_e\rangle}\right)\frac{1}{gap}\right), \tag{25}$$

or additively as

$$\Omega\left(\frac{1}{gap} + \log_2\left(\frac{1}{\langle P_e\rangle}\right)\right) \tag{26}$$

in the near neighborhood of capacity. Here we use the Ω notation to denote lower-bounds in the order sense of [37]. This conjecture is based on a graphical argument for the message-passing decoding of sparse-graph codes over the BEC. The intuition was that the bound should also hold for general memoryless channels, since the BEC is the channel with the simplest decoding.

Recently, the authors in [38] were able to formalize and prove a part of the Khandekar-McEliece conjecture for three important families of sparse-graph codes, namely the LDPC codes, the Accumulate-Repeat-Accumulate (ARA) codes, and the Irregular-Repeat Accumulate (IRA) codes. Using some remarkably simple bounds, the authors demonstrate that the number of iterations usually scales as $\Omega(\frac{1}{gap})$ for Binary Erasure Channels (BECs). If, however, the fraction of degree-2 nodes for these codes converges to zero, then the bounds in [38] become trivial. The authors note that all the known traditionally capacity-achieving sequences of these code families have a non-zero fraction of degree-2 nodes.

In addition, the bounds in [38] do not imply that the number of decoding iterations must go to infinity as $\langle P_e \rangle \to 0$. So the conjecture is not yet fully resolved. We can observe however that both of the conjectured bounds on the number of decoding iterations have only a singly-exponential dependence of the probability of error on the number of iterations. The multiplicative bound (26) behaves like a block or convolutional code with an error-exponent of $K \times gap$ and so, by the arguments of Section II-B.3, is not compatible with such codes being weakly capacity achieving in our sense. However, it turns out that the additive bound (25) *is* compatible with being weakly capacity achieving. This is because the main role of the double-exponential in our derivation is to allow a second logarithm to be taken that decoupled the term depending on the transmit power from the one that depends on the probability of error. The conjectured additive bound (25) has that form already.

C. 'Gap' to capacity

In the vicinity of capacity, the complication is that for any finite probability of bit-error, it is in principle possible to communicate at rates **above** the channel capacity. Before transmission, the k bits could be lossily compressed using a source code to $\approx (1 - h_b(\langle P_e \rangle))k$ bits. The channel code could then be used to protect these bits, and the resulting codeword transmitted over the channel. After decoding the channel code, the receiver could in principle use the source decoder to recover the message bits with an acceptable average probability of bit error. Therefore, for fixed $\langle P_e \rangle$, the maximum achievable rate is $\frac{C}{1-h_b(\langle P_e \rangle)}$.

Consequently, the appropriate *total gap* is $\frac{C}{1-h_b(\langle P_e \rangle)} - R$, which can be broken down as sum of two 'gap's

$$\frac{C}{1 - h_b(\langle P_e \rangle)} - R = \left\{ \frac{C}{1 - h_b(\langle P_e \rangle)} - C \right\} + \{C - R\} \tag{27}$$

The first term goes to zero as $\langle P_e \rangle \to 0$ and the second term is the intuitive idea of gap to capacity.

The traditional approach of error exponents is to study the behavior as the gap is fixed and $\langle P_e \rangle \to 0$. Considering the error exponent as a function of the gap reveals something about how difficult it is to approach capacity. However, as we have seen in the previous section, our bounds predict double-exponential improvements in the probability of error with the number of iterations. In that way, our bounds share a qualitative feature with the trivial bound of Section V-A.

It turns out that the bounds of Theorems 1 and 2 do not give very interesting results if we fix $\langle P_e \rangle > 0$ and let $R \to C$. We need $\langle P_e \rangle \to 0$ alongside $R \to C$. To capture the intuitive idea of gap, which is just the second term in (27), we want to be able to assume that the effect of the second term dominates the first. This way, we can argue that the decoding complexity increases to infinity as $gap \to 0$ and not just because $\langle P_e \rangle \to 0$. For this, it suffices to consider $\langle P_e \rangle = gap^\beta$ for $\beta > 1$. Our proof actually gives a result for $\langle P_e \rangle = gap^\beta$ for any $\beta > 0$.

D. Lower bound on iterations for regular decoding in the vicinity of capacity

Theorems 1 and 2 can be expanded asymptotically in the vicinity of capacity to see the order scaling of the required neighborhood size with the gap to capacity. Essentially, this shows that the neighborhood size must grow at least proportional to $\frac{1}{gap^2}$ unless the average probability of bit error is dropping so slowly with gap that the dominant gap is actually the $\left(\frac{C}{1-h_b(\langle P_e \rangle)} - C\right)$ term in (27).

Theorem 4: For the problem as stated in Section III, we obtain the following lower bounds on the required neighborhood size n for $\langle P_e \rangle = gap^\beta$ and $gap \to 0$.

For the BSC,

- For $\beta < 1$, $n = \Omega\left(\frac{\log_2(1/gap)}{gap^{2\beta}}\right)$.
- For $\beta \geq 1$, $n = \Omega\left(\frac{\log_2(1/gap)}{gap^2}\right)$.

For the AWGN channel,

- For $\beta < 1$, $n = \Omega\left(\frac{1}{gap^{2\beta}}\right)$.
- For $\beta \geq 1$, $n = \Omega\left(\frac{1}{gap^2}\right)$.

Proof: We give the proof here in the case of the BSC with some details relegated to the Appendix. The AWGN case follows analogously, with some small modifications that are detailed in Appendix IV.

Let the code for the given BSC P have rate R. Consider BSC channels G, chosen so that $C(G) < R < C(P)$, where $C(\cdot)$ maps a BSC to its capacity in bits per channel use. Taking $\log_2(\cdot)$ on both sides of (7) (for a fixed g),

$$\log_2\left(\langle P_e\rangle_P\right) \geq \log_2\left(h_b^{-1}\left(\delta(G)\right)\right) - 1 - nD\left(g||p\right) - \epsilon\sqrt{n}\log_2\left(\frac{g(1-p)}{p(1-g)}\right). \tag{28}$$

Rewriting (28),

$$nD\left(g||p\right) + \epsilon\sqrt{n}\log_2\left(\frac{g(1-p)}{p(1-g)}\right) + \log_2\left(\langle P_e\rangle_P\right) - \log_2\left(h_b^{-1}\left(\delta(G)\right)\right) + 1 \geq 0. \tag{29}$$

This equation is quadratic in $\sqrt{n}$. The LHS potentially has two roots. If both the roots are not real, then the expression is always positive, and we get a trivial lower bound of $\sqrt{n} \geq 0$. Therefore, the cases of interest are when the two roots are real. The larger of the two roots is a lower bound on $\sqrt{n}$.

Denoting the coefficient of n by $a = D\left(g||p\right)$, that of $\sqrt{n}$ by $b = \epsilon\log_2\left(\frac{g(1-p)}{p(1-g)}\right)$, and the constant terms by $c = \log_2\left(\langle P_e\rangle_P\right) - \log_2\left(h_b^{-1}\left(\delta(G)\right)\right) + 1$ in (29), the quadratic formula then reveals

$$\sqrt{n} \geq \frac{-b + \sqrt{b^2 - 4ac}}{2a}. \tag{30}$$

Since the lower bound holds for all g satisfying $C(G) < R = C - gap$, we substitute $g^* = p + gap^r$, for some $r < 1$ and small gap. This choice is motivated by examining Figure 12. The constraint $r < 1$ is imposed because it ensures $C(g^*) < R$ for small enough gap.

Lemma 1: In the limit of $gap \to 0$, for $g^* = p + gap^r$ to satisfy $C(g^*) < R$, it suffices that r be less than 1.

Proof:

$$\begin{aligned} C(g^*) &= C(p + gap^r) \\ &= C(p) + gap^r \times C'(p) + o(gap^r) \\ &\leq C(p) - gap = R, \end{aligned}$$

for small enough gap and $r < 1$. The final inequality holds since $C(p)$ is a monotonically-decreasing concave-$\cap$ function for a BSC with $p < \frac{1}{2}$ whereas gap^r increases faster than any linear function of gap when gap is small enough. ■

In steps, we now Taylor-expand the terms on the LHS of (29) about $g = p$.

Lemma 2 (Bounds on $h_b(p)$ and $h_b^{-1}(p)$ from [39]): For all $d > 1$, and for all $x \in [0, \frac{1}{2}]$ and $y \in [0, 1]$

$$h_b(x) \geq 2x \tag{31}$$

$$h_b(x) \leq 2x^{1-1/d}d/\ln(2) \tag{32}$$

$$h_b^{-1}(y) \geq y^{\frac{d}{d-1}}\left(\frac{\ln(2)}{2d}\right)^{\frac{d}{d-1}} \tag{33}$$

$$h_b^{-1}(y) \leq \frac{1}{2}y. \tag{34}$$

Proof: See Appendix III-A. ■

Lemma 3:

$$\frac{d}{d-1}r\ \log_2\left(gap\right) - 1 + K_1 + o(1) \leq \log_2\left(h_b^{-1}\left(\delta(g^*)\right)\right) \leq r\ \log_2\left(gap\right) - 1 + K_2 + o(1) \tag{35}$$

where $K_1 = \frac{d}{d-1}\left(\log_2\left(\frac{h_b'(p)}{C(p)}\right) + \log_2\left(\frac{\ln(2)}{d}\right)\right)$ where $d > 1$ is arbitrary and $K_2 = \log_2(\frac{h_b'(p)}{C(p)})$.

Proof: See Appendix III-B. ■

Lemma 4:

$$D(g^*||p) = \frac{gap^{2r}}{2p(1-p)\ln(2)}(1 + o(1)). \tag{36}$$

Proof: See Appendix III-C. ■

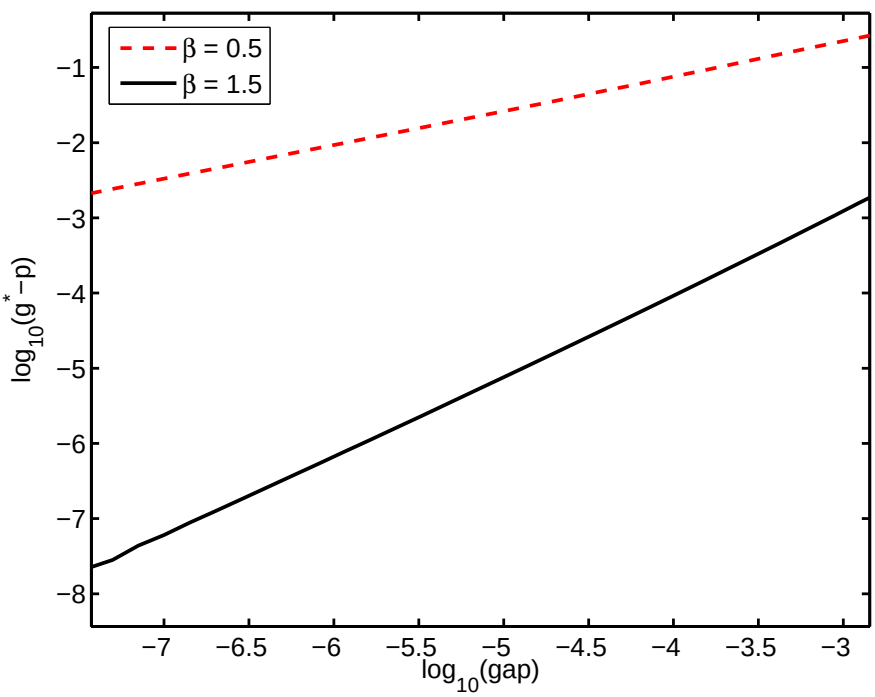

Fig. 12. The behavior of g^*, the optimizing value of g for the bound in Theorem 1, with gap. We plot $\log(g_{\text{opt}} - p)$ vs $\log(gap)$. The resulting straight lines inspired the substitution of $g^* = p + gap^r$.

Lemma 5:

$$\log_2\left(\frac{g^*(1-p)}{p(1-g^*)}\right) = \frac{gap^r}{p(1-p)\ln(2)}(1+o(1)).$$

Proof: See Appendix III-D. ■

Lemma 6:

$$\sqrt{\frac{r}{K(p)}}\sqrt{\log_2\left(\frac{1}{gap}\right)}(1+o(1)) \le \epsilon \le \sqrt{\frac{rd}{(d-1)K(p)}}\sqrt{\log_2\left(\frac{1}{gap}\right)}(1+o(1))$$

where $K(p)$ is from (10).

Proof: See Appendix III-E. ■

If $c < 0$, then the bound (30) is guaranteed to be positive. For $\langle P_e \rangle_P = gap^\beta$, the condition $c < 0$ is equivalent to

$$\beta \log_2(gap) - \log_2\left(h_b^{-1}\left(\delta(g^*)\right)\right) + 1 < 0 \tag{37}$$

Since we want (37) to be satisfied for all small enough values of gap, we can use the approximations in Lemma 3–6 and ignore constants to immediately arrive at the following sufficient condition

$$\begin{aligned} \beta \log_2(gap) - \frac{d}{d-1} r \log_2(gap) &< 0 \\ \text{i.e. } r &< \frac{\beta(d-1)}{d}, \end{aligned}$$

where d can be made arbitrarily large. Now, using the approximations in Lemma 3 and Lemma 5, and substituting them into (30), we can evaluate the solution of the quadratic equation.

As shown in Appendix III-F, this gives us the following lower bound on n.

$$n \ge \Omega\left(\frac{\log_2(1/gap)}{gap^{2r}}\right) \tag{38}$$

for any $r < \min\{\beta, 1\}$. Theorem 4 follows. ■

The lower bound on neighborhood size n can immediately be converted into a lower bound on the minimum number of computational iterations by just taking $\log_\alpha(\cdot)$. Note that this is not a comment about the degree of a potential sparse graph that defines the code. This is just about the maximum degree of the decoder's computational nodes and is a bound on the number of computational iterations required to hit the desired average probability of error.

It turns out to be easy to show that the upper bound of Theorem 3 gives rise to the same $\frac{1}{gap^2}$ scaling on the neighborhood size. This is because the random-coding error exponent in the vicinity of the capacity agrees with the sphere-packing error exponent which just has the quadratic term coming from the KL divergence. However, when we translate it from neighborhoods to iterations, the two bounds asymptotically differ by a factor of 2 that comes from (22).

The lower bounds are plotted in Figure 13 for various different values of β and reveal a $\log \frac{1}{gap}$ scaling to the required number of iterations when the decoder has bounded degree for message passing. This is much larger than the trivial lower bound of $\log\log \frac{1}{gap}$ but is much smaller than the Khandekar-McEliece conjectured $\frac{1}{gap}$ or $\frac{1}{gap}\log_2\left(\frac{1}{gap}\right)$ scaling for the number of iterations required to traverse such paths toward certainty at capacity.

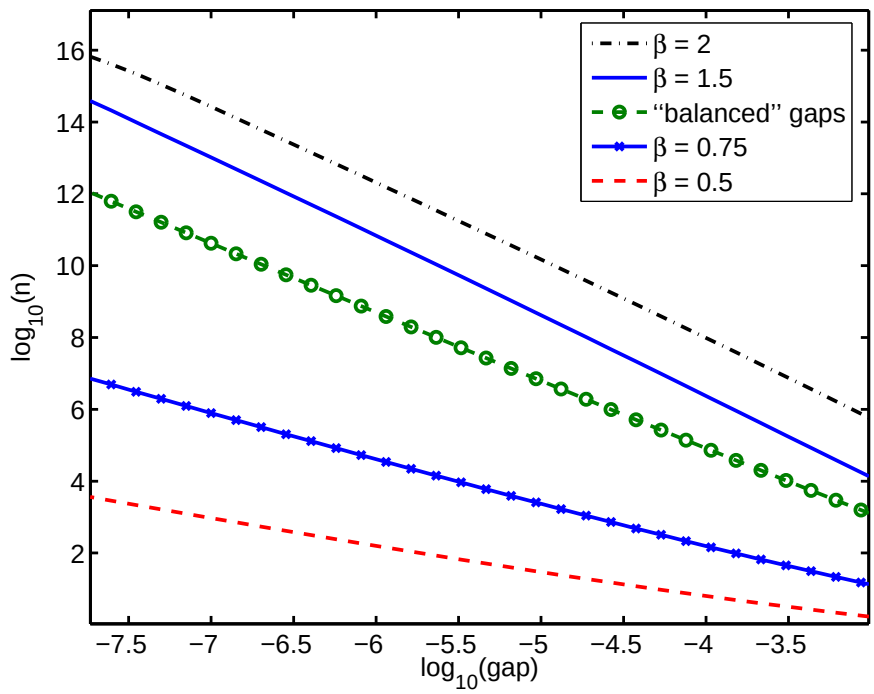

Fig. 13. Lower bounds for neighborhood size vs the gap to capacity for $\langle P_e \rangle = gap^\beta$ for various values of β. The curve titled "balanced" gaps shows the behavior for $\frac{C}{1-h_b(\langle P_e \rangle)} - C = C - R$, that is, the two 'gaps' are equal. The curves are plotted by brute-force optimization of (7), but reveal slopes that are as predicted in Theorem 4.

VI. Conclusions and Future Work

In this work, we use the inherently local nature of message-passing decoding algorithms to derive lower bounds on the number of iterations. It is interesting to note that with so few assumptions on the decoding algorithm and the code structure, the number of iterations still diverges to infinity as $gap \to 0$. As compared to [40] where a similar approach is adopted, the bounds here are stronger, and indeed tight in an order-sense for the decoding model considered. To show the tightness (in order) of these bounds, we derived corresponding upper bounds that behave similar to the lower bounds, but these exploit a loophole in our complexity model. Our model only considers the limitations induced by the internal communication structure of the decoder — it does not restrict the computational

power of the nodes within the decoder. Even so, there is still a significant gap between our upper and lower bounds in terms of the constants and we suspect this is largely related to the known looseness of the sphere-packing bound [41], as well as our coarse bounding of the required graph diameter. Our model also does not address the power requirements of encoding.

Because we assume little about the code structure, the bounds here are much more optimistic than those in [38]. However, it is unclear to what extent the optimism of our bound is an artifact. After all, [28] does get double-exponential reductions in probability of error with additional iterations, but for a family of codes that does not seem to approach capacity. This suggests that an investigation into expander codes might help resolve this question since expander codes can approach capacity, be decoded using a circuit of logarithmic depth (like our iterations), and achieve error exponents with respect to the overall block length [42]. It may very well be that expanders or expander-like codes can be shown to be weakly capacity achieving in our sense.

For any kind of capacity-achieving code, we conjecture that the optimizing transmit power will be the sum of three terms

$$P_T^* = C^{-1}(R) + \text{Tech}(\vec{\xi}, \alpha, E_{node}, R) \pm A(\langle P_e \rangle, R, \vec{\xi}, \alpha, E_{node}).$$

- $C^{-1}(R)$ is the prediction from Shannon's capacity.
- $\text{Tech}(\vec{\xi}, \alpha, E_{node}, R)$ is the minimum extra transmit power that needs to be used asymptotically to help reduce the difficulty of encoding and decoding for the given application and implementation technology. Solving (20) and subtracting $C^{-1}(R)$ gives a heuristic target value to aim for, but it remains an open problem to get a tight estimate for this term.
- $A(\langle P_e \rangle, R, \vec{\xi}, \alpha, E_{node})$ is an amount by which we should increase or reduce the transmit power because we are willing to tolerate some finite probability of error and the non-asymptotic behavior is still significant. This term should go to zero as $\langle P_e \rangle \to 0$.

Understanding the second term $\text{Tech}(\vec{\xi}, \alpha, E_{node}, R)$ above is what is needed to give principled answers regarding how close to capacity should the transmitter operate.

The results here indicate that strongly capacity-achieving coding systems are not possible if we use the given model of iterative decoding. There are a few possibilities worth exploring.

1) Our model of iterative decoding left out some real-world computational capability that could be exploited to dramatically reduce the required power consumption. There are three natural candidates here.
   - *Selective and adaptive sleep:* In the current model, all computational nodes are actively consuming power for all the iterations. If there was a way for computational nodes to adaptively turn themselves off and use **no power** while sleeping, then the results might change. We suspect that bounding the performance of such systems will require some sort of neighborhood-oriented analogies to the bounds for variable-block-length coding [43], [44].
   - *Dynamically reconfigurable circuits:* In the current model, the connectivity structure of computational nodes is fixed and considered as unchanging wiring. If there was a way for computational nodes to dynamically rewire who their neighbors are (for example by moving themselves in the combined spirit of [12], [45], [46]), this might change the results.
   - *Feedback:* In [16], a general scheme is presented that achieves an infinite computational error exponent by exploiting noiseless channel-output feedback as well as an infinite amount of common randomness. If such a scheme could be implemented, it would presumably be strongly capacity achieving as both the transmission and processing power could remain finite while having arbitrarily low average probability of bit error. However, we are unaware if either this scheme or any of the encoding strategies that claim to deliver "linear-time" encoding and decoding with an error exponent (e.g. [42], [47]) are actually implementable in a way that uses finite total power.
2) Strong or even weakly capacity-achieving communication systems may be possible using infallible computational entities but may be impossible to achieve using unreliable computational nodes that must burn more power (i.e. raise the voltages) to be more reliable [48].
3) Either strongly or weakly capacity-achieving communication systems might be impossible on thermodynamic grounds. Decoding in some abstract sense is related to the idea of cooling a part of a system [49]. Since an implementation can be considered a collection of Maxwell Daemons, this might be useful to rule out certain models of computation as being aphysical.

Finally, the approach here should be interesting if extended to a multiuser context where the prospect of causing interference makes it less easy to improve reliability by just increasing the transmit power. There, it might give some interesting answers as to what kind of computational efficiency is needed to make it asymptotically worth using multiterminal coding theory.

APPENDIX I
PROOF OF THEOREM 1: LOWER BOUND ON $\langle P_e \rangle_P$ FOR THE BSC

The idea of the proof is to first show that the average probability of error for any code must be significant if the channel were a much worse BSC. Then, a mapping is given that maps the probability of an individual error event under the worse channel to a lower-bound on its probability under the true channel. This mapping is shown to be convex-$\cup$ in the probability of error and this allows us to use this same mapping to get a lower-bound to the average probability of error under the true channel. We proceed in steps, with the lemmas proved after the main argument is complete.

Proof: Suppose we ran the given encoder and decoder over a test channel G instead.

Lemma 7 (Lower bound on $\langle P_e \rangle$ under test channel G.): If a rate-R code is used over a channel G with $C(G) < R$, then the average probability of bit error satisfies

$$\langle P_e \rangle_G \geq h_b^{-1}\left(\delta(G)\right) \tag{39}$$

where $\delta(G) = 1 - \frac{C(G)}{R}$. This holds for any channel model G, not just BSCs.

Proof: See Appendix I-A. ■

Let $\mathbf{b}_1^k$ denote the entire message, and let $\mathbf{x}_1^m$ be the corresponding codeword. Let the common randomness available to the encoder and decoder be denoted by the random variable U, and its realizations by u.

Consider the i-th message bit B_i. Its decoding is performed by observing a particular decoding neighborhood[10] of channel outputs $\mathbf{y}_{\text{nbd},i}^n$. The corresponding channel inputs are denoted by $\mathbf{x}_{\text{nbd},i}^n$, and the relevant channel noise by $\mathbf{z}_{\text{nbd},i}^n = \mathbf{x}_{\text{nbd},i}^n \oplus \mathbf{y}_{\text{nbd},i}^n$ where $\oplus$ is used to denote modulo 2 addition. The decoder just checks whether the observed $\mathbf{y}_{\text{nbd},i}^n \in \mathcal{D}_{y,i}(0,u)$ to decode to $\widehat{B}_i = 0$ or whether $\mathbf{y}_{\text{nbd},i}^n \in \mathcal{D}_{y,i}(1,u)$ to decode to $\widehat{B}_i = 1$.

For given $\mathbf{x}_{\text{nbd},i}^n$, the error event is equivalent to $\mathbf{z}_{\text{nbd},i}^n$ falling in a decoding region $\mathcal{D}_{z,i}(\mathbf{x}_{\text{nbd},i}^n, \mathbf{b}_1^k, u) = \mathcal{D}_{y,i}(1 \oplus b_i, u) \oplus \mathbf{x}_{\text{nbd},i}^n$. Thus by the linearity of expectations, (39) can be rewritten as:

$$\frac{1}{k}\sum_i \frac{1}{2^k}\sum_{\mathbf{b}_1^k}\sum_u \Pr(U=u)\Pr_G(\mathbf{Z}_{\text{nbd},i}^n \in \mathcal{D}_{z,i}(\mathbf{x}_{\text{nbd},i}^n(\mathbf{b}_1^k,u), \mathbf{b}_1^k, u)) \geq h_b^{-1}\left(\delta(G)\right). \tag{40}$$

The following lemma gives a lower bound to the probability of an event under channel P given a lower bound to its probability under channel G.

Lemma 8: Let A be a set of BSC channel-noise realizations $\mathbf{z}_1^n$ such that $\Pr_G(A) = \delta$. Then

$$\Pr_P(A) \geq f\left(\delta\right) \tag{41}$$

where

$$f(x) = \frac{x}{2}\, 2^{-nD(g\|p)} \left(\frac{p(1-g)}{g(1-p)}\right)^{\epsilon(x)\sqrt{n}} \tag{42}$$

is a convex-$\cup$ increasing function of x and

$$\epsilon(x) = \sqrt{\frac{1}{K(g)}\log_2\left(\frac{2}{x}\right)}. \tag{43}$$

Proof: See Appendix I-B. ■

Applying Lemma 8 in the style of (40) tells us that:

[10]For any given decoder implementation, the size of the decoding neighborhood might be different for different bits i. However, to avoid unnecessary complex notation, we assume that the neighborhoods are all the same size n corresponding to the largest possible neighborhood size. This can be assumed without loss of generality since smaller decoding neighborhoods can be supplemented with additional channel outputs that are ignored by the decoder.

$$\begin{aligned}\langle P_e\rangle_P &= \frac{1}{k}\sum_i \frac{1}{2^k}\sum_{\mathbf{b}_1^k}\sum_u \Pr(U=u)\Pr_P\left(\mathbf{Z}^n_{\text{nbd},i}\in\mathcal{D}_{z,i}(\mathbf{x}^n_{\text{nbd},i}(\mathbf{b}_1^k,u),\mathbf{b}_1^k,u)\right)\\ &\geq \frac{1}{k}\sum_i \frac{1}{2^k}\sum_{\mathbf{b}_1^k}\sum_u \Pr(U=u) f(\Pr_G\left(\mathbf{Z}^n_{\text{nbd},i}\in\mathcal{D}_{z,i}(\mathbf{x}^n_{\text{nbd},i}(\mathbf{b}_1^k,u),\mathbf{b}_1^k,u)\right)). \qquad (44)\end{aligned}$$

But the increasing function $f(\cdot)$ is also convex-$\cup$ and thus (44) and (40) imply that

$$\begin{aligned}\langle P_e\rangle_P &\geq f(\frac{1}{k}\sum_i \frac{1}{2^k}\sum_{\mathbf{b}_1^k}\sum_u \Pr(U=u)\Pr_G\left(\mathbf{Z}^n_{\text{nbd},i}\in\mathcal{D}_{z,i}(\mathbf{x}^n_{\text{nbd},i}(\mathbf{b}_1^k,u),\mathbf{b}_1^k,u)\right))\\ &\geq f(h_b^{-1}(\delta(G))).\end{aligned}$$

This proves Theorem 1. ■

At the cost of slightly more complicated notation, by following the techniques in [16], similar results can be proved for decoding across any discrete memoryless channel by using Hoeffding's inequality in place of the Chernoff bounds used here in the proof of Lemma 7. In place of the KL-divergence term $D(g||p)$, for a general DMC the arguments give rise to a term $\max_x D(G_x||P_x)$ that picks out the channel input letter that maximizes the divergence between the two channels' outputs. For output-symmetric channels, the combination of these terms and the outer maximization over channels G with capacity less than R will mean that the divergence term will behave like the standard sphere-packing bound when n is large. When the channel is not output symmetric (in the sense of [13]), the resulting divergence term will behave like the Haroutunian bound for fixed block-length coding over DMCs with feedback [50].

A. *Proof of Lemma 7: A lower bound on $\langle P_e\rangle_G$.*

Proof:

$$H(\mathbf{B}_1^k)-H(\mathbf{B}_1^k|\mathbf{Y}_1^m) = I(\mathbf{B}_1^k;\mathbf{Y}_1^m)\leq I(\mathbf{X}_1^m;\mathbf{Y}_1^m)\leq mC(G).$$

Since the $Ber(\frac{1}{2})$ message bits are iid, $H(\mathbf{B}_1^k)=k$. Therefore,

$$\frac{1}{k}H(\mathbf{B}_1^k|\mathbf{Y}_1^m)\geq 1-\frac{C(G)}{R}. \qquad (45)$$

Suppose the message bit sequence was decoded to be $\widehat{\mathbf{B}}_1^k$. Denote the error sequence by $\widetilde{\mathbf{B}}_1^k$. Then,

$$\mathbf{B}_1^k=\widetilde{\mathbf{B}}_1^k\oplus\widehat{\mathbf{B}}_1^k, \qquad (46)$$

where the addition $\oplus$ is modulo 2. The only complication is the possible randomization of both the encoder and decoder. However, note that even with randomization, the true message $\mathbf{B}_1^k$ is independent of $\widehat{\mathbf{B}}_1^k$ conditioned on $\mathbf{Y}_1^m$. Thus,

$$\begin{aligned}H(\widetilde{\mathbf{B}}_1^k|\mathbf{Y}_1^m) &= H(\widehat{\mathbf{B}}_1^k\oplus\mathbf{B}_1^k|\mathbf{Y}_1^m)\\ &= H(\widehat{\mathbf{B}}_1^k\oplus\mathbf{B}_1^k|\mathbf{Y}_1^m)+I(\mathbf{B}_1^k;\widehat{\mathbf{B}}_1^k|\mathbf{Y}_1^m)\\ &= H(\widehat{\mathbf{B}}_1^k\oplus\mathbf{B}_1^k|\mathbf{Y}_1^m)-H(\mathbf{B}_1^k|\mathbf{Y}_1^m,\widehat{\mathbf{B}}_1^k)+H(\mathbf{B}_1^k|\mathbf{Y}_1^m)\\ &= I(\widehat{\mathbf{B}}_1^k\oplus\mathbf{B}_1^k;\widehat{\mathbf{B}}_1^k|\mathbf{Y}_1^m)+H(\mathbf{B}_1^k|\mathbf{Y}_1^m)\\ &\geq H(\mathbf{B}_1^k|\mathbf{Y}_1^m)\\ &\geq k(1-\frac{C(G)}{R}).\end{aligned}$$

This implies

$$\frac{1}{k}\sum_{i=1}^k H(\widetilde{B}_i|\mathbf{Y}_1^m) \geq 1-\frac{C(G)}{R}. \qquad (47)$$

Since conditioning reduces entropy, $H(\widetilde{B}_i) \geq H(\widetilde{B}_i|\mathbf{Y}_1^m)$. Therefore,

$$\frac{1}{k}\sum_{i=1}^{k} H(\widetilde{B}_i) \geq 1 - \frac{C(G)}{R}. \tag{48}$$

Since $\widetilde{B}_i$ are binary random variables, $H(\widetilde{B}_i) = h_b(\langle P_{e,i}\rangle_G)$, where $h_b(\cdot)$ is the binary entropy function. Since $h_b(\cdot)$ is a concave-$\cap$ function, $h_B^{-1}(\cdot)$ is convex-$\cup$ when restricted to output values from $[0, \frac{1}{2}]$. Thus, (48) together with Jensen's inequality implies the desired result (39). ■

B. Proof of Lemma 8: a lower bound on $\langle P_{e,i}\rangle_P$ as a function of $\langle P_{e,i}\rangle_G$.

Proof: First, consider a strongly $G-$typical set of $\mathbf{z}^n_{\text{nbd},i}$, given by

$$\mathcal{T}_{\epsilon,G} = \{\mathbf{z}_1^n \; s.t. \; \sum_{i=1}^{n} z_i - ng \leq \epsilon\sqrt{n}\}. \tag{49}$$

In words, $\mathcal{T}_{\epsilon,G}$ is the set of noise sequences with weights smaller than $ng + \epsilon\sqrt{n}$. The probability of an event A can be bounded using

$$\begin{aligned}
\delta &= \Pr_G(\mathbf{Z}_1^n \in A) \\
&= \Pr_G(\mathbf{Z}_1^n \in A \cap \mathcal{T}_{\epsilon,G}) + \Pr_G(\mathbf{Z}_1^n \in A \cap \mathcal{T}^c_{\epsilon,G}) \\
&\leq \Pr_G(\mathbf{Z}_1^n \in A \cap \mathcal{T}_{\epsilon,G}) + \Pr_G(\mathbf{Z}_1^n \in \mathcal{T}^c_{\epsilon,G}).
\end{aligned}$$

Consequently,

$$\Pr_G(\mathbf{Z}_1^n \in A \cap \mathcal{T}_{\epsilon,G}) \geq \delta - \Pr_G(\mathcal{T}^c_{\epsilon,G}). \tag{50}$$

Lemma 9: The probability of the atypical set of Bernoulli-g channel noise $\{Z_i\}$ is bounded above by

$$\Pr_G\left(\frac{\sum_i^n Z_i - ng}{\sqrt{n}} > \epsilon\right) \leq 2^{-K(g)\epsilon^2} \tag{51}$$

where $K(g) = \inf\limits_{0<\eta\leq 1-g} \frac{D(g+\eta||g)}{\eta^2}$.

Proof: See Appendix I-C. ■

Choose ϵ such that

$$\begin{aligned}
2^{-K(g)\epsilon^2} &= \frac{\delta}{2} \\
\text{i.e. } \epsilon^2 &= \frac{1}{K(g)}\log_2\left(\frac{2}{\delta}\right).
\end{aligned} \tag{52}$$

Thus (50) becomes

$$\Pr_G(\mathbf{Z}_1^n \in A \cap \mathcal{T}_{\epsilon,G}) \geq \frac{\delta}{2}. \tag{53}$$

Let $n_{\mathbf{z}_1^n}$ denote the number of ones in $\mathbf{z}_1^n$. Then,

$$\Pr_G(\mathbf{Z}_1^n = \mathbf{z}_1^n) = g^{n_{\mathbf{z}_1^n}}(1-g)^{n-n_{\mathbf{z}_1^n}}. \tag{54}$$

This allows us to lower bound the probability of A under channel law P as follows:

$$\begin{aligned}
\Pr_P(\mathbf{Z}_1^n \in A) &\geq \Pr_P(\mathbf{Z}_1^n \in A \cap \mathcal{T}_{\epsilon,G}) \\
&= \sum_{\mathbf{z}_1^n \in A \cap \mathcal{T}_{\epsilon,G}} \frac{\Pr_P(\mathbf{z}_1^n)}{\Pr_G(\mathbf{z}_1^n)} \Pr_G(\mathbf{z}_1^n) \\
&= \sum_{\mathbf{z}_1^n \in A \cap \mathcal{T}_{\epsilon,G}} \frac{p^{n_{\mathbf{z}_1^n}}(1-p)^{n-n_{\mathbf{z}_1^n}}}{g^{n_{\mathbf{z}_1^n}}(1-g)^{n-n_{\mathbf{z}_1^n}}} \Pr_G(\mathbf{z}_1^n) \\
&\geq \frac{(1-p)^n}{(1-g)^n} \sum_{\mathbf{z}_1^n \in A \cap \mathcal{T}_{\epsilon,G}} \Pr_G(\mathbf{z}_1^n) \left(\frac{p(1-g)}{g(1-p)}\right)^{ng+\epsilon\sqrt{n}} \\
&= \frac{(1-p)^n}{(1-g)^n} \left(\frac{p(1-g)}{g(1-p)}\right)^{ng+\epsilon\sqrt{n}} \Pr_G(A \cap \mathcal{T}_{\epsilon,G}) \\
&\geq \frac{\delta}{2} 2^{-nD(g||p)} \left(\frac{p(1-g)}{g(1-p)}\right)^{\epsilon\sqrt{n}}.
\end{aligned}$$

This results in the desired expression:

$$f(x) = \frac{x}{2}\, 2^{-nD(g||p)} \left(\frac{p(1-g)}{g(1-p)}\right)^{\epsilon(x)\sqrt{n}}. \tag{55}$$

where $\epsilon(x) = \sqrt{\frac{1}{K(g)} \log_2\left(\frac{2}{x}\right)}$. To see the convexity of $f(x)$, it is useful to apply some substitutions. Let $c_1 = \frac{2^{-nD(g||p)}}{2} > 0$ and let $\xi = \sqrt{\frac{n}{K(g)\ln 2}} \ln(\frac{p(1-g)}{g(1-p)})$. Notice that $\xi < 0$ since the term inside the $\ln$ is less than 1. Then $f(x) = c_1 x \exp(\xi\sqrt{\ln 2 - \ln x})$.

Differentiating $f(x)$ once results in

$$f'(x) = c_1 \exp\left(\xi\sqrt{\ln(2) + \ln(\frac{1}{x})}\right)(1 + \frac{-\xi}{2\sqrt{\ln(2) + \ln(\frac{1}{x})}}). \tag{56}$$

By inspection, $f'(x) > 0$ for all $0 < x < 1$ and thus $f(x)$ is a monotonically increasing function. Differentiating $f(x)$ twice with respect to x gives

$$f''(x) = -\xi \frac{c_1 \exp\left(\xi\sqrt{\ln(2) + \ln(\frac{1}{x})}\right)}{2\sqrt{\ln(2) + \ln(\frac{1}{x})}} \left(1 + \frac{1}{2(\ln(2) + \ln(\frac{1}{x}))} - \frac{\xi}{2\sqrt{\ln(2) + \ln(\frac{1}{x})}}\right). \tag{57}$$

Since $\xi < 0$, it is evident that all the terms in (57) are strictly positive. Therefore, $f(\cdot)$ is convex-$\cup$. ■

C. Proof of Lemma 9: Bernoulli Chernoff bound

Proof: Recall that Z_i are iid Bernoulli random variables with mean $g \leq 1/2$.

$$\Pr\left(\frac{\sum_i (Z_i - g)}{\sqrt{n}} \geq \epsilon\right) = \Pr\left(\frac{\sum_i (Z_i - g)}{n} \geq \widetilde{\epsilon}\right) \tag{58}$$

where $\epsilon = \sqrt{n}\widetilde{\epsilon}$ and so $n = \epsilon^2/\widetilde{\epsilon}^2$. Therefore,

$$\Pr(\frac{\sum_i (Z_i - g)}{\sqrt{n}} \geq \epsilon) \leq [((1-g) + g\exp(s)) \times \exp(-s(g + \widetilde{\epsilon}))]^n \text{ for all } s \geq 0. \tag{59}$$

Choose s satisfying

$$\exp(-s) = \frac{g}{(1-g)} \times \left(\frac{1}{(g + \widetilde{\epsilon})} - 1\right). \tag{60}$$

It is safe to assume that $g+\widetilde{\epsilon} \leq 1$ since otherwise, the relevant probability is 0 and any bound will work. Substituting (60) into (59) gives

$$\Pr\left(\frac{\sum_i (Z_i - g)}{\sqrt{n}} \geq \epsilon\right) \quad \leq \quad 2^{-\frac{D(g+\widetilde{\epsilon}||g)}{\widetilde{\epsilon}^2}\epsilon^2}.$$

This bound holds under the constraint $\frac{\epsilon^2}{\widetilde{\epsilon}^2} = n$. To obtain a bound that holds uniformly for all n, we fix ϵ, and take the supremum over all the possible $\widetilde{\epsilon}$ values.

$$\begin{aligned}\Pr\left(\frac{\sum_i (Z_i - g)}{\sqrt{n}} \geq \epsilon\right) &\leq \sup_{0<\widetilde{\epsilon}\leq 1-g} \exp(-\ln(2)\frac{D(g+\widetilde{\epsilon}||g)}{\widetilde{\epsilon}^2}\epsilon^2) \\ &\leq \exp(-\ln(2)\epsilon^2 \inf_{0<\widetilde{\epsilon}\leq 1-g} \frac{D(g+\widetilde{\epsilon}||g)}{\widetilde{\epsilon}^2}),\end{aligned}$$

giving us the desired bound. ■

APPENDIX II
PROOF OF THEOREM 2: LOWER BOUND ON $\langle P_e \rangle_P$ FOR AWGN CHANNELS

The AWGN case can be proved using an argument almost identical to the BSC case. Once again, the focus is on the channel noise Z in the decoding neighborhoods [51]. Notice that Lemma 7 already applies to this channel even if the power constraint only has to hold on average over all codebooks and messages. Thus, all that is required is a counterpart to Lemma 8 giving a convex-$\cup$ mapping from the probability of a set of channel-noise realizations under a Gaussian channel with noise variance σ_G^2 back to their probability under the original channel with noise variance σ_P^2.

Lemma 10: Let A be a set of Gaussian channel-noise realizations $\mathbf{z}_1^n$ such that $\Pr_G(A) = \delta$. Then

$$\Pr_P(A) \geq f(\delta) \tag{61}$$

where

$$f(\delta) = \frac{\delta}{2}\ \exp(-nD(\sigma_G^2||\sigma_P^2) - \sqrt{n}(\frac{3}{2} + 2\ln\left(\frac{2}{\delta}\right))\left(\frac{\sigma_G^2}{\sigma_P^2} - 1\right)). \tag{62}$$

Furthermore, $f(x)$ is a convex-$\cup$ increasing function in δ for all values of $\sigma_G^2 \geq \sigma_P^2$.

In addition, the following bound is also convex whenever $\sigma_G^2 > \sigma_P^2\mu(n)$ with $\mu(n)$ as defined in (13).

$$f_L(\delta) = \frac{\delta}{2}\ \exp(-nD(\sigma_G^2||\sigma_P^2) - \frac{1}{2}\phi(n,\delta)\left(\frac{\sigma_G^2}{\sigma_P^2} - 1\right)) \tag{63}$$

where $\phi(n,\delta)$ is as defined in (16).

Proof: See Appendix II-A. ■

With Lemma 10 playing the role of Lemma 8, the proof for Theorem 2 proceeds identically to that of Theorem 1.

It should be clear that similar arguments can be used to prove similar results for any additive-noise models for continuous output communication channels. However, we do not believe that this will result in the best possible bounds. Instead, even the bounds for the AWGN case seem suboptimal because we are ignoring the possibility of a large deviation in the noise that happens to be locally aligned to the codeword itself.

A. *Proof of Lemma 10: a lower bound on $\langle P_{e,i} \rangle_P$ as a function of $\langle P_{e,i} \rangle_G$*

Proof: Consider the length-n set of G-typical additive noise given by

$$\mathcal{T}_{\epsilon,G} = \left\{\mathbf{z}_1^n : \frac{||\mathbf{z}_1^n||^2 - n\sigma_G^2}{n} \leq \epsilon\right\}. \tag{64}$$

With this definition, (50) continues to hold in the Gaussian case.

There are two different Gaussian counterparts to Lemma 9. They are both expressed in the following lemma.

Lemma 11: For Gaussian noise Z_i with variance σ_G^2,

$$\Pr(\frac{1}{n}\sum_{i=1}^{n}\frac{Z_i^2}{\sigma_G^2} > 1+\frac{\epsilon}{\sigma_G^2}) \leq \left((1+\frac{\epsilon}{\sigma_G^2})\exp(-\frac{\epsilon}{\sigma_G^2})\right)^{\frac{n}{2}}. \tag{65}$$

Furthermore

$$\Pr(\frac{1}{n}\sum_{i=1}^{n}\frac{Z_i^2}{\sigma_G^2} > 1+\frac{\epsilon}{\sigma_G^2}) \leq \exp(-\frac{\sqrt{n}\epsilon}{4\sigma_G^2}) \tag{66}$$

for all $\epsilon \geq \frac{3\sigma_G^2}{\sqrt{n}}$.

Proof: See Appendix II-B. ■

To have $\Pr(\mathcal{T}_{\epsilon,G}^c) \leq \frac{\delta}{2}$, it suffices to pick any $\epsilon(\delta,n)$ large enough.

So

$$\begin{aligned}\Pr_P(A) &\geq \int_{\mathbf{z}_1^n\in A\cap\mathcal{T}_{\epsilon,G}} f_P(\mathbf{z}_1^n)d\mathbf{z}_1^n \\ &= \int_{\mathbf{z}_1^n\in A\cap\mathcal{T}_{\epsilon,G}} \frac{f_P(\mathbf{z}_1^n)}{f_G(\mathbf{z}_1^n)} f_G(\mathbf{z}_1^n)d\mathbf{z}_1^n.\end{aligned} \tag{67}$$

Consider the ratio of the two pdf's for $\mathbf{z}_1^n \in \mathcal{T}_{\epsilon,G}$

$$\begin{aligned}\frac{f_P(\mathbf{z}_1^n)}{f_G(\mathbf{z}_1^n)} &= \left(\sqrt{\frac{\sigma_G^2}{\sigma_P^2}}\right)^n \exp\left(-\|\mathbf{z}_1^n\|^2\left(\frac{1}{2\sigma_P^2}-\frac{1}{2\sigma_G^2}\right)\right) \\ &\geq \exp\left(-(n\sigma_G^2+n\epsilon(\delta,n))\left(\frac{1}{2\sigma_P^2}-\frac{1}{2\sigma_G^2}\right)+n\ln\left(\frac{\sigma_G}{\sigma_P}\right)\right) \\ &= \exp\left(-\frac{\epsilon(\delta,n)n}{2\sigma_G^2}\left(\frac{\sigma_G^2}{\sigma_P^2}-1\right)-nD(\sigma_G^2||\sigma_P^2)\right)\end{aligned} \tag{68}$$

where $D(\sigma_G^2||\sigma_P^2)$ is the KL-divergence between two Gaussian distributions of variances σ_G^2 and σ_P^2 respectively. Substitute (68) back in (67) to get

$$\begin{aligned}\Pr_P(A) &\geq \exp\left(-\frac{\epsilon(\delta,n)n}{2\sigma_G^2}\left(\frac{\sigma_G^2}{\sigma_P^2}-1\right)-nD(\sigma_G^2||\sigma_P^2)\right)\int_{\mathbf{z}_1^n\in A\cap\mathcal{T}_{\epsilon,G}} f_G(\mathbf{z}_1^n)d\mathbf{z}_1^n \\ &\geq \frac{\delta}{2}\exp\left(-nD(\sigma_G^2||\sigma_P^2)-\frac{\epsilon(\delta,n)n}{2\sigma_G^2}\left(\frac{\sigma_G^2}{\sigma_P^2}-1\right)\right).\end{aligned} \tag{69}$$

At this point, it is necessary to make a choice of $\epsilon(\delta,n)$. If we are interested in studying the asymptotics as n gets large, we can use (66). This reveals that it is sufficient to choose $\epsilon \geq \sigma_G^2\max(\frac{3}{\sqrt{n}}, -4\frac{\ln(\delta)-\ln(2)}{\sqrt{n}})$. A safe bet is $\epsilon = \sigma_G^2\frac{3+4\ln(\frac{2}{\delta})}{\sqrt{n}}$ or $n\epsilon(\delta,n) = \sqrt{n}(3+4\ln(\frac{2}{\delta}))\sigma_G^2$. Thus (53) holds as well with this choice of $\epsilon(\delta,n)$.

Substituting into (69) gives

$$\Pr_P(A) \geq \frac{\delta}{2}\exp\left(-nD(\sigma_G^2||\sigma_P^2)-\sqrt{n}(\frac{3}{2}+2\ln\left(\frac{2}{\delta}\right))(\frac{\sigma_G^2}{\sigma_P^2}-1)\right).$$

This establishes the desired $f(\cdot)$ function from (62). To see that this function $f(x)$ is convex-$\cup$ and increasing in x, define $c_1 = \exp(-nD(\sigma_G^2||\sigma_P^2)-\sqrt{n}(\frac{3}{2}+2\ln(2))\left(\frac{\sigma_G^2}{\sigma_P^2}-1\right)-\ln(2))$ and $\xi = 2\sqrt{n}\left(\frac{\sigma_G^2}{\sigma_P^2}-1\right) > 0$. Then $f(\delta) = c_1\delta\exp(\xi\ln(\delta)) = c_1\delta^{1+\xi}$ which is clearly monotonically increasing and convex-$\cup$ by inspection.

Attempting to use (65) is a little more involved. Let $\widetilde{\epsilon} = \frac{\epsilon}{\sigma_G^2}$ for notational convenience. Then we must solve $(1+\widetilde{\epsilon})\exp(-\widetilde{\epsilon}) = (\frac{\delta}{2})^{\frac{2}{n}}$. Substitute $u = 1+\widetilde{\epsilon}$ to get $u\exp(-u+1) = (\frac{\delta}{2})^{\frac{2}{n}}$. This immediately simplifies to $-u\exp(-u) = -\exp(-1)(\frac{\delta}{2})^{\frac{2}{n}}$. At this point, we can immediately verify that $(\frac{\delta}{2})^{\frac{2}{n}} \in [0,1]$ and hence by the definition of the Lambert W function in [34], we get $u = -W_L(-\exp(-1)(\frac{\delta}{2})^{\frac{2}{n}})$. Thus

$$\widetilde{\epsilon}(\delta,n) = -W_L(-\exp(-1)(\frac{\delta}{2})^{\frac{2}{n}})-1. \tag{70}$$

Substituting this into (69) immediately gives the desired expression (63). All that remains is to verify the convexity. Let $v = \frac{1}{2}\left(\frac{\sigma_G^2}{\sigma_P^2} - 1\right)$. As above, $f_L(\delta) = \delta c_2 \exp(-nv\widetilde{\epsilon}(\delta, n))$. The derivatives can be taken using very tedious manipulations involving the relationship $W_L'(x) = \frac{W_L(x)}{x(1+W_L(x))}$ from [34] and can be verified using computer-aided symbolic calculation. In our case $-\widetilde{\epsilon}(\delta, n) = (W_L(x) + 1)$ and so this allows the expressions to be simplified.

$$f_L'(\delta) = c_2 \exp(-nv\widetilde{\epsilon})(2v + 1 + \frac{2v}{\widetilde{\epsilon}}). \tag{71}$$

Notice that all the terms above are positive and so the first derivative is always positive and the function is increasing in δ. Taking another derivative gives

$$f_L''(\delta) = c_2 \frac{2v(1+\widetilde{\epsilon})\exp(-nv\widetilde{\epsilon})}{\delta\widetilde{\epsilon}}\left[1 + 4v + \frac{4v}{\widetilde{\epsilon}} - \frac{4}{n\widetilde{\epsilon}} - \frac{2}{n\widetilde{\epsilon}^2}\right]. \tag{72}$$

Recall from (70) and the properties of the Lambert W_L function that $\widetilde{\epsilon}$ is a monotonically decreasing function of δ that is $+\infty$ when $\delta = 0$ and goes down to 0 at $\delta = 2$. Look at the term in brackets above and multiply it by the positive $n\widetilde{\epsilon}^2$. This gives the quadratic expression

$$(4v + 1)n\widetilde{\epsilon}^2 + 4(vn - 1)\widetilde{\epsilon} - 2. \tag{73}$$

This (73) is clearly convex-$\cup$ in $\widetilde{\epsilon}$ and negative at $\widetilde{\epsilon} = 0$. Thus it must have a single zero-crossing for positive $\widetilde{\epsilon}$ and be strictly increasing there. This also means that the quadratic expression is implicitly a strictly *decreasing* function of δ. It thus suffices to just check the quadratic expression at $\delta = 1$ and make sure that it is non-negative. Evaluating (70) at $\delta = 1$ gives $\widetilde{\epsilon}(1, n) = T(n)$ where $T(n)$ is defined in (14).

It is also clear that (73) is a strictly increasing linear function of v and so we can find the minimum value for v above which (73) is guaranteed to be non-negative. This will guarantee that the function f_L is convex-$\cup$. The condition turns out to be $v \geq \frac{2+4T-nT^2}{4nT(T+1)}$ and hence $\sigma_G^2 = \sigma_P^2(2v+1) \geq \frac{\sigma_G^2}{2}(1 + \frac{1}{T+1} + \frac{4T+2}{nT(T+1)})$. This matches up with (13) and hence the Lemma is proved. ■

B. Proof of Lemma 11: Chernoff bound for Gaussian noise

Proof: The sum $\sum_{i=1}^n \frac{Z_i^2}{\sigma_G^2}$ is a standard χ^2 random variables with n degrees of freedom.

$$\begin{aligned}
& \Pr\left(\frac{1}{n}\sum_{i=1}^n \frac{Z_i^2}{\sigma_G^2} > 1 + \frac{\epsilon}{\sigma_G^2}\right) \\
\leq_{(a)} \quad & \inf_{s>0}\left(\frac{\exp(-s(1+\frac{\epsilon}{\sigma_G^2}))}{\sqrt{1-2s}}\right)^n \\
\leq_{(b)} \quad & \left(\sqrt{1+\frac{\epsilon}{\sigma_G^2}}\exp(-\frac{\epsilon}{2\sigma_G^2})\right)^n && (74) \\
= \quad & \left((1+\frac{\epsilon}{\sigma_G^2})\exp(-\frac{\epsilon}{\sigma_G^2})\right)^{\frac{n}{2}} && (75)
\end{aligned}$$

where (a) follows using standard moment generating functions for χ^2 random variables and Chernoff bounding arguments and (b) results from the substitution $s = \frac{\epsilon}{2(\sigma_G^2+\epsilon)}$. This establishes (65).

For tractability, the goal is to replace (74) with a exponential of an affine function of $\frac{\epsilon}{\sigma_G^2}$. For notational convenience, let $\widetilde{\epsilon} = \frac{\epsilon}{\sigma_G^2}$. The idea is to bound the polynomial term $\sqrt{1+\widetilde{\epsilon}}$ with an exponential as long as $\widetilde{\epsilon} > \epsilon^*$.

Let $\epsilon^* = \frac{3}{\sqrt{n}}$ and let $K = \frac{1}{2} - \frac{1}{4\sqrt{n}}$. Then it is clear that

$$\sqrt{1+\widetilde{\epsilon}} \leq \exp(K\widetilde{\epsilon}) \tag{76}$$

as long as $\widetilde{\epsilon} \geq \epsilon^*$. First, notice that the two agree at $\widetilde{\epsilon} = 0$ and that the slope of the concave-$\cap$ function $\sqrt{1+\widetilde{\epsilon}}$ there is $\frac{1}{2}$. Meanwhile, the slope of the convex-$\cup$ function $\exp(K\widetilde{\epsilon})$ at 0 is $K < \frac{1}{2}$. This means that $\exp(K\widetilde{\epsilon})$ starts out below $\sqrt{1+\widetilde{\epsilon}}$. However, it has crossed to the other side by $\widetilde{\epsilon} = \epsilon^*$. This can be verified by taking the logs

of both sides of (76) and multiplying them both by 2. Consider the LHS evaluated at ϵ^* and lower-bound it by a third-order power-series expansion

$$\ln(1+\frac{3}{\sqrt{n}}) \quad \leq \quad \frac{3}{\sqrt{n}} - \frac{9}{2n} + \frac{9}{n^{3/2}}.$$

meanwhile the RHS of (76) can be dealt with exactly:

$$\begin{aligned} 2K\epsilon^* &= (1-\frac{1}{2\sqrt{n}})\frac{3}{\sqrt{n}} \\ &= \frac{3}{\sqrt{n}} - \frac{3}{2n}. \end{aligned}$$

For $n \geq 9$, the above immediately establishes (76) since $\frac{9}{2n} - \frac{3}{2n} = \frac{3}{n} \geq \frac{9}{n\sqrt{9}}$. The cases $n = 1, 2, 3, 4, 5, 6, 7, 8$ can be verified by direct computation. Using (76), for $\widetilde{\epsilon} > \epsilon^*$ we have:

$$\begin{aligned} \Pr(\mathcal{T}^c_{\epsilon,G}) &\leq [\exp(K\widetilde{\epsilon})\exp(-\frac{1}{2}\widetilde{\epsilon})]^n \\ &= \exp(-\frac{\sqrt{n}}{4}\widetilde{\epsilon}). \end{aligned} \tag{77}$$

■

APPENDIX III
APPROXIMATION ANALYSIS FOR THE BSC

A. Lemma 2

Proof: (31) and (34) are obvious from the concave-$\cap$ nature of the binary entropy function and its values at 0 and $\frac{1}{2}$.

$$\begin{aligned} h_b(x) &= x\log_2(1/x) + (1-x)\log_2(1/(1-x)) \\ &\leq_{(a)} 2x\log_2(1/x) = 2x\ln(1/x)/\ln(2) \\ &\leq_{(b)} 2xd(\frac{1}{x^{1/d}} - 1)/\ln(2) \quad \forall d > 1 \\ &\leq 2x^{1-1/d}d/\ln(2). \end{aligned}$$

Inequality (a) follows from the fact that $x^x < (1-x)^{1-x}$ for $x \in (0, \frac{1}{2})$. For inequality (b), observe that $\ln(x) \leq x - 1$. This implies $\ln(x^{1/d}) \leq x^{1/d} - 1$. Therefore, $\ln(x) \leq d(x^{1/d} - 1)$ for all $x > 0$ since $\frac{1}{d} \leq 1$ for $d \geq 1$.

The bound on $h_b^{-1}(x)$ follows immediately by identical arguments. ■

B. Lemma 3

Proof:

First, we investigate the small gap asymptotics for $\delta(g^*)$, where $g^* = p + gap^r$ and $r < 1$.

$$\begin{aligned} \delta(g^*) &= 1 - \frac{C(g^*)}{R} \\ &= 1 - \frac{C(p+gap^r)}{C(p) - gap} \\ &= 1 - \frac{C(p) - gap^r h_b'(p) + o(gap^r)}{C(p)(1 - gap/C(p))} \\ &= 1 - (1 - \frac{h_b'(p)}{C(p)}gap^r + o(gap^r)) \times (1 + gap/C(p) + o(gap)) \\ &= \frac{h_b'(p)}{C(p)}gap^r + o(gap^r). \end{aligned} \tag{78}$$

Plugging (78) into (34) and using Lemma 2 gives

$$\log_2\left(h_b^{-1}\left(\delta(g^*)\right)\right) \leq \log_2\left(\frac{h_b'(p)}{2C(p)}gap^r + o(gap^r)\right) \tag{79}$$

$$= \log_2\left(\frac{h_b'(p)}{2C(p)}\right) + r\log_2(gap) + o(1) \tag{80}$$

$$= r\log_2(gap) - 1 + \log_2\left(\frac{h_b'(p)}{C(p)}\right) + o(1) \tag{81}$$

and this establishes the upper half of (35).

To see the lower half, we use (33):

$$\begin{aligned}
\log_2\left(h_b^{-1}\left(\delta(g^*)\right)\right) &\geq \frac{d}{d-1}\left(\log_2\left(\delta(g^*)\right) + \log_2\left(\frac{\ln 2}{2d}\right)\right) \\
&= \frac{d}{d-1}\left(\log_2\left(\frac{h_b'(p)}{C(p)}gap^r + o(gap^r)\right) + \log_2\left(\frac{\ln 2}{2d}\right)\right) \\
&= \frac{d}{d-1}\left(r\log_2(gap) + \log_2\left(\frac{h_b'(p)}{C(p)}\right) + o(1) + \log_2\left(\frac{\ln 2}{2d}\right)\right) \\
&= \frac{d}{d-1}r\ \log_2(gap) - 1 + K_1 + o(1)
\end{aligned}$$

where $K_1 = \frac{d}{d-1}\left(\log_2\left(\frac{h_b'(p)}{C(p)}\right) + \log_2\left(\frac{\ln(2)}{d}\right)\right)$ and $d > 1$ is arbitrary. ■

C. Lemma 4

Proof:

$$\begin{aligned}
D(g^*||p) &= D(p + gap^r||p) \\
&= 0 + 0 \times gap^r + \frac{1}{2}\frac{gap^{2r}}{p(1-p)\ln(2)} + o(gap^{2r})
\end{aligned}$$

since $D(p||p) = 0$ and the first derivative is also zero. Simple calculus shows that the second derivative of $D(p+x||p)$ with respect to x is $\frac{\log_2(e)}{(p+x)(1-p-x)}$. ■

D. Lemma 5

Proof:

$$\begin{aligned}
\log_2\left(\frac{g^*(1-p)}{p(1-g^*)}\right) &= \log_2\left(\frac{1-p}{p}\right) + \log_2\left(\frac{g^*}{1-g^*}\right) \\
&= \log_2\left(\frac{1-p}{p}\right) + \log_2(g^*) - \log_2(1-g^*) \\
&= \log_2\left(\frac{1-p}{p}\right) + \log_2(p + gap^r) - \log_2(1-p-gap^r) \\
&= \log_2\left(\frac{1-p}{p}\right) + \log_2(p) + \log_2\left(1 + \frac{gap^r}{p}\right) - \log_2(1-p) - \log_2\left(1 - \frac{gap^r}{1-p}\right) \\
&= \frac{gap^r}{p\ln(2)} + \frac{gap^r}{(1-p)\ln(2)} + o(gap^r) \\
&= \frac{gap^r}{p(1-p)\ln(2)} + o(gap^r) \\
&= \frac{gap^r}{p(1-p)\ln(2)}(1 + o(1)).
\end{aligned}$$

■

E. Lemma 6

Proof: Expand (9):

$$\begin{aligned}
\epsilon &= \sqrt{\frac{1}{K(p+gap^r)}}\sqrt{\log_2\left(\frac{2}{h_b^{-1}(\delta(G))}\right)} \\
&= \sqrt{\frac{1}{\ln(2)K(p+gap^r)}}\sqrt{\ln(2)-\ln(h_b^{-1}(\delta(G)))} \\
&\geq \sqrt{\frac{1}{\ln(2)K(p+gap^r)}}\sqrt{\ln(2)-r\ln(gap)+\ln(2)-K_2\ln(2)+o(1)} \\
&= \sqrt{\frac{1}{\ln(2)K(p+gap^r)}}\sqrt{r\ln(\frac{1}{gap})+(2-K_2)\ln(2)+o(1)} \\
&= \sqrt{\frac{1}{\ln(2)K(p+gap^r)}}\sqrt{r\ln(\frac{1}{gap})}(1+o(1)).
\end{aligned}$$

and similarly

$$\begin{aligned}
\epsilon &= \sqrt{\frac{1}{\ln(2)K(p+gap^r)}}\sqrt{\ln(2)-\ln(h_b^{-1}(\delta(G)))} \\
&\leq \sqrt{\frac{1}{\ln(2)K(p+gap^r)}}\sqrt{(2-K_2)\ln(2)+\frac{d}{d-1}r\ln\left(\frac{1}{gap}\right)+o(1)} \\
&= \sqrt{\frac{rd}{\ln(2)(d-1)K(p+gap^r)}}\sqrt{\ln\left(\frac{1}{gap}\right)}(1+o(1)).
\end{aligned}$$

All that remains is to show that $K(p+gap^r)$ converges to $K(p)$ as $gap \to 0$. Examine (10). The continuity of $\frac{D(g+\eta||g)}{\eta^2}$ is clear in the interior $\eta \in (0, 1-g)$ and for $g \in (0, \frac{1}{2})$. All that remains is to check the two boundaries. $\lim_{\eta\to 0}\frac{D(g+\eta||g)}{\eta^2} = \frac{1}{g(1-g)\ln 2}$ by the Taylor expansion of $D(g+\eta||g)$ as done in the proof of Lemma 4. Similarly, $\lim_{\eta\to 1-g}\frac{D(g+\eta||g)}{\eta^2} = D(1||g) = \log_2\left(\frac{1}{1-g}\right)$. Since K is a minimization of a continuous function over a compact set, it is itself continuous and thus the limit $\lim_{gap\to 0} K(p+gap^r) = K(p)$.

Converting from natural logarithms to base 2 completes the proof. ■

F. Approximating the solution to the quadratic formula

In (30), for $g = g^* = p + gap^r$,

$$\begin{aligned}
a &= D(g^*||p) \\
b &= \epsilon\log_2\left(\frac{g^*(1-p)}{p(1-g^*)}\right) \\
c &= \log_2\left(\langle P_e\rangle_P\right) - \log_2\left(h_b^{-1}(\delta(g^*))\right) + 1.
\end{aligned}$$

The first term, a, is approximated by Lemma 4 so

$$a = gap^{2r}(\frac{1}{2p(1-p)\ln(2)} + o(1)). \tag{82}$$

Applying Lemma 5 and Lemma 6 reveals

$$\begin{aligned} b &\leq \sqrt{\frac{rd}{(d-1)K(p)}}\sqrt{\log_2\left(\frac{1}{gap}\right)}\frac{gap^r}{p(1-p)\ln(2)}(1+o(1)) \\ &= \frac{1}{p(1-p)\ln(2)}\sqrt{\frac{rd}{(d-1)K(p)}}\sqrt{gap^{2r}\log_2\left(\frac{1}{gap}\right)}(1+o(1)) \qquad (83) \\ b &\geq \frac{1}{p(1-p)\ln(2)}\sqrt{\frac{r}{K(p)}}\sqrt{gap^{2r}\log_2\left(\frac{1}{gap}\right)}(1+o(1)). \qquad (84) \end{aligned}$$

The third term, c, can be bounded similarly using Lemma 3 as follows,

$$\begin{aligned} c &= \beta\log_2(gap) - \log_2\left(h_b^{-1}(\delta(g^*))\right) + 1 \\ &\leq (\frac{d}{d-1}r - \beta)\log_2\left(\frac{1}{gap}\right) + K_3 + o(1) \qquad (85) \\ &\geq (r-\beta)\log_2\left(\frac{1}{gap}\right) + K_4 + o(1). \qquad (86) \end{aligned}$$

for a pair of constants K_3, K_4. Thus, for gap small enough and $r < \frac{\beta(d-1)}{d}$, we know that $c < 0$.

The lower bound on $\sqrt{n}$ is thus

$$\begin{aligned} \sqrt{n} &\geq \frac{\sqrt{b^2-4ac}-b}{2a} \\ &= \frac{b}{2a}\left(\sqrt{1-\frac{4ac}{b^2}}-1\right). \qquad (87) \end{aligned}$$

Plugging in the bounds (82) and (84) reveals that

$$\frac{b}{2a} \geq \frac{\sqrt{\log_2\left(\frac{1}{gap}\right)}}{gap^r}\sqrt{\frac{r}{K(p)}}(1+o(1)) \qquad (88)$$

Similarly, using (82), (84), (85), we get

$$\begin{aligned} \frac{4ac}{b^2} &\leq \frac{4gap^{2r}\left(\frac{1}{p(1-p)\ln(2)}\right)\times\left[(\frac{d}{d-1}r-\beta)\log_2\left(\frac{1}{gap}\right)+K_3\right](1+o(1))}{(\frac{1}{p(1-p)\ln(2)})^2\frac{r}{K(p)}gap^{2r}\log_2\left(\frac{1}{gap}\right)(1+o(1))} \\ &= 4p(1-p)K(p)\ln(2)\left[\frac{d}{d-1}-\frac{\beta}{r}\right]+o(1). \qquad (89) \end{aligned}$$

This tends to a negative constant since $r < \frac{\beta(d-1)}{d}$.

Plugging (88) and (89) into (87) gives:

$$\begin{aligned} n &\geq [\sqrt{\frac{r}{K(p)}}\frac{\log_2\left(\frac{1}{gap}\right)}{gap^r}(1+o(1))\left(\sqrt{1+4p(1-p)\ln(2)K(p)\left[\frac{\beta}{r}-\frac{d}{d-1}\right]+o(1)}-1\right)]^2 \\ &= [\frac{\log_2\left(\frac{1}{gap}\right)}{gap^r}]^2\frac{1}{K(p)}\left(\sqrt{r+4p(1-p)\ln(2)K(p)\left[\beta-\frac{rd}{d-1}\right]}-\sqrt{r}\right)^2(1+o(1)) \\ &= \Omega\left(\frac{(\log_2(1/gap))^2}{gap^{2r}}\right) \qquad (90) \end{aligned}$$

for all $r \leq \min\{\frac{\beta}{\frac{d}{d-1}}, 1\}$. By taking d arbitrarily large, we arrive at Theorem 4 for the BSC.

APPENDIX IV
APPROXIMATION ANALYSIS FOR THE AWGN CHANNEL

Taking logs on both sides of (11) for a fixed test channel G,

$$\ln(\langle P_e\rangle_P) \geq \ln(h_b^{-1}(\delta(G))) - \ln(2) - nD(\sigma_G^2||\sigma_P^2) - \sqrt{n}(\frac{3}{2} + 2\ln 2 - 2\ln(h_b^{-1}(\delta(G)))\left(\frac{\sigma_G^2}{\sigma_P^2} - 1\right), \tag{91}$$

Rewriting this in the standard quadratic form using

$$a = D(\sigma_{G^*}^2||\sigma_P^2), \tag{92}$$

$$b = (\frac{3}{2} + 2\ln 2 - 2\ln(h_b^{-1}(\delta(G))))\left(\frac{\sigma_G^2}{\sigma_P^2} - 1\right), \tag{93}$$

$$c = \ln(\langle P_e\rangle_P) - \ln(h_b^{-1}(\delta(G))) + \ln(2). \tag{94}$$

it suffices to show that the terms exhibit behavior as $gap \to 0$ similar to their BSC counterparts.

For Taylor approximations, we use the channel G^*, with corresponding noise variance $\sigma_{G^*}^2 = \sigma_P^2 + \zeta$, where

$$\zeta = gap^r\left(\frac{2\sigma_P^2(P_T + \sigma_P^2)}{P_T}\right). \tag{95}$$

Lemma 12: For small enough gap, for ζ as in (95), if $r < 1$ then $C(G^*) < R$.

Proof: Since $C(P) - gap = R > C(G^*)$, we must satisfy

$$gap \leq \frac{1}{2}\log_2\left(1 + \frac{P_T}{\sigma_P^2}\right) - \frac{1}{2}\log_2\left(1 + \frac{P_T}{\sigma_P^2 + \zeta}\right).$$

So the goal is to lower bound the RHS above to show that (95) is good enough to guarantee that this is bigger than the gap. So

$$\begin{aligned}
&= \frac{1}{2}\left(\log_2\left(1 + \frac{\zeta}{\sigma_P^2}\right) - \log_2\left(1 + \frac{\zeta}{\sigma_P^2 + P_T}\right)\right) \\
&= \frac{1}{2}\left(\log_2\left(1 + 2gap^r(1 + \frac{\sigma_P^2}{P_T})\right) - \log_2\left(1 + 2gap^r\frac{\sigma_P^2}{P_T}\right)\right) \\
&\geq \frac{1}{2}\left(\frac{c_s}{\ln(2)}2gap^r(1 + \frac{\sigma_P^2}{P_T}) - \frac{1}{\ln(2)}2gap^r\frac{\sigma_P^2}{P_T}\right) \\
&= gap^r\frac{1}{\ln(2)}\left(c_s - (1 - c_s)\frac{\sigma_P^2}{P_T}\right).
\end{aligned} \tag{96}$$

For small enough gap, this is a valid lower bound as long as $c_s < 1$. Choose c_s so that $1 < c_s < \frac{\sigma_P^2}{P_T + \sigma_P^2}$. For ζ as in (95), the LHS is $gap^r K$ and thus clearly having $r < 1$ suffices for satisfying (96) for small enough gap. This is because the derivative of gap^r tends to infinity as $gap \to 0$. ■

In the next Lemma, we perform the approximation analysis for the terms inside (92), (93) and (94).

Lemma 13: Assume that $\sigma_{G^*}^2 = \sigma_P^2 + \zeta$ where ζ is defined in (95).

(a)

$$\frac{\sigma_{G^*}^2}{\sigma_P^2} - 1 = gap^r\left(\frac{2(P_T + \sigma_P^2)}{P_T}\right). \tag{97}$$

(b)

$$\ln(\delta(G^*)) = r\ln(gap) + o(1) - \ln(C(P)). \tag{98}$$

(c)

$$\ln(h_b^{-1}(\delta(G^*))) \geq \frac{d}{d-1}r\ln(gap) + c_2, \tag{99}$$

for some constant c_2 that is a function of d.

$$\ln(h_b^{-1}(\delta(G^*))) \leq r\ln(gap) + c_3, \tag{100}$$

for some constant c_3.

(d)

$$D(\sigma^2_{G^*}||\sigma^2_P) = \frac{(P_T+\sigma^2_P)^2}{P_T^2} gap^{2r}(1+o(1)). \tag{101}$$

Proof: *(a)* Immediately follows from the definitions and (95).

(b) We start with simplifying $\delta(G^*)$

$$\begin{aligned}
\delta(G^*) &= 1-\frac{C(G^*)}{R} \\
&= \frac{C-gap-\frac{1}{2}\log_2\left(1+\frac{P_T}{\sigma^2_{G^*}}\right)}{C-gap} \\
&= \frac{\frac{1}{2}\log_2\left(1+\frac{P_T}{\sigma^2_P}\right)-\frac{1}{2}\log_2\left(1+\frac{P_T}{\sigma^2_P+\zeta}\right)-gap}{C-gap} \\
&= \frac{\frac{1}{2}\log_2\left((\frac{\sigma^2_P+P_T}{\sigma^2_P})(\frac{\sigma^2_P+\zeta}{P_T+\sigma^2_P+\zeta})\right)-gap}{C-gap} \\
&= \frac{\frac{1}{2}\log_2\left(1+\frac{\zeta}{\sigma^2_P}\right)-\frac{1}{2}\log_2\left(1+\frac{\zeta}{P_T+\sigma^2_P}\right)-gap}{C-gap} \\
&= \frac{\frac{1}{2}\frac{\zeta}{\sigma^2_P}-\frac{1}{2}\frac{\zeta}{P_T+\sigma^2_P}+o(\zeta)-gap}{C-gap} \\
&= \frac{\frac{1}{2}\left(\frac{\zeta P_T}{\sigma^2_P(P_T+\sigma^2_P)}+o(\zeta)\right)-gap}{C-gap} \\
&= \frac{1}{C}(\frac{1}{2}\left(gap^r\frac{2\sigma^2_P(P_T+\sigma^2_P)}{P_T}\frac{P_T}{\sigma^2_P(P+\sigma^2_P)}+o(gap^r)\right)-gap)(1-\frac{gap}{C}+o(gap)) \\
&= \frac{gap^r}{C}(1+o(1)).
\end{aligned}$$

Taking $\ln(\cdot)$ on both sides, the result is evident.

(c) follows from *(b)* and Lemma 2.

(d) comes from the definition of $D(\sigma^2_{G^*}||\sigma^2_P)$ followed immediately by the expansion $\ln(\sigma^2_{G^*}/\sigma^2_P) = \ln(1+\zeta/\sigma^2_P) = \frac{\zeta}{\sigma^2_P}-\frac{1}{2}(\frac{\zeta}{\sigma^2_P})^2+o(gap^{2r})$. All the constant and first-order in gap^r terms cancel since $\frac{\sigma^2_{G^*}}{\sigma^2_P} = 1+\frac{\zeta}{\sigma^2_P}$. This gives the result immediately. ■

Now, we can use Lemma 13 to approximate (92), (93) and (94).

$$a = \frac{(P_T+\sigma^2_P)^2}{P_T^2} gap^{2r}(1+o(1)) \tag{102}$$

$$\begin{aligned}
b &= \left(\frac{3}{2}+2\ln 2-2\ln(h_b^{-1}(\delta(G)))\right) gap^r \frac{2(P_T+\sigma^2_P)}{P_T} \\
&\leq \frac{2d(P_T+\sigma^2_P)}{(d-1)P_T} r\ln(\frac{1}{gap}) gap^r(1+o(1))
\end{aligned} \tag{103}$$

$$b \geq \frac{2(P_T+\sigma^2_P)}{P_T} r\ln(\frac{1}{gap}) gap^r(1+o(1)) \tag{104}$$

$$c \leq (\frac{d}{d-1}r-\beta)\ln(\frac{1}{gap})(1+o(1)) \tag{105}$$

$$c \geq (r-\beta)\ln(\frac{1}{gap})(1+o(1)). \tag{106}$$

Therefore, in parallel to (88), we have for the AWGN bound

$$\frac{b}{2a} \geq \frac{rP_T}{(P_T+\sigma_P^2)}\left(\frac{\ln(\frac{1}{gap})}{gap^r}\right)(1+o(1)). \tag{107}$$

Similarly, in parallel to (89), we have for the AWGN bound

$$\frac{4ac}{b^2} \quad \leq \quad (1+o(1))\frac{1}{r^2}(\frac{d}{d-1}r-\beta)\frac{1}{\ln(\frac{1}{gap})}.$$

This is negative as long as $r < \frac{\beta(d-1)}{d}$, and so for every $c_S < \frac{1}{2}$ for small enough gap, we know that

$$\sqrt{1-\frac{4ac}{b^2}}-1 \quad \geq \quad c_s\frac{1}{r^2}(\beta-\frac{d}{d-1}r)\frac{1}{\ln(\frac{1}{gap})}(1+o(1)).$$

Combining this with (107) gives the bound:

$$n \quad \geq \quad (1+o(1))[c_s\frac{1}{r^2}(\beta-\frac{d}{d-1}r)\frac{1}{\ln(\frac{1}{gap})}\frac{rP_T}{P_T+\sigma_P^2}\left(\frac{\ln(\frac{1}{gap})}{gap^r}\right)]^2 \tag{108}$$

$$= \quad (1+o(1))[c_s\frac{P_T}{r(P_T+\sigma_P^2)}(\beta-\frac{d}{d-1}r)\left(\frac{1}{gap^r}\right)]^2. \tag{109}$$

Since this holds for all $0 < c_s < \frac{1}{2}$ and all $r < \min(1, \frac{\beta(d-1)}{d})$ for all $d > 1$, Theorem 4 for AWGN channels follows.

Acknowledgments

Years of conversations with colleagues in the Berkeley Wireless Research Center have helped motivate this investigation and informed the perspective here. Cheng Chang was involved with the discussions related to this paper, especially as regards the AWGN case. Sae-Young Chung (KAIST) gave valuable feedback at an early stage of this research and Hari Palaiyanur caught many typos in early drafts of this manuscript. Funding support from NSF CCF 0729122, NSF ITR 0326503, NSF CNS 0403427, and gifts from Sumitomo Electric.

References

[1] P. Grover and A. Sahai, "A general lower bound on the decoding complexity of sparse-graph codes," in *Submitted to the IEEE Workshop on Information Theory*, Porto, Portugal, May 2008.

[2] J. Massey, "Deep-space communications and coding: A marriage made in heaven," in *Advanced Methods for Satellite and Deep Space Communications: Lecture Notes in Control and Information Sciences 182*, J. Hagenauer, Ed. New York, NY: Springer, 1992, pp. 1–17.

[3] R. J. McEliece, *Are there turbo-codes on Mars?*, Chicago, IL, Jun. 2004.

[4] S. L. Howard, C. Schlegel, and K. Iniewski, "Error control coding in low-power wireless sensor networks: when is ECC energy-efficient?" *EURASIP Journal on Wireless Communications and Networking*, pp. 1–14, 2006.

[5] Y. M. Chee, C. J. Colbourn, and A. C. H. Ling, "Optimal memoryless encoding for low power off-chip data buses," 2007. [Online]. Available: doi:10.1145/1233501.1233575

[6] P. Agrawal, "Energy efficient protocols for wireless systems," in *IEEE International Symposium on Personal, Indoor, Mobile Radio Communication*, 1998, pp. 564–569.

[7] S Cui, AJ Goldsmith and A Bahai, "Energy Constrained Modulation Optimization," *IEEE Trans. Wireless Commun.*, vol. 4, no. 5, pp. 1–11, 2005.

[8] A. J. Goldsmith and S. B. Wicker, "Design challenges for energy constrained ad hoc wireless networks," *IEEE Trans. Wireless Commun.*, pp. 8–27, 2002.

[9] P. Massaad, M. Medard, and L. Zheng, "Impact of Processing Energy on the Capacity of Wireless Channels," in *International Symposium on Information Theory and its Applications (ISITA)*, 2004.

[10] S. Vasudevan, C. Zhang, D. Goeckel, and D. Towsley, "Optimal power allocation in wireless networks with transmitter-receiver power tradeoffs," *Proceedings of the 25th IEEE International Conference on Computer Communications INFOCOM*, pp. 1–11, Apr. 2006.

[11] P. J. M. Havinga and G. J. M. Smit, "Minimizing energy consumption for wireless computers in Moby Dick," in *IEEE International Conference on Personal Wireless Communications*, 1997, pp. 306–310.

[12] M. Bhardwaj and A. Chandrakasan, "Coding under observation constraints," in *Proceedings of the Allerton Conference on Communication, Control, and Computing*, Monticello, IL, Sep. 2007.

[13] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY: John Wiley, 1971.

[14] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic Press, 1981.

[15] M. S. Pinsker, "Bounds on the probability and of the number of correctable errors for nonblock codes," *Problemy Peredachi Informatsii*, vol. 3, no. 4, pp. 44–55, Oct./Dec. 1967.

[16] A. Sahai, "Why block-length and delay behave differently if feedback is present," *IEEE Trans. Inform. Theory*, Submitted. [Online]. Available: http://www.eecs.berkeley.edu/˜sahai/Papers/FocusingBound.pdf

[17] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, Jul./Oct. 1948.

[18] ——, "Coding theorems for a discrete source with a fidelity criterion," *IRE National Convention Record*, vol. 7, no. 4, pp. 142–163, 1959.

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

[20] C. E. Shannon, "The zero error capacity of a noisy channel," *IEEE Trans. Inform. Theory*, vol. 2, no. 3, pp. 8–19, Sep. 1956.

[21] A. Sahai and S. K. Mitter, "The necessity and sufficiency of anytime capacity for stabilization of a linear system over a noisy communication link. part I: scalar systems," *IEEE Trans. Inform. Theory*, vol. 52, no. 8, pp. 3369–3395, Aug. 2006.

[22] G. D. Forney, "Convolutional codes II. maximum-likelihood decoding," *Information and Control*, vol. 25, no. 3, pp. 222–266, Jul. 1974.

[23] ——, "Convolutional codes III. sequential decoding," *Information and Control*, vol. 25, no. 3, pp. 267–297, Jul. 1974.

[24] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inform. Theory*, vol. 13, no. 2, pp. 260–269, Apr. 1967.

[25] I. M. Jacobs and E. R. Berlekamp, "A lower bound to the distribution of computation for sequential decoding," *IEEE Trans. Inform. Theory*, vol. 13, no. 2, pp. 167–174, Apr. 1967.

[26] R. Gallager, "Low-Density Parity-Check Codes," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 1960.

[27] T. Richardson and R. Urbanke, *Modern Coding Theory*. Cambridge University Press, 2007.

[28] M. Lentmaier, D. V. Truhachev, K. S. Zigangirov, and D. J. Costello, "An analysis of the block error probability performance of iterative decoding," *IEEE Trans. Inform. Theory*, vol. 51, no. 11, pp. 3834–3855, Nov. 2005.

[29] T. J. Richardson and R. L. Urbanke, "Efficient encoding of low-density parity-check codes," *IEEE Trans. Inform. Theory*, vol. 47, no. 2, pp. 638–656, 2001.

[30] P. P. Sotiriadis, V. Tarokh, and A. P. Chandrakasan, "Energy reduction in VLSI computation modules: an information-theoretic approach," *IEEE Trans. Inform. Theory*, vol. 49, no. 4, pp. 790–808, Apr. 2003.

[31] N. Shanbhag, "A mathematical basis for power-reduction in digital VLSI systems," *IEEE Trans. Circuits Syst. II*, vol. 44, no. 11, pp. 935–951, Nov. 1997.

[32] T. Koch, A. Lapidoth, and P. P. Sotiriadis, "A channel that heats up," in *Proceedings of the 2007 IEEE Symposium on Information Theory*, Nice, France, 2007.

[33] ——, "A hot channel," in *2007 IEEE Information Theory Workshop (ITW)*, Lake Tahoe, CA, 2007.

[34] R. M. Corless, G. H. Gonnet, D. E. G. Hare, and D. E. Knuth, "On the Lambert W function," *Advances in Computational Mathematics*, vol. 5, pp. 329–359, 1996.

[35] A. Khandekar, "Graph-based codes and iterative decoding," Ph.D. dissertation, California Institute of Technology, Pasadena, CA, 2002.

[36] A. Khandekar and R. McEliece, "On the complexity of reliable communication on the erasure channel," in *IEEE International Symposium on Information Theory*, 2001.

[37] T. H. Cormen, C. E. Leiserson, and R. L. Rivest, *Introduction to Algorithmns*, 1989.

[38] I. Sason, "Bounds on the number of iterations for turbo-like ensembles over the binary erasure channel," *submitted to IEEE Transactions on Information Theory*, 2007.

[39] H. Palaiyanur, personal communication, Oct. 2007.

[40] P. Grover, "Bounds on the tradeoff between rate and complexity for sparse-graph codes," in *2007 IEEE Information Theory Workshop (ITW)*, Lake Tahoe, CA, 2007.

[41] I. Sason and S. Shamai, *Performance Analysis of Linear Codes under Maximum Likelihood Decoding: A tutorial*. Hanover, MA: Foundations and Trends in Communication and Information theory, NOW Publishing, 2006.

[42] A. Barg and G. Zemor, "Error exponents of expander codes," *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1725–1729, Jun. 2002.

[43] G. D. Forney, "Exponential error bounds for erasure, list, and decision feedback schemes," *IEEE Trans. Inform. Theory*, vol. 14, pp. 206–220, 1968.

[44] M. V. Burnashev, "Data transmission over a discrete channel with feedback, random transmission time," *Problemy Perdachi Informatsii*, vol. 12, no. 4, pp. 10–30, Oct./Dec. 1976.

[45] M. Grossglauser and D. Tse, "Mobility increases the capacity of adhoc wireless networks," *IEEE/ACM Trans. Networking*, vol. 10, pp. 477–486, Aug. 2002.

[46] C. Rose and G. Wright, "Inscribed matter as an energy-efficient means of communication with an extraterrestrial civilization," *Nature Letter*, pp. 47–49, Sep. 2004.

[47] V. Guruswami and P. Indyk, "Linear-time encodable/decodable codes with near-optimal rate," *IEEE Trans. Inform. Theory*, vol. 51, no. 10, pp. 3393–3400, Oct. 2005.

[48] L. R. Varshney, "Performance of LDPC codes under noisy message-passing decoding," *2007 Information Theory Workshop*, pp. 178–183, Sep. 2007.

[49] P. Ruján, "Finite temperature error-correcting codes," *Physical Review Letters*, vol. 70, no. 19, pp. 2968–2971, May 1993.

[50] E. A. Haroutunian, "Lower bound for error probability in channels with feedback," *Problemy Peredachi Informatsii*, vol. 13, no. 2, pp. 36–44, 1977.

[51] C. Chang, personal communication, Nov. 2007.